UNIVERSITY OF THESSALY, GREECE

Optimization and game theory techniques for energy-constrained networked systems and the smart grid

Doctor of Philosophy

in

Electrical & Computer Engineering

by

Lazaros Gkatzikis

July 2013

Dissertation Committee:

Prof. Leandros Tassiulas,

Assistant Prof. Iordanis Koutsopoulos

Associate Prof. Slawomir Stanczak

The Dissertation of Lazaros Gkatzikis is approved by:

_____

_____

_____
Committee Chairperson

University of Thessaly

ACKNOWLEDGMENTS

This thesis represents the culmination of research that was conducted towards my PhD degree from the Department of Electrical and Computer Engineering, University of Thessaly, Greece.

I would like to express my gratitude to all those who made this dissertation possible, and first of all to my supervisor, Assistant Professor Iordanis Koutsopoulos for giving me the opportunity to work in such a joyful and inspiring research unit. Without his guidance and constructive criticism this doctoral study would not have been possible. His patience and understanding have been a great support in this long journey. He has also given me the opportunity to participate in interesting European Union and National research projects, an experience that has broadened my research view. I would also like to thank the other members of the dissertation committee for their contribution and the examination committee for reading this thesis.

I would like to thank my friends and lab-mates, who provided help and made this long journey a lot more enjoyable and memorable. Particularly, I would like to thank Dr. Paris Flegkas, Dr. Georgios Paschos and Vasilis Sourlas for their help and fruitful cooperation in many aspects of my research.

Special thanks go to my family for their love and support. There are no words that can express my gratitude and appreciation for all they have done for me. The least I can do in return is to dedicate this thesis to them.

Dedicated to my family.

ΠΕΡΙΛΗΨΗ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

"Τεχνικές Βελτιστοποίησης και Θεωρίας Παιγνίων για Δικτυακά Συστήματα Περιορισμένης Ενέργειας και Ευφυή Δίκτυα Ηλεκτρισμού"

Λάζαρος Γκατζίκης

Τα τελευταία χρόνια παρατηρείται συνεχής αύξηση των ενεργειακών αναγκών της ανθρωπότητας. Επομένως, η αποδοτική αξιοποίηση της ενέργειας έχει καταστεί ένας βασικός στόχος της ερευνητικής κοινότητας. Οι τρέχουσες ερευνητικές προσπάθειες επικεντρώνονται σε δύο κατευθύνσεις, *i*) την βελτιστοποίηση της αποδοτικότητας και της αξιοπιστίας του δικτύου ηλεκτροδότησης και   *ii*) την βελτιστοποίηση της ενεργειακής αποδοτικότητας μεμονωμένα κάθε συσκευής. Η παρούσα διδακτορική διατριβή εξετάζει και τις δύο αυτές κατευθύνσεις.

Η αποδοτική και οικονομική λειτουργία του δικτύου ηλεκτρισμού απαιτεί το φορτίο να είναι κατανεμημένο ομοιόμορφα εντός της ημέρας. Επί του παρόντος η κοστολόγηση της καταναλισκόμενης ενέργειας στους τελικούς χρήστες είναι γενικά σταθερή κατά την διάρκεια μιας μέρας. Δεδομένου ότι η καθημερινή δραστηριότητα των ανθρώπων είναι παρόμοια, αυτό οδηγεί σε σημαντικά ανομοιόμορφη κατανομή της ζήτησης. Αντίθετα, η εφαρμογή χρονικά μεταβαλλόμενων τιμών (δυναμική τιμολόγηση) παρακινεί τους χρήστες να μεταβάλλουν το ημερήσιο προφίλ κατανάλωσης ενέργειας ώστε να ελαχιστοποιήσουν τον λογαριασμό ρεύματος. Αυτό ταυτόχρονα οδηγή σε πιο εξισορροπημένη συνολική κατανάλωση ενέργειας και συνεπώς καλύτερη λειτουργία του δικτύου. Σε αυτήν την κατεύθυνση, εισάγουμε ένα ρεαλιστικό μοντέλο της συμπεριφοράς των οικιακών καταναλωτών ρεύματος, μοντελοποιούμε την αντίδραση τους σε χρονικά μεταβαλλόμενες τιμές ρεύματος και αναπτύσσουμε μηχανισμούς δυναμικής τιμολόγησης βασισμένους στην θεωρία βελτιστοποίησης.

 Ο μεγάλος αριθμός όμως οικιακών χρηστών και η έλλειψη τεχνογνωσίας σχετικά με την δυναμική τιμολόγηση από την πλευρά του διαχειριστή δημιούργησε την ανάγκη για μια νέα οντότητα στην αγορά του ηλεκτρικού ρεύματος. Οι διαμεσολαβητές (aggregators) λειτουργούν ως ενδιάμεσοι που συντονίζουν τους χρήστες ώστε να μεταθέσουν χρονικά ή και να μειώσουν τις απαιτήσεις τους για ηλεκτρικό ρεύμα και μεταπωλούν αυτή την υπηρεσία στον διαχειριστή του συστήματος ηλεκτρισμού. Σε αυτό το πλαίσιο, προτείνουμε ένα ιεραρχικό μοντέλο τριών επιπέδων για την αγορά ηλεκτρισμού και έναν μηχανισμό τιμολόγησης για κάθε επίπεδο. Χρησιμοποιώντας ρεαλιστικά δεδομένα κατανάλωσης

δείχνουμε ότι ο προτεινόμενος μηχανισμός μπορεί να εγγυηθεί σημαντικά οικονομικά οφέλη για κάθε οντότητα της αγοράς (καταναλωτές, διαμεσολαβητές και ο διαχειριστής).

Παράλληλα με την βελτιστοποίηση της λειτουργίας του δικτύου ηλεκτρισμού, επιτακτική είναι και η βέλτιστη αξιοποίηση της ενέργειας σε επίπεδο συσκευής. Οι σύγχρονες φορητές συσκευές τροφοδοτούνται από μπαταρίες και συνεπώς έχουν περιορισμένες ενεργειακές δυνατότητες. Αντίθετα οι σύγχρονες εφαρμογές είναι ιδιαίτερα απαιτητικές και δεν μπορούν να εκτελεστούν τοπικά στην συσκευή. Επομένως, η εκτέλεση κάποιων εργασιών σε ένα απομακρυσμένο σύστημα (γενικά αναφέρεται ως υπολογιστικό νέφος) μπορεί να αυξήσει σημαντικά την αυτονομία και τις δυνατότητες μιας φορητής συσκευής. Στην διδακτορική αυτή διατριβή εξετάζουμε το πρόβλημα του βέλτιστου χρονοπρογραμματισμού εργασιών που γεννιούνται από φορητές συσκευές με διαφορετικό και άγνωστο ρυθμό στους διάφορους κόμβους του υπολογιστικού νέφους.

Επίσης, ο σχεδιασμός αλγορίθμων ανάθεσης πόρων σε ασύρματα συστήματα οφείλει να λαμβάνει υπόψη τους ενεργειακούς περιορισμούς των φορητών συσκευών. Σε αυτή την κατεύθυνση, η συγκεκριμένη διδακτορική διατριβή ερευνά το πρόβλημα της ασύρματης πρόσβασης στο μέσο για φορητές συσκευές και συγκεκριμένα την αντιστάθμιση (tradeoff) ενέργειας –ρυθμοαπόδοσης. Αρχικά μελετάμε την επίδραση του ανταγωνισμού για το κανάλι στην κατανάλωση ενέργειας ασύρματων κόμβων που υποστηρίζουν καταστάσεις ύπνου (sleep modes) και μοντελοποιούμε τον ανταγωνισμό των χρηστών για το μέσο ως ένα μη συνεργατικό παίγνιο δύο επιπέδων. Στο χαμηλότερο επίπεδο καθορίζεται η πρόσβαση στο μέσο, ενώ στο ανώτερο ο προγραμματισμός των περιόδων ύπνου και δείχνουμε ότι υπάρχει ένα μοναδικό σημείο ισορροπίας. Τέλος, εξετάζουμε το πρόβλημα ελέγχου ισχύος και της βέλτιστης ανάθεσης αναμεταδοτών (relays) σε περιβάλλον παρεμβολών. Αναγνωρίζοντας την σημασία της παρεμβολής στα κυψελωτά δίκτυα προτείνουμε αλγορίθμους για την μεγιστοποίηση του επιτεύξιμου ρυθμού μετάδοσης, που βασίζονται στην ιδέα της εύρεσης του ταιριάσματος μέγιστου βάρους σε ένα διμερή γράφο και την χρήση γεωμετρικού προγραμματισμού.

Ένα σημαντικό κοινό χαρακτηριστικό όλων αυτών των προβλημάτων είναι η αυτόνομη φύση των εμπλεκόμενων οντοτήτων. Κάθε χρήστης επιλέγει ανεξάρτητα τις ενέργειες του με γνώμονα τη μεγιστοποίηση του προσωπικού του οφέλους, μια επιλογή που γενικά δεν είναι η βέλτιστη συνολικά για το σύστημα. Για το σκοπό αυτό χρησιμοποιούμε τεχνικές βελτιστοποίησης και θεωρίας παιγνίων για να μοντελοποιήσουμε και να αναλύσουμε την συμπεριφορά και την αλληλεπίδραση των χρηστών.

ABSTRACT OF THE DISSERTATION

Optimization and game theory techniques for energy-constrained networked
systems and the smart grid

by

Lazaros Gkatzikis

Doctor of Philosophy, Graduate Program in Electrical and Computer Engineering
University of Thessaly, July   2013
Prof. Leandros Tassiulas, Chairperson

The energy needs of all sectors of our modern societies are constantly increasing. Indicatively, *annual worldwide demand for electricity has increased ten-fold within the last* 50 *years*. Thus, energy efficiency has become a major target of the research community. The ongoing research efforts is focused on two main threads, *i*) optimizing efficiency and reliability of the power grid and *ii*) improving energy efficiency of individual devices / systems. In this thesis we explore the use of optimization and game theory techniques towards both goals.

Stable and economic operation of the power grid calls for electricity demand to be uniformly distributed across a day. Currently, the price of electricity is fixed throughout a day for most users. Given also the highly correlated daily schedules of the users, this leads to unbalanced distribution of demand. However, the recent development of low-cost smart meters enables bidirectional communication between the electricity operator and each user, and hence introduces the option of dynamic pricing and demand adaptation (a.k.a. Demand Response - DR). *Dynamic pricing motivates home users to modify their electricity consumption profile so as to reduce their electricity bill*. Eventually, users by moving demand out of peak consumption periods lead to a more balanced total demand pattern and a more stable grid.

A DR scheme has to balance the contradictory interests of the utility operator and the users. On the one hand, the operator wants to minimize electricity generation cost. On the other hand, each user aims to maximizes a utility function that captures the trade-off between timely execution of demands and financial savings. In this thesis we focus on designing efficient DR schemes for the residential sector. Initially, we introduce a realistic model of user's response to time-varying prices and identify the operating constraints of home appliances that make optimal demand scheduling NP-Hard. Thus, we devise an optimization-based dynamic pricing mechanism and demonstrate its applicability through a day-ahead DR market. Our numerical results underline the potential of residential DR and verify that our scheme exploits DR benefits more efficiently compared to existing ones.

The large number of home users though and the fact that the utility operator generally lacks the know-how of designing and applying dynamic pricing at such a large scale introduced the need for

a new market entity. *Aggregators act as intermediaries that coordinate home users to shift or even curtail their demands and then resell this service to the utility operator.* In this direction, we introduce a three-level hierarchical model for the smart grid market and we devise the corresponding pricing mechanism for each level. The operator seeks to minimize the smart grid operational cost and offers rewards to aggregators toward this goal. Aggregators are profit-maximizing entities that compete to sell DR services to the operator. Finally, end-users are also self-interested and seek to optimize the tradeoff between earnings and discomfort. Based on realistic demand traces we demonstrate the dominant role of the utility operator and how its strategy affects the actual DR benefits. Although the proposed scheme guarantees significant financial benefits for each market entity, interestingly, users that are extremely willing to modify their consumption pattern do not derive the maximum financial benefit.

In parallel to optimizing the power grid, per device energy economy has become a goal of utmost performance. Contemporary mobile devices are battery powered and hence characterized by limited processing and energy resources. In addition, the latest mobile applications are particularly demanding and hence cannot be executed locally. Instead, *a mobile device can outsource its computationally intensive tasks to the cloud over its wireless access technology, so as to maximize both its lifetime and performance.* In this thesis, we explore task offloading and Virtual Machine (VM) migration mechanisms that minimize energy consumption and execution time. We identify that in order to decide whether offloading is beneficial, a mobile has also to consider the delay and energy cost induced from communication with the cloud. On the other hand, the challenge for the cloud is to optimally allocate the arising VMs to its servers. Providing quality of service guarantees is particularly challenging in the dynamic cloud environment, due to the time-varying bandwidth of the access links, the ever changing available processing capacity at each server and the time-varying data volume of each VM. Thus, we propose a mobile cloud architecture that brings the cloud closer to the user and online VM migration policies spanning fully uncoordinated ones, in which each user or server autonomously makes its migration decisions, up to cloud-wide ones.

The transceiver is one of the most power consuming components of a mobile wireless device. Since the medium access layer controls when a transmission takes place, it has significant impact on overall energy consumption and consequently on the lifetime of a device. In this direction, we investigate a scenario of competing wireless devices that support sleep modes and we provide a characterization of the energy-throughput tradeoff. We derive the optimal throughput under energy constraints and model contention for the wireless medium as a non-cooperative game, where the strategy of each user constitutes in its access probability and its sleep mode schedule. We show that the resulting game has a unique Nash Equilibrium Point and that energy constraints reduce the negative impact of selfish behaviour, leading to bounded price of anarchy. We devise also a modified medium access scheme, where the state of the medium can be sampled in the beginning of each frame and show that it leads to improved exploitation of the medium without any explicit cooperation.

Finally, we move to a scenario where concurrent transmissions over the same channel are not destructive but lead to reduced performance due to interference. In this context, we consider the

problem of power control and optimal relay assignment and we develop interference-aware sum-rate maximization algorithms. Our proposed schemes make use of a bipartite maximum weight matching formulation of the problem and geometric programming and are amenable to distributed implementation. We also identify the importance of interference for cell-edge users in cellular networks and demonstrate that our schemes bring together two main features of 4G systems, namely *interference management* and *relaying*.

Publications

The results of this research are included in the following publications.

Journals and Magazines

[J.01] L. Gkatzikis, I. Koutsopoulos and T. Salonidis, "The Role of Aggregators in Smart Grid Demand Response Markets," in IEEE JSAC- Special series on Smart Grid Communications, vol.31, no.7, pp 1247 - 1257, July 2013.

[J.02] L. Gkatzikis and I. Koutsopoulos, "Migrate or Not? Exploiting Dynamic Task Migration in Mobile Cloud Computing Systems," in IEEE Wireless Communications Magazine: Special Issue on Mobile cloud computing vol.20, no.3, June 2013.

[J.03] L. Gkatzikis, and I. Koutsopoulos, "Mobiles on Cloud Nine: Efficient Task Migration Policies for Cloud Computing Systems," under review in Elsevier Computer Networks, Special Issue on Communications and Networking in the Cloud.

[J.04] L. Gkatzikis, T. Salonidis, N. Hegde and L. Massoulie, "The Impact of Shiftable Demands on Residential Demand Response," to be submitted in IEEE Transactions in Smart Grid.

Conferences

[C.01] L. Gkatzikis, G. Iosifidis, I. Koutsopoulos and L. Tassiulas, "Collaborative Placement and Use of Storage Resources in the Smart Grid," under review in *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2013 .

[C.02] L. Gkatzikis, G.S. Paschos and I. Koutsopoulos, "Medium Access Games: The impact of energy constraints," in proc. of International Conference on Network Games, Control and Optimization (NETGCOOP), 2011.

[C.03] L. Gkatzikis and I. Koutsopoulos, "Low Complexity Algorithms for Relay Selection and Power Control in Interference-Limited Environments," in proc. of Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2010, Avignon, France.

In addition, our research efforts within the same period led to the following publications that are not directly related to this thesis.

Journals

[J.05] Vasilis Sourlas, Lazaros Gkatzikis, Paris Flegkas and Leandros Tassiulas, "Autonomic Cache Management and Performance Limits in Information-Centric Networks," to appear in IEEE Transaction on Network and Service Management (TNSM), 2013.

[J.06] I. Koutsopoulos, L. Tassiulas and L. Gkatzikis, "Client and Server Games and Nash Equilibria in Peer-to-Peer Networks," under review in Elsevier Computer Networks.

Conferences

[C.04] P. Mannersalo, G.S. Paschos and L. Gkatzikis, "Geometrical Bounds on the Efficiency of Wireless Network Coding," to appear in proc. of Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2013.

[C.05] L. Gkatzikis, T. Salonidis, N. Hegde and L. Massoulie, "Electricity Markets Meet the Home through Demand Response," in proc. of IEEE Conference on Decision and Control (CDC), Maui, Hawai 2012.

[C.06] L. Gkatzikis, T. Tryfonopoulos and I. Koutsopoulos, "An Efficient Probing Mechanism for Next Generation Mobile Broadband Systems," in proc. of IEEE Wireless Communications and Networking Conference (WCNC), Paris, France, 2012.

[C.07] Vasilis Sourlas, Paris Flegkas, Lazaros Gkatzikis and Leandros Tassiulas, "Autonomic Cache Management in Information-Centric Networks," in 13th IEEE/IFIP Network Operations and Management Symposium (NOMS 2012), pp. 121-129, Hawaii, USA, April 2012.

[C.08] Vasilis Sourlas, Lazaros Gkatzikis and Leandros Tassiulas, "On-Line Storage Management with Distributed Decision Making for Content-Centric Networks," in 7th Conference on Next Generation Internet (NGI) 2011, pp. 1-8, Kaiseslautern, Germany, June 2011.

[C.09] I. Koutsopoulos, L. Tassiulas and L. Gkatzikis, "Client and Server Games in Peer-to-Peer Networks," in proc. of IEEE International Workshop on QoS (IWQoS), 2009, Charleston, SC, USA.

Contents

List of Figures

List of Tables

List of Abbreviations

AaF	....	Amplify and Forward
AP	....	Access Point
BS	....	Base Station
CSI	....	Channel State Information
DaF	....	Decode and Forward
DR	....	Demand Response
ICT	....	Information and Communications Technology
ISO	....	Independent System Operator
LOS	....	Line Of Sight
LTE	....	Long-Term Evolution
MCC	....	Mobile Cloud Computing
MIMO	....	Multiple Input Multiple Output
MNO	....	Mobile Network Operator
MS	....	Mobile Station
MWM	....	Maximum Weighted Matching
NEP	....	Nash Equilibrium Point
QoS	....	Quality of Service
OFDMA	....	Orthogonal Frequency Division Multiple Access
PoA	....	Price of Anarchy
RS	....	Relay Station
SINR	....	Signal to Interference and Noise Ratio
SLA	....	Service Level Agreement
VM	....	Virtual Machine
WiFi	....	Wireless Fidelity

Chapter 1

Introduction

1.1 Motivation

In the last decades, the energy needs of all sectors of our modern societies have risen significantly. Indicatively, annual worldwide demand for electricity has increased ten-fold within the last 50 years and has almost doubled in the last decade [1]. Thus, energy efficiency has become a major concern of the research community. The ongoing research efforts are classified into two main threads, *i*) optimizing the efficiency and reliability of the power grid and *ii*) improving energy efficiency of individual devices or systems per se.

1.1.1 Efficiency of the power grid

Electricity market is about to undergo a paradigm shift. From a centralized model of operation, where all the decision making is performed by the utility operator, we are moving towards a complex market structure. Traditionally, the utility operator has been responsible for generation, transmission and distribution of electric power as well as pricing of consumed electricity to the end-users. The recent deregulation trend though promotes the separation of these activities. Typically, distribution and transmission are undertaken by non-profit operators, whereas generation and retailing are left in the hands of free market. The structure of the electricity market is depicted in Fig. 1.1.

The role of the utility operator (a.k.a. Independent System Operator - ISO) has been reduced to selecting and utilizing the least expensive resources to meet energy demand, a principle referred to as *economic dispatch*. Utility is responsible for predicting the total consumption for each period of the following day and conducts a day-ahead wholesale auction where a number of competing generators offer their electricity output. Each generator makes a bid indicating the amount of energy that it can provide at each given time period and the corresponding price. The best offers that meet predicted demand are accepted. Generally, the market clearing price is the one of the highest bid accepted, in an attempt to guarantee fairness among the generators. Evidently, since electricity supply and demand have to be constantly balanced, higher demand calls for additional and more expensive powerplants to be activated.
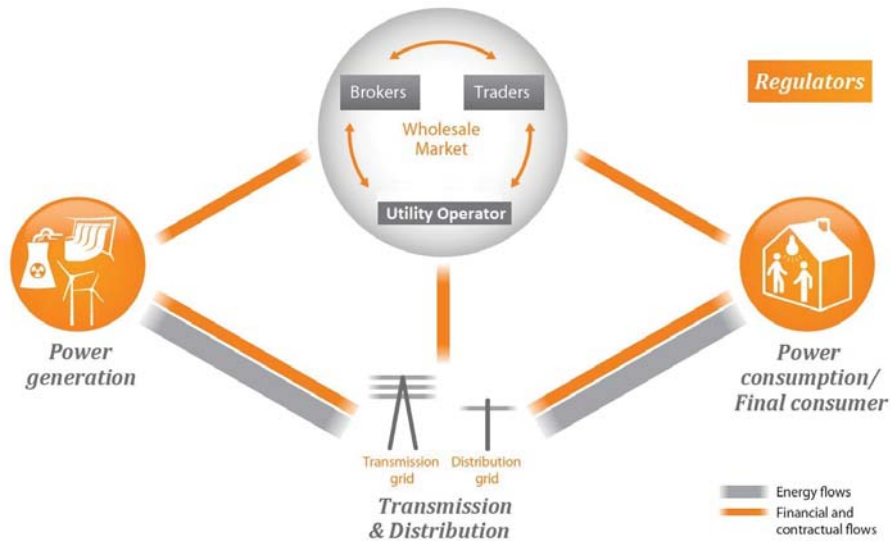
Figure 1.1: The operation of the electricity market.

The latest advancements in the electricity market though, such as market liberalization and availability of low-cost smart meters, introduced the alternative option of dynamic demand adaptation. In this direction, *demand response (DR) programs provide incentives to the users, usually in the form of dynamic pricing, to reduce their electricity consumption in peak demand periods*. Compared to activating additional power plants, DR can take place in a much faster timescale, almost in real time, and leads to a more stable power grid system, significantly lower electricity generation cost and reduced CO_2 emissions. In general, DR programs pursue the win-win situation where end-users enjoy a reduced electricity bill, while the operator enjoys reduced operating cost.

Although DR has been successfully applied in the industry, its application in the residential sector is limited. According to a recent report [2] the main forces constraining rapid growth of residential DR market are consumer backlash to smart meters, the cost of delivering demand response to residential customers, and the lack of knowledge and time from the users' side to respond to dynamic prices.

On the other hand, compared to the large industrial loads, small residential ones can provide more reliable and faster response and are spatially distributed. Given also that the residential sector accounts for a significant portion of the total electricity consumption (31% in UK), DR mechanisms for the home hold enormous potential. Currently, most homes are enrolled in static pricing programs ($> 98\%$) and schedule their demands for the most convenient time. Given the highly correlated daily schedules of home users, this leads to unbalanced electricity demand throughout the day (Fig. 1.2).

This thesis aims to design and implement efficient DR mechanisms that will pass residential demand response from the realm of theory to practice. In particular, we focus on DR market models that exploit dynamic pricing to incentivize users either to shift their demands out of peak consumption periods or even to curtail their total consumption. The ultimate objective is to devise realistic DR models that capture the particularities of residential demands and efficient pricing schemes that lead to

2
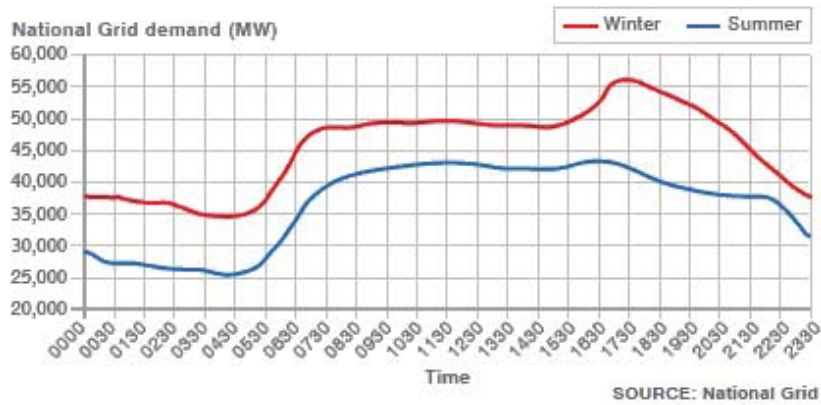
Figure 1.2: Daily demand distribution for a typical winter / summer day in UK.

a more balanced demand pattern. Based on the derived models we quantify the potential of residential DR in terms of financial savings for both the home-users and the operator.

1.1.2 Energy efficiency of networked systems

Apart from optimizing the operation of the power grid itself, minimizing energy consumption of mobile devices and networked systems is an active area of research [3]. Currently, Information and Communications Technologies (ICT) account for almost 5% of worldwide electricity consumption, with communication networks and datacenters being the major electricity consumers (Fig. 1.3). To counteract this, industry has devoted its efforts on designing energy-efficient hardware, either in the form of low-consumption mobile devices or efficient datacenters [6].
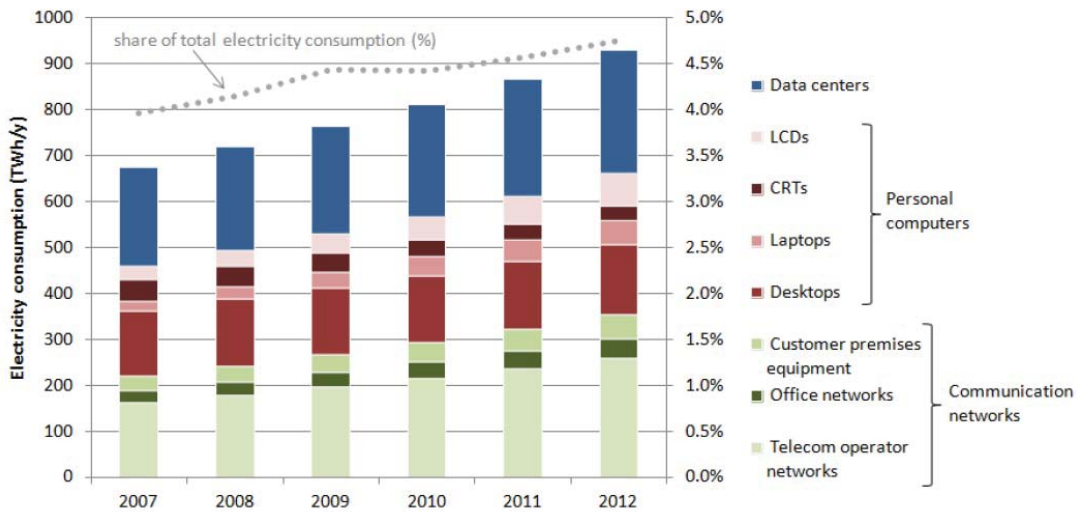
Figure 1.3: Worldwide electricity consumption of ICT (source:[4]).

3
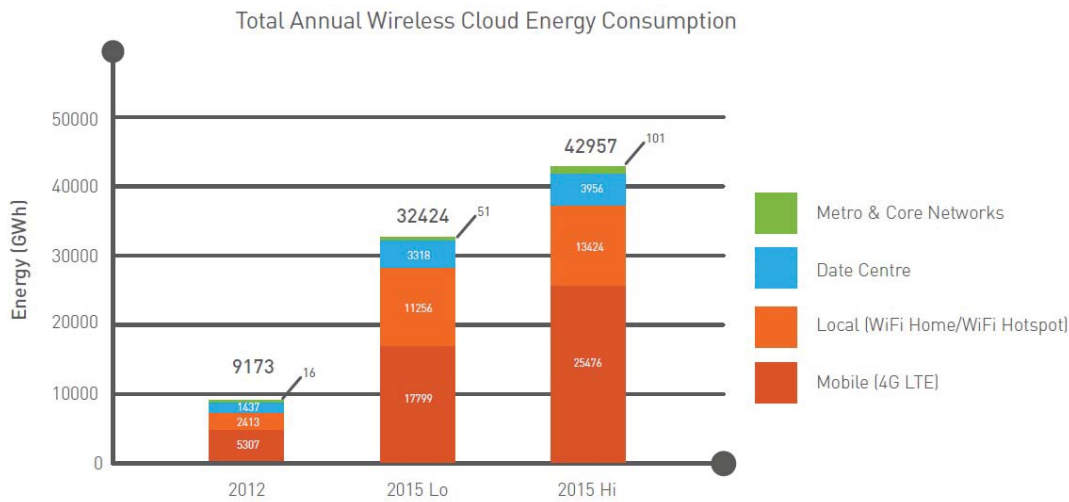
Figure 1.4: Annual energy consumption broken down into the components of the wireless cloud ecosystem in 2012 and two projections (low and high) for 2015 (source:[7])

Recently though, the authors of [7] showed that *wireless access is responsible for* 90% *of total energy consumption of the wireless cloud* (Fig. 1.4). Based on the emerging trend of cloud services being accessed via wireless communication networks such as WiFi and 4G LTE, they suggest that energy efficiency of wireless access networks should be pursued instead. In this thesis, we explore energy efficiency at the level of mobile device, wireless access and cloud infrastructure.

Energy efficiency can be viewed as a problem of maximizing the performance of a wireless system under a given energy budget. Mobile devices are battery powered and hence are characterized by limited energy capabilities. Thus, designing communication protocols and resource allocation mechanisms that take full advantage of the available energy is a goal of utmost importance. In this direction, we attempt a characterization of the energy-performance tradeoff in wireless access networks. In particular, we consider scenarios ranging from Mobile Cloud Computing, where energy-constrained mobile devices offload their tasks to the cloud, up to wireless communication scenarios, where we investigate how contention for the shared wireless medium and energy constraints affect the achievable throughput.

Contention for the medium may be realized in any of the following two forms. In case of probabilistic medium access, any two concurrent transmissions to a single receiver lead to a collision and hence energy gets wasted. On the other hand, concurrent transmissions between remote communication pairs over the same frequency band are generally not destructive, but cause interference to each other and hence lead to reduced performance. In this thesis we explore both cases and show that some form of explicit or implicit coordination of transmissions has to be enforced, so as to optimally exploit the available energy.

1.2 Synopsis

In this thesis we propose game theoretic models and optimization-based mechanisms for the smart grid and for energy-constrained systems. The common ground of the investigated scenarios is the autonomous nature of the participating entities. Each user acts independently towards maximizing her own utility, a strategy that generally does not coincide with the socially optimal. Thus, we investigate mechanisms that align the interests of the involved entities and lead the system to more efficient operating points.

In **Chapter 2** we provide an introduction to demand response and the structure of the smart grid market. Then, we introduce a novel DR model, where the operator sets the prices and multiple home users respond by scheduling their demands. The objective of the operator is to minimize electricity generation cost, whereas each user maximizes a utility function that captures the trade-off between timely execution of demands and financial savings. We show that optimal demand scheduling is an NP-hard problem and derive an efficient price setting strategy that results to more balanced demand throughout the day. We exploit price discrimination to further improve user utility and use per appliance demand traces to demonstrate that our scheme aligns the interests of the operator and the home users.

In **Chapter 3** we extend our DR market model in order to capture the role of the aggregators that act as intermediaries between the utility operator and the home users. In the proposed market, the operator seeks to minimize the smart grid operational cost and offers rewards to aggregators toward this goal. Profit-maximizing aggregators compete to sell DR services to the operator and provide compensation to end-users in order to modify their preferable consumption pattern. Finally, end-users seek to optimize the tradeoff between earnings received from the aggregator and discomfort from having to modify their pattern. Initially, we investigate a benchmark scenario where all the decision making is performed by the operator under full information. Next, we present how the proposed model could be realised in the form of a hierarchical DR market and investigate the required information exchange. We also use daily demand traces to quantify the benefits arising from the proposed scheme.

Next, we move to the setting of per device energy economy and investigate the throughput-energy tradeoff. In particular, **Chapter 4** is devoted to efficient VM (Virtual Machine) migration mechanisms for the mobile cloud computing paradigm, where mobile devices outsource their computationally intensive tasks to the cloud. Mobile devices aim to minimize their energy consumption. On the other hand, the challenge for the cloud provider is to minimize task execution and data download time to the user, whose location changes due to mobility. In this direction, we model interaction of tasks / VMs within the cloud and propose migration mechanisms that capture the interaction of co-located tasks. We also demonstrate how VM migrations can be exploited towards a more energy-efficient cloud. We verify numerically that VM migrations can provide significant performance benefits compared to one-shot VM placement schemes.

In **Chapter 5** we consider random medium access schemes for devices that support sleep modes, i.e. turning off electronic compartments for energy saving. Due to hardware limitations, sleep mode

transitions cannot occur at the medium access timescale, but only at a slower timescale. Each terminal can choose when to turn on/off and its probability to transmit on an arbitrary slot. Thus, we develop a two level model, consisting of a fast timescale for transmission scheduling and a slower timescale for the sleep mode transitions. We take a game theoretic approach to model the user interactions and show that the energy constraints modify the medium access problem significantly. Interestingly, energy constraints lead to bounded price of anarchy and hence introduce some form of coordination. Our results give valuable insights on the energy–throughput tradeoff for contention based systems.

In **Chapter 6** we consider the scenario of an interference-limited wireless network, where multiple source-destination pairs compete for the same pool of relay nodes. In an attempt to maximize the sum-rate performance of the system, we address the joint problem of relay assignment and power control. Initially, we study the autonomous scenario, where each source greedily selects the strategy (transmission power and relay) that maximizes its individual rate, leading to a simple one-shot algorithm of linear complexity. Then, we propose a more sophisticated algorithm of polynomial complexity that is amenable to distributed implementation through appropriate message passing. We also generalize our proposed algorithms so as to fit to the LTE-Advanced framework.

Finally, in **Chapter 7** we summarize the findings of our work, we provide guidelines for the design of practical energy-efficient systems and we discuss open research problems in the area.

Chapter 2

Energy Efficiency in the Smart Grid - Part I: The Impact of Shiftable Demands on Demand Response

In this chapter, we explore the potential of demand response (DR) towards a more stable power grid. Most existing DR programs are tailored specifically to the needs of industrial and commercial clients. Residential appliances though are characterized by significantly different and diverse operating constraints. A major differentiating factor is the *shiftable nature of residential demands* that introduces the need for a novel DR approach. In this direction, we introduce a model for the day-ahead electricity market, where the operator sets the prices and multiple home users respond by scheduling their demands. The objective of the operator is to minimize electricity generation cost, whereas each user maximizes her utility function, which captures the trade-off between timely execution of demands and financial savings.

We propose a DR model that captures the diverse energy characteristics of different home appliances and derive an efficient price setting strategy that leads to a more balanced demand pattern throughout the day. We exploit price discrimination to further improve user utility and use per appliance demand traces to demonstrate that our scheme aligns the interests of the operator and the home users. Since most home users lack the knowledge and time to respond to dynamic prices, automated DR schemes will eventually have to be developed. Any such scheme though would rely on an estimation of the actual user utility. In this direction, we quantify the impact of estimation accuracy on the DR benefits of each market entity.

2.1 Introduction to Demand Response

For the proper operation of the power grid, electricity supply and demand have to be constantly balanced. Currently, this is achieved by activating additional powerplants to increase available supply and meet the arising demand. The latest advancements in the electricity market though, such as

market liberalization and increased availability of smart meters, introduced the alternative option of dynamic demand adaptation. In this direction, demand response (DR) programs provide incentives to the users, usually in the form of dynamic pricing, to reduce their electricity consumption in peak demand periods. Compared to activating additional power plants, DR can take place in a much faster timescale, almost in real time, and leads to significantly lower electricity generation cost and reduced CO_2 emissions.

Although DR has been successfully applied in the industry, its application in the residential sector is limited. Compared to the large industrial loads, small residential ones can provide more reliable and faster response and are spatially distributed. Given also that the residential sector accounts for a significant portion of the total electricity consumption (31% in UK) DR mechanisms for the home have enormous potential. Currently, most homes are enrolled in static pricing programs and schedule their demands for the most convenient time. Given the highly correlated daily schedules of home users, this leads to unbalanced electricity demand throughout the day.

Dynamic pricing motivates home users to change their electricity consumption behavior and spread demand throughout the day. A more balanced demand distribution can lead to significant benefits for all the involved entities. In particular, the gain for the operator is twofold, namely a more stable electricity network and reduced generation cost, since only the least expensive power plants have to be used to cover demand. Home users on the other hand can shift their demands to low price periods and reduce their electricity bill. However, this shifting usually causes inconvenience to the user, since it forces changes in her daily schedule. Given the day–ahead prices, each user has to balance inconvenience and financial savings and schedule her demands towards maximizing net utility.

However, the broader market penetration of residential DR is hampered by the lack of knowledge and time of the users to respond to dynamic prices [20]. Indicatively, the authors of [8] studied the behaviour of home users charged by hourly time-varying prices and identified that most exhibit an inherent discomfort in observing the price evolution throughout the day and adjusting their energy consumption profile correspondingly. To counteract this, we propose an automated DR mechanism that requires minimal human intervention. In the home side, an estimation of the appliances to be operated the following day and the exact time of activation is performed based on historical data. As we show, the accuracy of this estimation is of crucial importance for the market operation. We introduce parametric utility functions to capture the fact that each user evaluates the inconvenience/-cost tradeoff in a different way and that different types of demand (appliances) exhibit different price elasticity. The smart meter negotiates the following day's schedule with the operator and proposes a schedule to the user. In the end, it is the user that decides whether to follow the proposed schedule partially or totally.

An important building block of any DR scheme is also the strategy of the price setting entity, namely the utility operator. Most currently applied pricing schemes are either based on fixed time-of-use rates, and hence are not sufficiently adaptive, or simply expose users to the wholesale market prices, causing thus extreme price volatility and uncertainty. Instead, we propose a negotiation mech-

anism that takes place in advance. Hence, the home users are not only aware of the exact pricing pattern for the following day but also of the exact cost of the proposed schedules.

In this chapter, we investigate the interaction between an operator, multiple competing generators and a set of price taking home users in a day–ahead market. Our contributions can be summarized as follows,

- we introduce a realistic electricity DR model that captures the diverse set of home appliances,

- we identify the conditions that turn optimal demand scheduling into an NP-Hard problem. We show that this is the case for most types of residential demand, in contrast to the simplifying assumption of totally splittable demands generally made,

- we propose a price–setting mechanism that converges within a small number of iterations and aligns the interests of the operator and the home users,

- we show that price discrimination can be used to improve the total utility of the home users,

- we quantify the impact of inaccurate estimation of user utilities on the performance of demand response.

The rest of the chapter is organized as follows. Section 2.2 provides an overview of the related work. In Section 2.3 we introduce our DR model and show that optimal demand scheduling is NP-hard. Section 2.4 describes the operation of a day–ahead market that coordinates demands through dynamic pricing. In Section 2.5 we perform a classification of home appliances and use demand traces to evaluate the benefits arising from the proposed DR scheme. Section 2.6 concludes our study.

2.2 DR Related Work

In a series of works the problem of demand response is mapped to the well investigated problem of selfish routing over parallel links, with each link corresponding to a timeslot. The authors of [22, 26] consider a set of users that schedule demands within a finite horizon of several slots. Each user receives some utility per unit of energy consumed and pays the corresponding electricity cost to the operator. The difference of the two is the net utility, that the user wants to maximize by controlling the amount of power allocated at each timeslot. The price setting entity on the other hand aims at maximizing the social welfare of the system. Under convexity assumptions, they derive convergent distributed algorithms based on auction mechanisms and dual decomposition methods respectively.

In a similar framework, a two time scale wholesale electricity market is considered in [11], where renewable sources provide electricity at zero cost, while additional power may be purchased either through a day-ahead market or in real-time. For the case of deterministic demands and social welfare maximization the optimal strategy is to set prices equal to the marginal cost. Under the assumption that consumers respond by adjusting their demand schedule, the system converges to the social optimum. A simplified market model is considered in [18], where indicative utility functions for different

types of appliances are proposed. In place of deterministic knowledge of prices, [20] proposes a prediction mechanism that uses historical pricing data.

All the above works [11, 18, 20, 22, 26] share the simplifying assumptions that power is a continuous control variable that can be arbitrarily allocated at the slots and the utility is a concave function of the consumed power. However, this would require appliances that operate at any power level. In this work we showed that the problem of optimal demand scheduling becomes NP-hard once a realistic model for the demands is introduced. Besides, the scenario of maximizing social welfare, though tractable since the optimal pricing comes directly from the fundamental theorem of welfare economics, it does not capture the main objective of the operator *i.e.,* a balanced consumption pattern throughout the day. To the best of our knowledge, [10] is the only work that adopts a model of shiftable demands similar to ours; it neglects though the reciprocal interaction of pricing and demand scheduling that is inherent in any DR scheme. Here, we consider also the DR benefits at the user side, we introduce price discrimination as a means to improve the total utility of the users, and investigate the market operation under inaccurate estimation of the user utility functions.

Currently only a small percentage of households have enrolled into dynamic pricing programs, where prices are set according to the day-ahead wholesale market. However, as market penetration increases, the stability of electricity markets becomes doubtful. The authors of [25] show that making the cost minimizing individuals aware of the real-time prices may lead to an unstable closed loop feedback system and consequently to extreme price volatility. Thus, they proposed in [24] a pricing mechanism that charges electricity usage beyond the predicted one by some a posteriori calculated price. Such an approach cannot be easily applied in practice though, since home users require to know the price per unit of consumed energy in advance. Instead, we propose a negotiation-based mechanism that provides incentives to the end-users to enroll to DR programs, but also avoids instability.

In order to quantify the benefits of any DR mechanism a realistic model of the home energy consumption is required. In this direction, the authors of [9] derive a continuous time k-state Markovian model for the total consumption of a house based on real life measurements. They derived 12 reference models, each corresponding to a different period of the day and a different type of house. A detailed model that generates per appliance daily consumption traces was derived also in [23], based on a survey conducted in the United Kingdom in 2000. The model was successfully tested against actual consumption data from 22 houses, yielding close results. We extend this model by incorporating parametric utility functions that capture the elasticity of different appliances.

2.3 DR Model and Problem Formulation

We consider an electricity market consisting of a set of home users and a single operator. The operator applies dynamic pricing by dividing the day ahead into T equal time slots. The timeslot duration determines the market granularity. For example, a coarse-grained market would consist of $T = 24$ hourly intervals, whereas in a fine-grained one the price could change even in a minute timescale. The goal of the operator is to minimize electricity generation cost by balancing electricity consumption
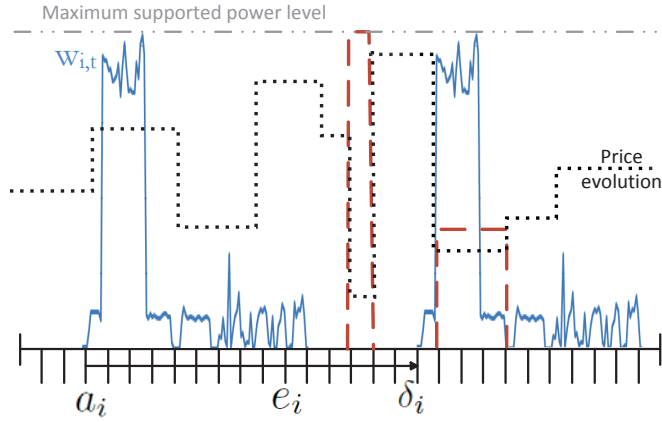
Figure 2.1: Optimal washing machine schedule for a cost minimizing user under the shiftable (solid blue) and the splittable (dashed red) models

throughout the day, while the users attempt to maximize their utility.

2.3.1 Motivation behind the DR model of shiftable demands

Initially, we motivate why a new, realistic model for demand response is required. Most home appliances are characterized by a specific consumption pattern, consisting of cycles of operation that have to be executed with a specific order and without any interruption. For example the solid blue curve of Fig. 2.1 corresponds to a typical consumption pattern of a washing machine. In this work, we propose that *most home demands are of shiftable nature i.e.,*users respond to dynamic prices by shifting their demands to an earlier or a later time. In contrast, most works [18, 20, 22, 26] in the field of demand response are based on the simplifying assumption of splittable demands, where electricity consumption per timeslot is the control variable. If this were true, at each timeslot we would be able to operate the washing machine at any power level. Such a model of splittable demands violates two basic constraints though; first, that most appliances do not support a continuous range of operating power levels and, second, that they cannot be preempted. These are hard constraints imposed by the operation of the appliance and hence any DR schedule that violates them is infeasible.

Although we showed that the splittable assumption leads to infeasible schedules, one may claim that it provides a close approximation of the actual case of shiftable demands. We depict in Fig. 2.1 the best response of a cost minimizing user for her washing machine under both models. The dotted line represents the evolution of price throughout the day. In the shiftable case the cost minimizing shift is selected, leading to the rightmost solid blue demand curve. The splittable approach is depicted with the dashed red line, where the appliance is activated within the lowest price timeslots. Although the total energy consumption is the same, the resulting demand pattern differs significantly. Concluding, the splittable approximation is neither realistic nor accurate, at least for the user side.

11

2.3.2 Scheduling of residential demands under dynamic pricing

We consider a set $\mathcal{N}$ of N residential users issuing a set of demands $\mathcal{D}$ for the following day. Each demand would be ideally (under flat pricing) scheduled according to most convenient schedule. Hence, demand $i \in \mathcal{D}$ is characterized by a power consumption profile, represented as a sequence of power levels $w_{i,t} \geq 0$ and a total power requirement of $W_i = \sum_{t=1}^{T} w_{i,t}$. Such a profile can be constructed either by explicit user input or automated measurement-based approaches [10]. Dynamic pricing motivates users to move demand out of peak consumption periods. We categorize demands into shiftable that cannot be interrupted and splittable ones that support a continuous range of operating levels up to a maximum of W_{max}.

For shiftable demand i, we denote with $a_i \in \{1, 2, \ldots, T\}$ the most convenient starting time and e_i the corresponding ending timeslot. We assume that only specific shifts are feasible, *e.g.*, the dishes have to be washed by dinner time. Let $\mathcal{F}_i \subseteq [-a_i + 1, T - e_i]$ denote the set of integer feasible shifts δ_i for demand i. Each demand individually responds to price vector $\boldsymbol{p} = \{p_t\}_{1 \leq t \leq T}$ with the shift $\delta_i^* \in \mathcal{F}_i$ that maximizes net utility, namely a function of the amount of shift and the electricity cost:

Response of shiftable demand i (max net payoff):

$$\max_{\delta_i \in \mathcal{F}_i} (1 - \theta_i) U_i(\delta_i) - \theta_i \sum_{t=a_i}^{e_i} w_{i,t} p_{t+\delta_i} \tag{2.1}$$

The first term, $U_i(\delta_i)$, captures the satisfaction received from the timely execution of demand i and is generally decreasing in $|\delta_i|$, since the larger the shift the higher the inconvenience caused to the user. The second term corresponds to the electricity cost. The *net* utility of timely execution less the associated electricity cost is represented by the difference of these two terms. The trade-off between them is modeled by the parameter θ_i, which depends on both user behavior and type of demand. A high value of θ represents either a user that mainly cares about minimizing cost as opposed to maximizing her utility (convenience) or shift-insensitive types of demand, such as the laundry machine, that yield the same utility irrespective of the time of day. As a result, parameter θ_i is indicative of the elasticity of each demand. In this work, we introduce also feasibility related parameter $\beta_i \in [0, 1]$ and define the feasible shift space as the integer subset $\mathcal{F}_i \subseteq \beta_i \times [-a_i + 1, T - e_i]$. A low value of β captures demands of tight activation constraints.

For each splittable demand i, we have to find the optimal portion of demand $x_{i,t}$ of timeslot t, *i.e.*,:

Response of splittable demand i (max net payoff):

$$\underset{\{x_{i,t}:1 \leq t \leq T\}}{\text{minimize}} \quad (1 - \theta_i) \sum_{t=1}^{T} V_{it}(x_{i,t}) + \theta_i \sum_{t=1}^{T} x_{i,t} p_t \tag{2.2}$$

$$\text{s.t.} \quad 0 \leq x_{i,t} \leq W_{max} \quad \forall t \in \{1, \ldots T\}, \tag{2.3}$$

$$\sum_{t=1}^{T} x_{i,t} = W_i, \tag{2.4}$$

where the *disutility* function $V(\cdot)$ captures the dissatisfaction caused to the user whenever modifying

her most convenient schedule. In general, $V(\cdot)$ is a convex function of demand since the differential dissatisfaction of a user increases as the distance from the most convenient schedule increases. Indicatively, we may use the absolute distance from the ideal consumption level of timeslot t $x_{i,t}^0$ to capture dissatisfaction, *i.e.,*:

$$V_{it}(x_{it}) = (x_{i,t} - x_{i,t}^0)^2 \qquad (2.5)$$

2.3.3 The role of the utility operator as the price setting entity

The operator aims at minimizing the electricity generation cost, which for each timeslot t is an increasing and convex function $c_t : \mathbb{R}^+ \to \mathbb{R}^+$ of the total demand [18]. Its exact form is determined by the operating cost of the power plants that have to be activated to meet demand. We perform an estimation of the generation cost as a function of the demand in Section 2.5 based on publicly available price/demand data.

Although the operator has no control over the demands of the individuals, it affects indirectly their distribution within a day by applying dynamic pricing. If p_{flat} denotes the price of the default flat pricing scheme, each day the operator has to solve the following optimization problem:

Operator's problem (min operational cost):

$$\min_{\boldsymbol{p}} \quad \sum_{t=1}^{T} c_t\left(\chi_t(\boldsymbol{p})\right) \qquad (2.6)$$

$$\text{s.t.} \quad 0 \le \bar{p} \le p_{flat} \qquad \forall i \in \mathcal{D},$$

where χ_t is the total demand at timeslot t resulting from price vector $\boldsymbol{p}$. The constraint implies that the average price $\bar{p}$ of any dynamic pricing program has to be at most equal to that of flat pricing, so as to attract users to enroll. In addition, this constraint ensures a fair comparison setting for different dynamic pricing schemes.

Without loss of generality we consider an average price bound normalized to unity ($p_{flat} = 1$), since we are only interested in the distribution of the prices throughout the day and not on their exact values.

Remark 2.3.1. *Although a user's demand scheduling affects the pricing strategy of the operator, its impact on the resulting prices is negligible, since each residential user controls only a tiny percentage of the total demand. This justifies our assumption of price taking users, who simply respond to the announced prices.*

2.3.4 What makes optimal demand scheduling NP-Hard?

Several works derive optimal polynomial algorithms for demand scheduling under certain assumptions. Here, we identify the characteristics of home appliances that turn cost minimization under dynamic pricing into NP-hard. Consider the simplified version of our problem where the operator has direct control over demands, *i.e.,* it can shift each demand at will. For rectangular demands of

arbitrary duration but equal power requirements, this is an NP-hard bin packing instance [16]. Here, we generalize to demands of arbitrary pattern and consider the impact of pricing on complexity.

Consider demands of single slot duration, but arbitrary power requirements. Without loss of generality, we can relax the feasibility constraint $\delta_i \in \mathcal{F}_i$ by assuming that each demand can be scheduled throughout the horizon T. This can be directly mapped to the multiprocessor scheduling problem (a.k.a. minimum makespan scheduling), where a set of jobs have to be scheduled on a set of identical machines, so as to minimize the maximum load, *i.e.,* the sum of the processing times of jobs assigned to each machine. This has been shown in [13] to be NP-Hard (by a partition reduction).

In our case, the T identical machines correspond to the T timeslots and the jobs are the demands of different power requirements. Our objective is the minimization of generation cost, which is a convex and increasing function of the total demand within a timeslot. Thus, the optimal scheduling is the one that minimizes the maximum power consumption within a timeslot. This equivalence proves the NP-Hardness of our problem for demands that can be preempted arbitrarily, leading to demand quanta of single slot duration.

We showed that scheduling single slot demands of arbitrary power or arbitrarily long demands of equal power [16] is NP-Hard. Next, we consider the special case of arbitrarily shaped demands that can be split in both dimensions, time and power. If all the demands support the same range of power levels (discrete or continuous) and preemptive scheduling, the problem can be mapped to the load balancing one of [15], which for convex cost functions can be solved optimally in polynomial time. Thus, we deduce that polynomial complexity holds only for demands that can be split across both time and power. Otherwise, the non-splittable nature of demands makes demand scheduling NP-hard.

Finally, we introduce pricing. If the utility function of each user is invertible and known to the operator we have:

$$\exists \boldsymbol{p}_i : \delta_i^* = k, \forall k \in \mathcal{F}_i \tag{2.7}$$

In this case, by setting a different price to each user, the problem becomes equivalent to controlling the demands directly. Thus, even under complete information optimal pricing is NP-Hard. Since the utility functions is private information that is not disclosed, we propose a DR scheme that is based only on the aggregate demand pattern.

2.4 A Novel Electricity Market for Efficient Demand Response

In this section we investigate how an automated DR mechanism can be incorporated into the electricity market. We assume that each user is price taking and autonomously selects her strategy. We initially focus on the usual scenario of nondiscriminatory pricing that guarantees fairness across the users. However, later we relax this assumption and show that properly designed price discrimination can improve the utility of the system.
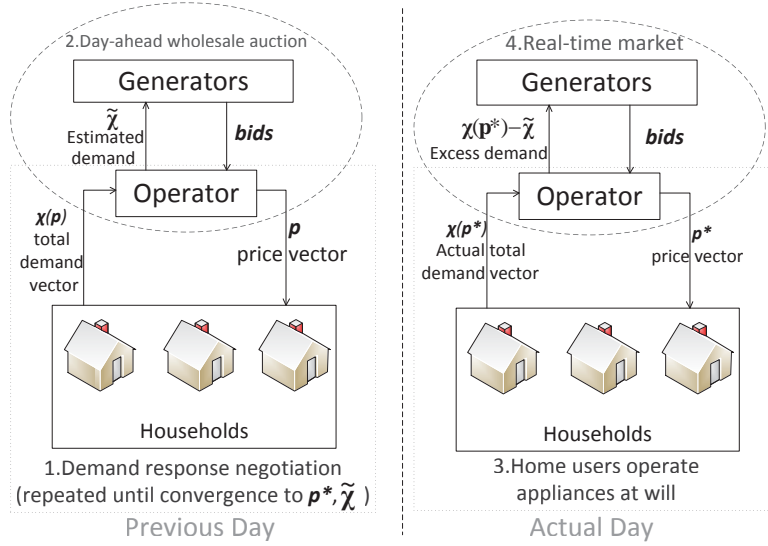
Figure 2.2: The operation of the proposed market

2.4.1 Operation of the proposed market

The envisioned market consists of an operator, multiple competing generators and a set of end-users that have enrolled to a dynamic pricing program. It takes place within two consecutive days, the previous and the actual day, and its operation is depicted in Fig. 2.2. Initially, the operator deduces the ideal daily consumption pattern of each household, by announcing a flat price for the following day. The smart meter of each home, based on locally available historical data, estimates the demands to be scheduled for the next day and communicates its total demand pattern to the operator. The day ahead estimation of demand is beyond the scope of this work. The interested reader may refer to [10].

Next, a negotiation phase starts, where the operator calculates the estimated aggregate demand of all the homes, updates its pricing strategy by solving (2.6) and announces the new day-ahead prices. Each smart meter in turn responds by calculating the optimal demand schedule to the announced prices through (2.1). For this purpose the smart meter has to be aware of the utility function of each demand. This could be provided either directly through user input during the installation phase or through a training process. The negotiation phase takes place before the actual day and is repeated until convergence.

In the end, the operator has at its disposal an estimation of the total demand for the following day. Based on this day ahead prediction a wholesale auction is conducted, where a number of competing generators offer their electricity output. Each generator makes a bid indicating the amount of energy that it can provide at a given time period and the corresponding price. The best offers are accepted to meet the predicted demand. Generally, the market clearing price is the one of the highest bid accepted, in an attempt to guarantee fairness among the generators.

In the house premises, the smart meter proposes the optimal schedule to the user for the actual day. It is at her discretion though, whether to follow it totally, partially or even not at all. Generally, this depends on the accuracy of the estimation of the day ahead demands and of the utility of the user. A perfect prediction would lead to an equilibrium where all the users would follow the proposed schedule exactly, since any deviation from it would reduce their utility. In a real system though, due to imperfect prediction of demand, a deficit of energy may appear within specific timeslots of the actual day. This deficit has to be covered from the real-time market, generally at a higher cost. In Section 2.5, we quantify the impact of the prediction accuracy on the market operation.

2.4.2 Day-ahead DR negotiation mechanism

The operator applies day–ahead dynamic pricing to minimize electricity generation cost, which for a given timeslot is a convex and increasing function of the total demand. If we assume that the same cost function holds for each timeslot, the optimal price vector is the one that leads to the most balanced demand for the given constraints of the users. Ideally, the total demand would be $\chi_t^* = \mu = \frac{1}{T} \sum_{i \in \mathcal{D}} W_i$ at every timeslot.

Obviously, χ_t is a decreasing function of the corresponding price p_t, but increasing in the price of any other slot. Thus, χ_t is a function of the whole vector $\boldsymbol{p}$, and the exact dependence is determined by the utility functions of the users. Since the operator is unaware of the utility functions, it cannot deduce the exact impact of a price increase/decrease on the demand distribution. Such information can only be deduced by observing the user responses to the announced prices. For this purpose, we propose the day–ahead negotiation phase of Algorithm 1. For notational simplicity we have discarded the function arguments.

Generally, when the gradient of the objective function is unknown or cannot be calculated, randomized pattern search methods are used [17]. Such methods perform steps of random size in random search directions in an attempt to find a local optimum. In contrast, we follow a deterministic search direction and adjust only the step size. The strategy of the operator is to increase price at the peak demand periods and reduce it at the off-peak periods. Thus, although the actual gradient is not available, we use the total demand vector $\boldsymbol{\chi}$ as an estimator of the gradient.

The operator orders slots according to their total demand (L5,L6) and decreases price by an adjustable step size ε at the lowest demand slots (L8), while price is equally increased in the highest ones (L9). L4 ensures that prices do not become negative. However, by discarding L4 and the last condition of L7 we may enable this option. L11-L13 correspond to the best response strategy of the users. Each user responds by adjusting demand through (2.1), which for arbitrary prices and demand patterns requires a search over the feasible horizon and through (2.2) for its splittable demands which is of waterfilling nature.

Operator's strategy extends to lines L14-L18. In order to ensure convergence, we apply a pattern search like approach for the adaptation of step size. A price update is adopted (L15) only if it leads to a reduction of the objective function, *i.e.,*the generation cost. Otherwise, the algorithm backtracks and the step size ε is reduced (L17). Obviously, the rate of convergence depends on the step size related

16

Algorithm 1 The negotiation phase of our electricity market

Input: ε: step size, $q < 1$: step adaptation parameter

1: $\mu \leftarrow \frac{1}{T} \sum_{i \in \mathcal{D}} W_i$ // average consumption per slot
2: $\boldsymbol{p} \leftarrow \boldsymbol{1}$ // start from flat pricing
3: **repeat** // iteration k
4:      $\boldsymbol{r} \leftarrow \min\{\varepsilon, \boldsymbol{p}^{(k)}\}$ // the operator updates prices
5:      $\mathcal{G}_{min} = \{\text{timeslot indices ordered in increasing } \chi_t^{(k-1)}\}$
6:      $\mathcal{G}_{max} = \{\text{timeslot indices ordered in decreasing } \chi_t^{(k-1)}\}$
7:      **for all** $t \in \mathcal{G}_{min} \cap \{t : \chi_t^{(k-1)} < \mu\} \cap \{t : p_t^{(k-1)} > 0\}$ **do**
8:          $y_t \leftarrow p_t^{(k)} - r_t$
9:          $y_{\tilde{t}} \leftarrow p_t^{(k)} + r_t$ // $\tilde{t}$ the corresponding index of $\mathcal{G}_{max}$
10:     **end for**
11:     **for all** $i \in \mathcal{D}$ **do** // each demand responds
12:        calculate $\delta_i^{(k)}$ from (2.1) or $\{x_{i,t}\}$ from (2.2)
13:     **end for**
14:     **if** $\Delta c \leftarrow c^{(k)} - c^{(k-1)} < 0$ **then** // adopt solution
15:        $\boldsymbol{p}^k \leftarrow \boldsymbol{y}$
16:     **else** // decrease step size
17:        $\varepsilon \leftarrow q\varepsilon$
18:     **end if**
19: **until** $\Delta c \simeq 0$ // convergence check

parameters q and ε. Small enough values of q reduce the number of iterations required for convergence, but may lead to inefficient local minima. Nevertheless, our extensive simulations indicate that generally less than 50 iterations are required.

This approach requires no knowledge of market parameters, such as the utility functions and the types of demands, that may be either private information or of stochastic nature. All the decision making is performed based only on the total demand pattern, information that is communicated to the operator during the negotiation phase. Note that in case of time varying generation cost, the resulting generation cost vector should serve as the gradient instead.

2.4.3 Discriminatory pricing as a means of improving users' utility

Price discrimination has been used extensively by companies as a means of maximizing revenue [27]. In the smart grid scenario, differentiated prices enable the operator to exploit the elasticity of demands of the users more efficiently. In this work, we investigate how price discrimination can be used to improve the utility of the system as a whole. Generally, in order to fully exploit the potential of price discrimination, the operator has to be aware of the elasticity of each user, which is private information though. However, the operator can estimate elasticity by observing the response of each user to the announced prices during the negotiation phase.

Note that updating the price according to L7-L9 of Algorithm 1 for a user that is irresponsive to price changes leads to a decrease of the utility without offering any improvement to the generation cost. Thus, we propose that once the operator detects that $\boldsymbol{h}_j^{(k)} = \{h_{j,t}^{(k)} : 1 \leq t \leq T\}$,the aggregate demand of household j, remains unaltered at iteration k, it should be removed from negotiation.

17

In order to identify whether the demand of a household has changed in iteration k, we define the similarity index $I_j^{(k)}$ given by:

$$I_j^{(k)} = \frac{\boldsymbol{h}_j^{(k)} \cdot \boldsymbol{h}_j^{(k-1)}}{\left\|\boldsymbol{h}_j^{(k)}\right\| \left\|\boldsymbol{h}_j^{(k-1)}\right\|}, \tag{2.8}$$

where $< \cdot >$ is the inner product operator and $\left\|\boldsymbol{h}_j^{(k)}\right\| = \sqrt{\sum_{t=1}^{T} (h_{j,t}^{(k)})^2}$ is the ℓ^2-norm. The resulting similarity index ranges from -1 for opposite vectors up to 1 for aligned ones. We also define similarity threshold $\gamma > 0$, a design parameter that determines the aggressiveness of price discrimination. The conservative approach of a high value of γ maximizes the number of users participating in negotiations and hence leads to limited discrimination. By replacing L15 of Algorithm 1 with Algorithm 2 we derive a price discriminatory strategy.

Algorithm 2 Discriminatory pricing strategy

Input: γ: similarity threshold
1: **for all** $j \in \mathcal{N}$ **do** // user selection
2:     **if** $I_j^{(k)} < \gamma$ **then**
3:         $\boldsymbol{p}_j^{(k)} \leftarrow \boldsymbol{y}$
4:     **else**
5:         discard user j from negotiation
6:     **end if**
7: **end for**

Notice that due to price discrimination, the price for each user j at iteration k is different and is denoted by $\boldsymbol{p}_j^{(k)}$. In addition, since the search step size ε is decreasing, such a discriminatory strategy does not harm the induced generation cost significantly.

2.4.4 A lower bound of the operator's cost

For comparison purposes we consider the relaxed version of the original problem, where scheduling of all demands is performed directly by the operator under the assumption of splittable demands. This can be formally expressed as:

$$\min_{\{x_{i,t}\}} \quad \sum_{t=1}^{T} c_t \left(\sum_{i \in \mathcal{D}} x_{i,t} \right) \tag{2.9}$$

$$\text{s.t.} \quad x_{i,t} \geq 0 \qquad \forall i \in \mathcal{D}, t \in \{1, \dots T\}$$

$$\sum_{t=a_i+\delta_i}^{e_i+\delta_i} x_{i,t} = W_i \qquad \forall i \in \mathcal{D}, \delta_i \in \mathcal{F}_i$$

Since the cost function is convex, this is equivalent to the constrained load balancing problem of [15], which can be solved in polynomial time. In particular, it leads to the most balanced assignment and provides a lower bound of the actual generation cost. Notice that in practice due to the non-splittable nature of demands this bound is generally not achievable.

(a) Day–ahead prices as a function of demand   (b) Real-time prices as a function of demand
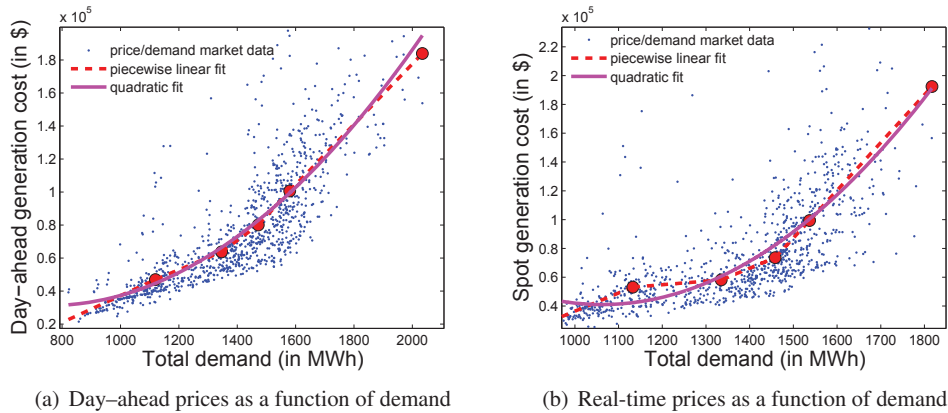
Figure 2.3: Derivation of a generation cost function for DA and Real-time market

Notice that in practice due to the non-splittable nature of demands this bound is generally not achievable. Besides, this centralized approach actually assumes that all the demand scheduling is performed directly by the operator without considering the utilities of the users. Since this simplifying assumption has been extensively used in the field of DR scheduling, we will use it as a performance benchmark.

2.5 Numerical Results

The main performance metrics of any DR mechanism are the total utility of the end users and the electricity generation cost. In order to estimate the latter, we use historical demand/price data from the New England market. Next, we build a realistic DR dataset based on the per appliance demand traces of [23] to quantify performance of proposed schemes. Since different appliances exhibit different elasticity, we classify household appliances according to their DR behavior and show that most support only two power states, ON and OFF, and cannot be interrupted, hence fall within the class of shiftable demands.

2.5.1 Estimation of generation cost

In order to quantify the benefits of demand response we need a representative function for the electricity generation cost, which has to be a convex and increasing function of the total demand at each timeslot. For this purpose, we perform curve fitting on historical price and the corresponding total demand data from the day–ahead electricity market of New England [12] by using piecewise linear functions and quadratic polynomials. In Fig. 2.3(a) we depict the data used and the derived cost functions. In order to derive an expression of the electricity cost at the real-time market, we perform the same procedure on the corresponding real-time price/demand data (Fig. 2.3(b)). In the following evaluations, we use the derived quadratic expressions to estimate the generation cost for different

Table 2.1: DR related characteristics of home appliances

Appliance	Preemptive	Power levels	Elastic	θ	β	Power(W)
Refrigerator	✓	$2 - \infty$	✓	0.8	0.1	300
Hi-Fi, PC, TV	X	2	X	0.1	0.1	100, 150, 150
Iron,Vacuum cleaner	X	2	X	1	1	1000, 2000
Printer	X	2	X	0.5	1	300
Hob, Oven	X	2	X	1	1	2200, 2000
Microwave, Kettle	X	2	X	1	0.1	1440, 2000
Dishwasher, Washing machine	X	2	X	1	1	1200, 600
Tumble dryer, Washer dryer	X	2	X	1	1	2800, 800
Water heating	X	2	X	0.3	1	4000
Electric space heating	✓	$2 - \infty$	✓	1	-	2000
Lights	X	$2 - \infty$	✓	0.4	-	400

values of demand:

$$c_{DA} = 0.066\chi^2 - 65\chi + 34967 \tag{2.10}$$

$$c_{RA} = 0.266\chi^2 - 567\chi + 343645 \tag{2.11}$$

Notice that, as expected, the cost per MWh is significantly higher in the spot market.

2.5.2 DR behaviour of home appliances

Each household contains numerous appliances of diverse characteristics. We perform a classification of the demands based on three basic attributes related to demand response:

1. **Preemptive vs. Non-preemptive scheduling**: The former can be stalled arbitrarily many times, whereas the latter can be scheduled only in consecutive timeslots.

2. **Multiple vs. Single operating power level**: Some appliances support multiple power levels of operation, while others only two states, ON and OFF. A small number of appliances support a continuous power range, but these are not broadly available at home.

3. **Elastic vs. Inelastic total energy requirement**: Some appliances may allow for partial fulfillment of their energy requirements. Our model can easily capture this case by using utility functions that also depend on the amount of energy consumed.

In this work we consider the set of appliances of [23]. We classify them in the aforementioned categories in Table 2.1. Additionally, in the last three columns we depict the ranges for the elasticity related parameters θ and β that we used in our simulations and the typical power consumption for each of the devices. Parameter θ captures the timely–execution/cost–saving tradeoff, whereas β is indicative of how loose the deadline of each appliance is. The depicted values are the relative maximum elasticity considered for each type of appliance. Thus, in order to capture the diverse behavioral profile of different users, we perform uniform sampling in the range $[0, \theta_i]$ and $[0, \beta_i]$.

Notice that the vast majority of home appliances is of shiftable nature. Only two types can be preempted, and even these generally support a limited number of power levels. Only heating appliances operating on a continuous power range became available to home users. In order to capture this ongoing change, we assume that all houses are equipped with heating/cooling appliances of inverter-type. The rest of the appliances are considered shiftable.

Next, we quantify the performance of the proposed market and show that the widely used splittable model leads to an overestimation of the benefits of demand response. For a fair comparison with the flat pricing scenario, we do not consider any tasks of elastic total energy requirement. Thus, independently of the pricing strategy applied, the total consumption within a day remains the same.

2.5.3 Quantifying the benefits of demand response

In order to evaluate the performance of the proposed DR scheme, we consider also the proportional pricing scheme of [21, 19], where day-ahead prices are set proportionally to the total demand. For comparison purposes, in each figure we depict also the derived splittable lower bound that respects the deadlines though.

In Fig. 2.4 we depict the distribution of total demand throughout a day for a system consisting of 200 households. In Fig. 2.4(a) we depict the case of flat pricing. Next, we demonstrate in Fig. 2.4(b) that applying proportional pricing smooths demand distribution slightly. Fig. 2.4(c) depictsthe demand pattern resulting from our scheme and Fig. 2.4(d) corresponds to the centralized relaxation of splittable demands. We notice that our approach leads to a much more balanced demand. The small spikes observed come from the non-splittable nature of demands, and hence are not apparent in the relaxed approximation.

(a) Flat pricing   (b) Proportional pricing   (c) Proposed pricing   (d) Lower bound
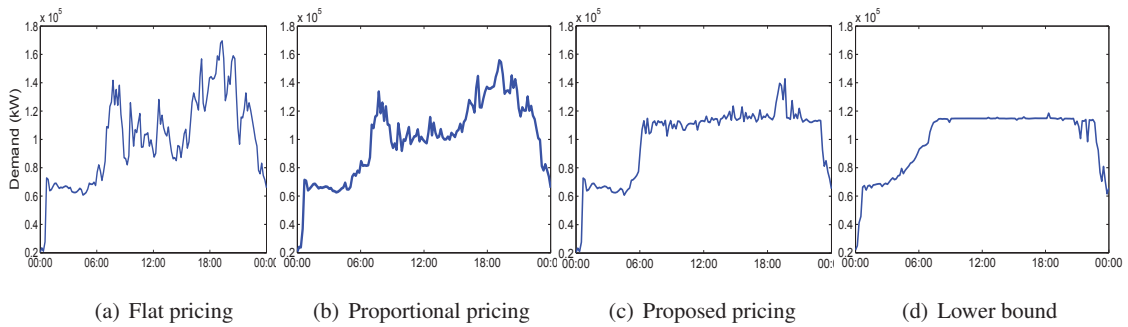
Figure 2.4: Demand distribution throughout the day

In an attempt to quantify DR benefits, we depict in Fig. 2.5 the resulting generation cost for different values of demand elasticity parameters β and θ. Fig. 2.5(a) depicts the impact of the deadline–related parameter β on the generation cost for an average value of θ of 0.5 showing that a small deadline horizon (captured by small values of β) is sufficient to guarantee most of the DR gains. We notice that our scheme outperforms the proportional one significantly. Fig. 2.5(b) quantifies the impact of balancing parameter θ on the generation cost. Our scheme approaches the lower bound

(a) Horizon parameter uniformly distributed in $(0, \beta)$

(b) Utility / payment tradeoff parameter uniformly distributed in $(0, \theta)$
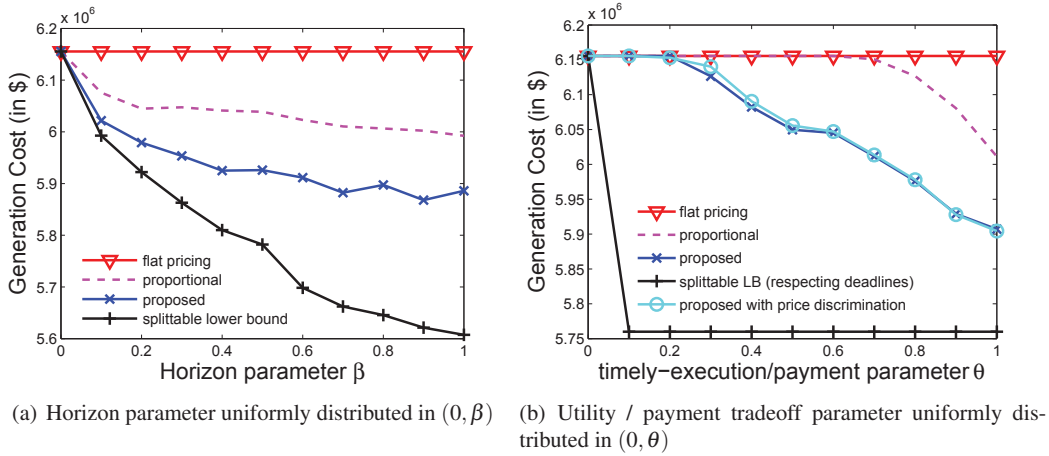
Figure 2.5: Generation cost for different system settings
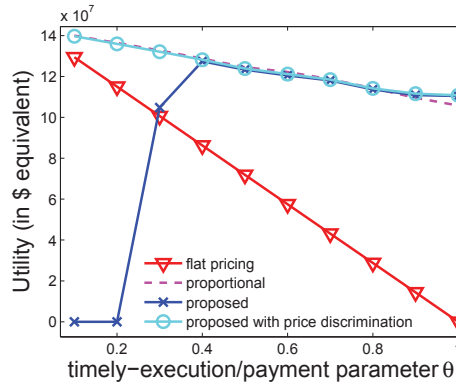
Figure 2.6: System utility for different values of parameter θ

as the elasticity parameter θ increases, whereas the proportional scheme requires significantly more elasticity in order to achieve some cost reduction.

Until now we concentrated on the benefits of the operator, which due to its leading role in the market may impose its cost minimizing strategy. However, the transition to a new status requires that all the involved entities derive some gain. Otherwise the end users would have no incentive to enroll to a dynamic pricing program. Thus, we focus on the impact of the proposed market on the utility of the end-users. In Fig. 2.6 depicts the total utility achieved for different values of θ. The depicted values have been normalized, such that the utility of a cost minimizing user under flat pricing is zero. Whereas all the dynamic pricing schemes generally improve the utility of the users, the proposed scheme causes a degradation of the utility in the low elasticity regime. This comes from the fact that we apply the same price vector to all the users, and hence inelastic ones will experience increased prices and consequently decreased utility, and motivates us to charge different prices to different users.

22

In this direction we investigate also the benefits of the proposed discriminatory scheme. In Fig. 2.6 we observe that our price discriminatory strategy improves the system utility significantly for the cases of low elasticity, where are original approaches tend to overcharge users without deriving any significant benefits. But what is the impact on generation cost. From Fig. 2.5(b) we deduce that the modified scheme performs identically in terms of generation cost and hence the improvement at the user side comes free of charge.

(a) Generation cost
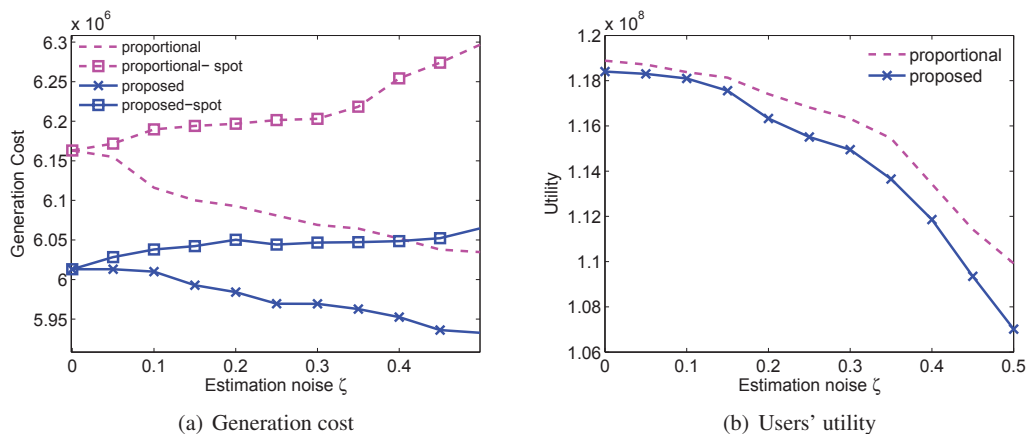
(b) Users' utility

Figure 2.7: Impact on performance of estimation noise uniformly distributed in $(-\zeta, \zeta)$

Until now we assumed that the estimation of the user's utility is perfectly accurate and consequently that the user follows exactly the announced demand schedule. In practice though, the utility function is estimated based on historical data of the user behaviour and hence can not be perfect. In order to capture this phenomenon, we assume that the negotiation takes place according to a noisy estimate $\hat{\theta} = \theta + \zeta$ of the actual elasticity parameter θ and demonstrate resulting performance in Fig. 2.7. Under the assumption that the operator can buy the additional electricity required at the day–ahead price, we show in Fig. 2.7(a) that the impact of noise ζ on generation cost is negligible. However, in reality this additional energy has to be purchased through the real-time (spot) market at a significantly higher price, leading to increased generation cost. As we observe in Fig. 2.7(b), estimation error causes also a significant decrease in the utility of the home users. Our results indicate the importance of accurate utility estimation and that the significance of users following the aggreed schedule.

2.6   Conclusions

Our analysis provides a better understanding of DR schemes and their role on the ongoing electricity market restructuring. We investigated the price setting strategy of the operator and the best response of home users. By using realistic demand traces, we showed that, compared to existing DR schemes, the proposed one exploits more efficiently the DR benefits and leads to a win–win situation for all involved entities.

Given that not all users are similarly responsive to dynamic prices, we showed how price discrimination can improve total user utility. Finally, we quantified the impact of inaccurate prediction of user utility on the operation of the market. Interestingly, the erroneous estimation of utilities has a negligible impact on generation cost, but leads to significantly reduced user utility. Notice that although we mainly focused on the residential sector, most of the derived results hold in general.

Chapter 3

Energy Efficiency in the Smart Grid - Part II: The Role of Aggregators in Demand Response Markets

In the previous chapter, we outlined the importance of further penetration of DR schemes in the residential sector. The design of efficient Demand Response (DR) mechanisms for the residential sector though entails significant challenges, mainly due to the large number of home users and the negligible impact of each of them on the market. To counteract this, we introduce in this chapter a hierarchical market model for the smart grid where a set of competing aggregators act as intermediaries between the utility operator and the home users. The operator seeks to minimize the smart grid operational cost and offers rewards to aggregators toward this goal. Profit-maximizing aggregators compete to sell DR services to the operator and provide compensation to end-users in order to modify their preferable consumption pattern. Finally, end-users seek to optimize the tradeoff between earnings received from the aggregator and discomfort from having to modify their pattern.

Based on this market model, we first address the benchmark scenario of a cost-minimizing operator that has full information about user demands. Then, we consider a DR market, where all entities are self-interested and non-cooperative. The proposed market scheme captures the diverse objectives of the involved entities and, compared to flat pricing, guarantees significant benefits for each. Using realistic demand traces, we quantify the arising DR benefits. Interestingly, users that are extremely willing to modify their consumption pattern do not derive maximum benefit.

3.1 Introduction to hierarchical DR markets

Recent advances in smart metering technology enable bidirectional communication between the utility operator and the end-users and facilitate the option of dynamic load adaptation. Toward this direction, demand-response (DR) programs provide incentives to major consumers of electricity, usually in the form of monetary rewards, to reduce their electricity consumption in peak-demand periods. DR

can take place at a very fast timescale, almost real-time, it leads to a more stable power grid system and it significantly reduces electricity generation cost and CO_2 emissions [8].

Although DR has been successfully applied in the industry sector [28], its application in the residential sector is a more challenging task. First, if the existing DR schemes were applied in the residential sector, the operator would derive most of the DR benefits for itself, since each individual amounts to a small portion of the total demand and hence has limited negotiation power. Second, the sheer number of home users introduces scalability issues. Third, the utility operator in general lacks the know-how of designing and applying DR mechanisms at such a large scale.

Aggregators are new entities in the electricity market that act as mediators / brokers between users and the utility operator. Aggregators possess the technology to perform DR and are responsible for the installation of the communication and control devices (i.e. smart meters) at end-user premises. Since each aggregator represents a significant amount of total demand in the DR market, it can negotiate on behalf of the home users with the operator more efficiently. The current role of aggregators amounts to paying a monthly fee to the contracted end-users (mainly industrial ones) in order to gain direct control of their appliances [28, 29]. Thus, in case of a peak-demand emergency they can turn off the energy-intensive appliances of the users, e.g. air-conditioning, for a short period.

In this work, we devise a hierarchical DR market model for the residential sector that captures the interaction of home users, the utility operator, and several competing aggregators through a day-ahead DR market. Our model enables efficient demand response in the presence of self-interested entities and structures the DR market in three levels.

At the upper level, the utility operator provides monetary rewards to the aggregators for their DR services. Its objective is to minimize its own operational cost. At the middle level, the aggregators provide DR services to the operator by presenting a total demand profile that minimizes the cost of the operator to support it. The aggregators seek to achieve this demand profile by providing monetary incentives to home users to modify their demand pattern through a day-ahead market. The objective of each aggregator is to maximize its own net profit, namely the income received from the operator minus the compensation it provides to home users.

At the lower level, home users negotiate with the aggregators to receive monetary compensation in order to modify their consumption pattern. Users trade off their inconvenience that arises from deviation from their preferable or customary usage patterns for a lower electricity bill. Given the day-ahead compensation, the objective of each user is to determine a consumption pattern that strikes an optimal balance of this tradeoff by maximizing a net payoff function. We capture this tradeoff through an inelasticity parameter that is proportional to the inconvenience caused. The market operation that involves all three types of self-interested entities is summarized in Fig. 3.1.

The introduction of aggregators in the market introduces novel challenges on how the arising DR benefits span the entire chain of utility operator, aggregators and end-users. To the best of our knowledge very few recent works have considered the role of the aggregator in the future electricity market e.g. [30],[31], but do not focus on the scenario of residential demands. Our proposed hierarchical market and the corresponding DR mechanism guarantee that no deficit exists in the market and lead
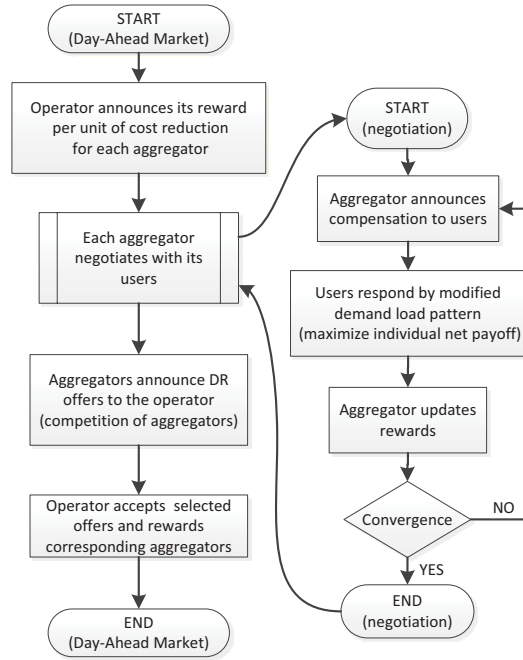
Figure 3.1: The operation of the proposed day ahead market.

to significant benefits for all market participants. Our contributions can be summarized as follows:

- We formulate the objectives of the utility operator, the competing aggregators and the home users in a hierarchical DR market.

- We characterize the operation for the benchmark scenario of a DR market where the operator has full information of all DR-related parameters.

- We devise a DR market scheme, where all entities execute non-cooperative strategies and show the interaction of the entities, when each acts according to its own objective.

- Using realistic demand traces, we show that our hierarchical DR scheme guarantees a portion of the DR benefits for each market entity. Thus, despite the dominant role of the operator, both the aggregators and the end-users can achieve significant financial benefits through negotiation. Interestingly, increasing the willingness of users to modify their consumption pattern (lower inelasticity) reduces the grid operational cost, but is not always beneficial for the aggregators and home users.

The rest of this chapter is organized as follows. Section 3.2 provides an overview of related work. In Section 3.3 we introduce the hierarchical DR model and formulate the objectives of the involved entities. Section 3.4 describes the benchmark scenario of a day–ahead market where all the decision making is performed by the operator under full information about the demands of end-users. In Section 3.5 we present how our model could be applied in a realistic hierarchical market and

investigate the required information exchange. In Section 3.6 we use daily demand traces to quantify the benefits arising from the proposed market structure. Section 3.7 concludes our study.

3.2 Related Work

Numerous works in the field of DR (e.g. [11, 22, 26]) assume a social-welfare maximizing operator that communicates and negotiates directly with utility-maximizing end-users. In this case, according to the fundamental theorem of welfare economics, the optimal strategy is to set prices equal to the marginal cost of supply. However, this does not hold for our scenario, where a self-interested operator seeks to minimize its operational cost. This has been considered lately in [32] that investigates the problem of demand scheduling from the operator's perspective. The authors devise a stochastic model for demand requests which arrive according to a Poisson process, they have exponentially distributed power requirements, and they need to be activated within a deadline. They derive a threshold-based demand scheduling policy that is asymptotically optimal as the deadline expiration rate goes to zero. In contrast to our work, [32] assumes that the operator has direct control over the demands and hence they do not consider incentives for end-users. The DR scenario of a cost-minimizing operator that incentivizes home users to shift their demands through dynamic pricing has been considered in [14].

In addition, several works cast the problem of demand response as a Walrasian auction, where prices are set so as to match supply and demand, and use tatonnement mechanisms for its solution. Indicatively, the work [34] considers price-taking residential customers that schedule their consumption throughout the day and a utility operator that sets prices to maximize social welfare. A similar tatonnement process has been proposed in [35] for the stochastic version of this problem.

All the works mentioned above, require extensive message exchange between the operator and the end-users. Hence, scalability issues may arise in a large scale deployment. Hierarchical market structures, where aggregators serve as DR intermediates, appear as a promising approach to deal with scalability issues. To the best of our knowledge only a few works have considered the role of the aggregator in the future electricity market. Recently, [36] proposed that the aggregator should coordinate electricity generation and demand response. Given a reference supply / demand imbalance signal or an indirect signal in the form of time varying market price, the aggregator has to solve a convex quadratic program for each time slot. The authors of [30] propose a simplified hierarchical market mechanism, where the operator sets a target demand curtailment and the aggregator provides compensation to the end-users to meet this target at minimum payment. Each consumer is a price-taker and bids its supply function in order to minimize its disutility. In this setting, two bidding mechanisms are proposed that converge to the optimum. Both these works do not consider the incentives that would enable such a mechanism and consequently do not capture the interaction of competing aggregators.

To the best of our knowledge [31] is the only work that considers the interaction of several aggregators, but for the microgrid scenario. A two stage market model is proposed, where in the first stage the price paid by the operator and the electricity provided by each aggregator is decided through a tatonnement process. For the lower stage a supply function bidding mechanism is proposed, where
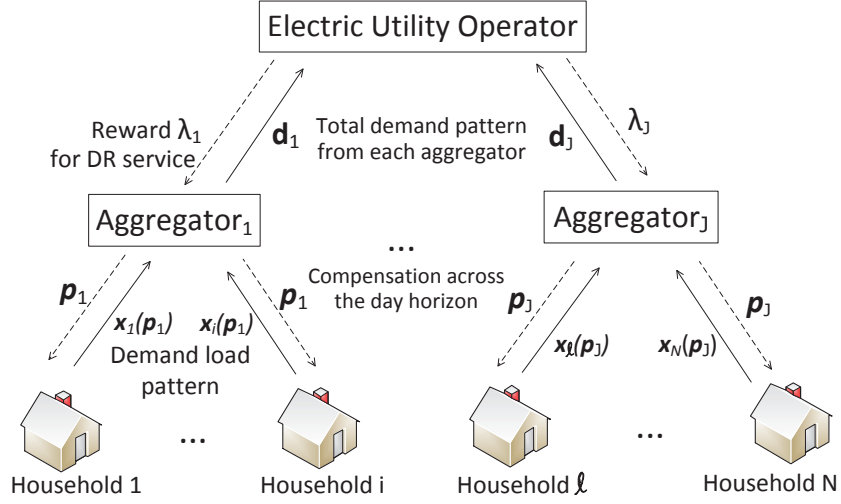
Figure 3.2: The hierarchical structure of the proposed day-ahead market: *i*) at the lower level the users modify their demand pattern according to the compensation advertised by the aggregators, *ii*) at the middle level the aggregators determine their compensation strategy so as to maximize their net profit, given the rewards offered to them by the operator, *iii*) at the upper level the operator computes the reward per unit of cost reduction for each aggregator so as to minimize its operational cost.

the microgrids bid their supply functions and the aggregator sets the price so as to maximize its profit. While [31] focuses on microgrid generation within a particular time slot, here we focus on the scheduling of residential demands across time.

The importance of DR in the residential sector was also quantified recently in [37], where it was shown that a slight extension of 10% in the total operation time of residential demands may reduce peak consumption by 125MW. Here, we consider similar framework capturing the discomfort caused to the users and the incentives required to achieve such DR benefits. In order to quantify the benefits of a residential DR scheme, a realistic model of the home energy consumption is required. For this purpose, we use the generator of per appliance consumption daily patterns of [23] and we extend it by incorporating demand elasticity parameters.

3.3 Hierarchical System Model and Problem Formulation

We consider an electricity market consisting of an electric utility operator, a set of aggregators $\mathcal{A} = \{1, 2, \ldots J\}$ and a set of residential users $\mathcal{N} = \{1, 2, \ldots N\}$, as the one depicted in Fig. 3.2. All the involved entities are self-interested and rational. Each user participates in the market through an a priori determined aggregator on a contractual basis, i.e., users do not move between aggregators. We focus on the day-ahead market and assume that the day is divided into T equal periods. We denote with $\mathcal{T} = \{1, 2, \ldots T\}$ the corresponding set of time slots. Throughout the text, we use the terms demand and load interchangeably.

3.3.1 The role of the utility operator

In a day-ahead market, the ensemble of users issues a cumulative demand of W Watts. Under the current practice of flat pricing, a fixed price q_f is charged per Watt of consumption, independently of the period of the day. Thus, within a day the operator derives a total income of $q_f W$ from the users payments.

In order to meet demand, the operator has either to activate costly powerplants or to purchase electricity from third parties. The cost of generating electricity power is generally assumed to vary with time, due to the time-varying availability of supply, e.g. from renewable sources. However, for any given time slot t cost is a strictly increasing and convex function $c_t(y_t)$ of the corresponding total load $y_t \geq 0$ [32].

The operator seeks to calculate the cumulative daily load profile vector $\mathbf{y} = \{y_t : t \in \mathcal{T}\}$ that maximizes its revenue. If the operator could directly control the loads, its objective would be formally expressed as:

$$\max_{\mathbf{y}} \quad q_f W - \sum_{t \in \mathcal{T}} c_t(y_t) \tag{3.1}$$

$$\text{s.t.} \quad \sum_{t \in \mathcal{T}} y_t = W.$$

Since the received income is fixed, (3.1) can be transformed into a convex cost minimization problem, which can be easily solved through convex optimization. Indicatively, for time-invariant cost functions, the optimal solution is a perfectly balanced load across the day.

Since the operator has no means to exercise direct control over the user demands, alternative means of indirect control such as dynamic pricing have been proposed, e.g. in [26, 11]. The large number of home users and the fact that each of them has negligible impact on the total cost makes such an approach challenging though. This challenge motivates the introduction of the aggregator as an intermediary entity between the operator and the end users.

In our model, the operator provides monetary rewards $\boldsymbol{\lambda} = \{\lambda_j \geq 0 : j \in \mathcal{A}\}$ to the aggregators so that they perform DR on its behalf. Notice that in this case the load vector $\mathbf{y}$ is dependent on $\boldsymbol{\lambda}$. In particular, the operator is willing to provide a portion $\hat{\lambda} = \sum_{j \in \mathcal{A}} \lambda_j$ of its DR gain to the aggregators. The DR gain Δc is the reduction of the power generation cost that results from reward $\boldsymbol{\lambda}$ and is given by

$$\Delta c(\mathbf{y}(\boldsymbol{\lambda})) = \sum_{t \in \mathcal{T}} \Delta c_t(y_t(\boldsymbol{\lambda})) = \sum_{t \in \mathcal{T}} \left[c_t^0 - c_t(y_t(\boldsymbol{\lambda})) \right], \tag{3.2}$$

where c_t^0 is the power generation cost at timeslot t if no DR is applied.

Thus, in the presence of aggregators, problem (3.1) becomes:

Operator's problem (min operational cost):

$$\min_{\boldsymbol{\lambda}} \quad \sum_{t \in \mathcal{T}} c_t \left(y_t(\boldsymbol{\lambda}) \right) + \hat{\lambda} \Delta c \left(\boldsymbol{y}(\boldsymbol{\lambda}) \right) \tag{3.3}$$

$$\text{s.t.} \quad 0 \le \hat{\lambda} \le 1$$

$$\lambda_j \ge 0 \qquad \forall j \in \mathcal{A}.$$

The objective function of the operator captures all its expenses in a DR market, i.e. both power generation cost and its reward to the aggregators for their services. Notice that the reward provided to the aggregators depends only on the quality of their aggregate DR services. The exact way that reward is allocated to the aggregators is addressed in Section 3.5.

3.3.2 The role of the aggregators

Since home users cannot negotiate directly with the operator, they enroll in DR programs provided by an aggregator that aggregates several small residential DR assets into a larger unit, in order to increase their negotiation power. Given that each user is assigned to an aggregator through a contract, we denote with $\mathcal{N}_j$ the corresponding set of users under aggregator j. The role of the aggregator is twofold: *i*) to provide DR services to the operator and *ii*) to guarantee a reduced electricity bill to the end-users.

Each aggregator tries to shape the load pattern of its users and receives compensation for the cost savings incurred to the operator due to this shaping. We assume that an aggregator incentivizes users to modify their power consumption profile through dynamic compensation per unit of power. The strategy of aggregator j constitutes of the compensation vector $\boldsymbol{p}_j = \{p_{jt} : t \in \mathcal{T}\}$. Let $\boldsymbol{d}_j = \{d_{jt} : t \in \mathcal{T}\}$ denote the cumulative load of aggregator j at time slot t, over all its users $\mathcal{N}_j$, that results from compensation $\boldsymbol{p}_j$.

From aggregator's j point of view, the DR gain Δc of (3.2) depends on its own compensation strategy $\boldsymbol{p}_j$, but also on the compensation strategy of all aggregators other than j denoted by $\boldsymbol{P}_{-j} = (\boldsymbol{p}_1, \dots \boldsymbol{p}_{j-1}, \boldsymbol{p}_{j+1}, \dots \boldsymbol{p}_J)$. The same holds also for the actual reward received by aggregator j, since power generation cost at time slot t is a function of the corresponding total load $y_t = \sum_{j \in \mathcal{A}} d_{jt}$.

The objective of aggregator j is to maximize its net profit by solving the following optimization problem:

Aggregator's j problem (max net profit):

$$\max_{\boldsymbol{p}_j} \quad \lambda_j \Delta c(\boldsymbol{p}_j, \boldsymbol{P}_{-j}) - \sum_{t \in \mathcal{T}} p_{jt} d_{jt}(\boldsymbol{p}_j) \tag{3.4}$$

$$\text{s.t.} \quad p_{jt} \ge 0 \qquad \forall t \in \mathcal{T}. \tag{3.5}$$

The first term corresponds to the reward received from the operator, while the second term is the compensation provided to the users.

31

3.3.3 Residential demand scheduling

At the home premises, under the current model of flat pricing, users tend to use their appliances at the most convenient time throughout the day, driven by their personal preference. For example, most people activate cooling within the hottest period of a day, thus creating demand peaks. We define as $x_i^0 = \{x_{it}^0 : t \in \mathcal{T}\}$ the reference consumption profile of user i that reflects its preferable (i.e. most convenient) demand pattern in the absence of any DR incentives.

In our model, monetary compensation provided by the aggregators motivates users to move load out of peak consumption periods. We assume that end-users are price-taking since they control a negligible amount of demand and hence cannot affect the compensation strategy of the aggregator. Each user is characterized by a total daily electricity requirement of W_i Watts that has to be scheduled for the following day. We assume W_i to be fixed and independent of the provided rewards. This enables a fair comparison with other pricing schemes, such as flat pricing, since the assumption of curtailable demands further amplifies the benefits of any DR scheme. The operator charges a flat price q_f for each Watt of consumption, which incurs a fixed daily cost of $q_f W_i$ for user i.

For each user i we define control $x_i = \{x_{it} : t \in \mathcal{T}\}$ capturing the power consumption pattern of the following day. Clearly, for each aggregator j holds $d_{jt} = \sum_{i \in \mathcal{N}_j} x_{it}$. Each user $i \in \mathcal{N}_j$ aims to maximize its net payoff, i.e. the compensation received by the aggregator minus the incurred dissatisfaction, leading to the following problem of water-filling type:

User's i problem (max net payoff):

$$\max_{x_i} \quad \sum_{t \in \mathcal{T}} [x_{it} p_{jt} - V_{it}(x_{it})] \tag{3.6}$$

$$\text{s.t.} \quad x_{it} \geq 0, \qquad \forall t \in \mathcal{T} \tag{3.7}$$

$$\sum_{t \in \mathcal{T}} x_{it} = W_i, \tag{3.8}$$

where the *disutility* function $V_{it}(\cdot)$ captures the dissatisfaction caused due to deviation from the reference consumption. Function $V_{it}(\cdot)$ may be taken to be convex, since the differential dissatisfaction of a user increases as the amount of deviation from the reference power consumption increases. An indicative example of such a function that we also use throughout this chapter is

$$V_{it}(x_{it}) = v_i(x_{it} - x_{it}^0)^2. \tag{3.9}$$

We call $v_i \geq 0$ the inelasticity parameter of user i. Small enough values of v_i indicate elastic users that experience minimum dissatisfaction if their consumption pattern is modified. On the other hand, a large value indicates a user that is sensitive in changes.

3.4 A Benchmark Model for Utility-Aggregator-User Interaction

In this section, we consider a benchmark scenario of full information, where the utility operator has global knowledge of the system parameters, namely the reference power consumption profiles x_i, the inelasticity parameters v_i and the set of allocated users $\mathcal{N}_j$ to each aggregator j. This scenario captures the case where all the involved entities are willing to report their parameters to the operator.

This approach provides insight on how misaligned are the interests of the market entities and whether disclosing this information to the operator is beneficial for the lower levels. It also serves as a benchmark regarding the cost of the operator. We formulate the problem as a multilevel optimization problem [33] and provide a characterization of the solution for the scenario where *i*) at the lower level the users modify their demand pattern according to the compensation advertised by the aggregators, *ii*) at the middle level the aggregators determine their compensation strategy so as to maximize their net profit, given the rewards offered to them by the operator, and *iii*) at the upper level the operator computes the reward per unit of cost reduction to provide to the aggregators so as to minimize its operational cost.

3.4.1 Aggregator-user interaction

The aggregator would ideally like to control the user appliances directly, in order to impose the consumption pattern that maximizes its net profit. However, each demand of a user participates in demand response according to its own disutility function. Thus, the aggregator has to provide incentives to its end-users in the form of compensation in order to motivate them to modify their demand load pattern. In particular, for a given reward vector $\boldsymbol{\lambda}_j$ from the operator, each aggregator has to find its compensation strategy $\boldsymbol{p}_j$ that solves (3.4).

The optimal strategy of aggregator j depends on the aggregate demand profile of its users throughout the day, $\boldsymbol{d}_j$, which is directly affected by its compensation strategy $\boldsymbol{p}_j$ through the user problem (3.6). In order to calculate its optimal response, the aggregator needs to know the analytical form of $d_{jt}(\boldsymbol{p}_j) = \sum_{i\in\mathcal{N}_j} x_{it}(\boldsymbol{p}_j)$, which in our scenario is an increasing function of the corresponding reward p_{jt}, but it is decreasing in the reward provided for any other slot, i.e. $\frac{\partial d_{jt}(\boldsymbol{p}_j)}{\partial p_{jt}} \geq 0$ and $\frac{\partial d_{jt}(\boldsymbol{p}_j)}{\partial p_{j\tau}} \leq 0 \ \forall \tau \neq t$. This is expected since increasing compensation for a specific time slot motivates users to move load there, whereas increasing compensation in any other slot motivates users to move load out of it.

In the full information case, the aggregator is aware of the disutility functions of appliances and hence can calculate from the Langrange optimality conditions the best response for each demand. For a given reward vector $\boldsymbol{p}_j$, (3.6) is a convex optimization problem in x_i. Under the condition that none of the constraints (3.7) is active, we can derive an analytical expression for $d_{jt}(\boldsymbol{p}_j)$. The Langrangian of user i is then given by:

$$L(\boldsymbol{x}_i, \gamma_i) = \sum_{t\in\mathcal{T}} [x_{it} p_{jt} - V_{it}(x_{it})] + \gamma_i (\sum_{t\in\mathcal{T}} x_{it} - W_i), \tag{3.10}$$

where γ_i is the Lagrange multiplier corresponding to constraint (3.8). Notice that we have not included the terms of (3.7), since the corresponding KKT multipliers are taken to be zero. Then, due to convexity of the objective function, the KKT conditions are necessary and sufficient conditions for optimality. Thus, we may derive the following T optimality conditions:

$$\frac{\partial L(\boldsymbol{x}_i, \gamma_i)}{\partial x_{it}} = 0 \implies p_{jt}(x_{it}, \gamma_i) = \frac{\partial V_{it}(x_{it})}{\partial x_{it}} - \gamma_i, \quad \forall t \in \mathcal{T}. \tag{3.11}$$

From (3.11) we can calculate the demand portion x_{it} of each user i at each time slot t as an expression of $\boldsymbol{p}_j$ and γ_i. Strict convexity of the disutility function guarantees that its partial derivative is strictly increasing and hence invertible. Thus, for the disutility function in (3.9) we get:

$$x_{it}(p_{jt}, \gamma_i) = \frac{p_{jt} + \gamma_i}{2v_i} + x_{it}^0, \quad \forall t \in \mathcal{T}. \tag{3.12}$$

In addition, the equality constraint (3.8) needs to hold. By replacing x_{it} from (3.12) to (3.8), the aggregator can compute γ_i as an expression of $\boldsymbol{p}_j$ as $\gamma_i(\boldsymbol{p}_j) = -\frac{1}{T}\sum_{t\in\mathcal{T}} p_{jt} = -\bar{p}_j$, with $\bar{p}_j$ denoting the average compensation provided by aggregator j over the day horizon T. By replacing this into (3.12), we derive the actual distribution of demands as a function of the compensation vector:

$$\boldsymbol{x}_i(\boldsymbol{p}_j) = \frac{1}{2v_i}(\boldsymbol{p}_j - \bar{p}_j\boldsymbol{1}) + \boldsymbol{x}_i^0, \tag{3.13}$$

with $\boldsymbol{1}$ denoting the column vector of ones of dimension T.

Then, since there is no coupling among users, each aggregator j can derive $\boldsymbol{p}_j$ by replacing $d_{jt}(\boldsymbol{p}_j) = \sum_{i\in\mathcal{N}_j} x_{it}(\boldsymbol{p}_j)$ into (3.4), i.e. by solving:

$$\max_{\boldsymbol{p}_j} \quad \lambda_j \sum_{t=1}^{T} \Delta c_t \Big(\sum_{i\in\mathcal{N}_j} x_{it}(\boldsymbol{p}_j), \boldsymbol{P}_{-j} \Big) - \sum_{t=1}^{T} p_{jt} \sum_{i\in\mathcal{D}_j} x_{it}(\boldsymbol{p}_j) \tag{3.14}$$

$$\text{s.t.} \quad (3.5), \; x_{it}(\boldsymbol{p}_j) > 0, \quad \forall i \in \mathcal{N}_j, t \in \mathcal{T}.$$

Although, for notational simplicity we have assumed that each aggregator provides the same compensation to all its users, discriminatory compensation strategies can be easily captured by our model.

3.4.2 Operator-aggregator interaction

The operator has to find the monetary reward vector $\boldsymbol{\lambda}^*$ that minimizes its operational cost. However, the exact impact of its rewards on demand distribution is difficult to quantify, since it involves also the optimization problems of the lower two levels. In particular, the operator needs to know the analytical expression of $d_{jt}(\boldsymbol{p}_j(\lambda_j))$. Notice that the reward strategy of the operator determines the reward provided by the aggregator, which in turn determines the demands of the users.

This problem falls within the class of multi-level optimization problems, which are particularly challenging to solve [33]. In the previous section, we showed how the lower two levels can be merged into one optimization problem. However, the previously derived solution is a numerical one, while in

order to characterize the DR solution from the operator's point of view, we need an analytic expression for the optimal solution of the problem of each aggregator. Since in general such an expression cannot be derived, we calculate the reward strategy of the operator numerically and evaluate it in Section 3.6.

3.5   A non-cooperative market mechanism

In the previous section, we assumed that all the information about end-users (demand load and disutility functions) and about the total demand load under each aggregator is available at the operator side. We used this assumption to focus on the problem from the point of view of the utility operator. Such a scenario of information exchange could be plausible in the case that such exchanges are part of prior agreements between the involved entities.

In this section, we consider how the hierarchical market may function in a setting, where the aforementioned assumption of information exchange is relaxed. We propose a top-down approach where at the higher level the operator announces its reward strategy to the aggregators, while on the lower level each aggregator presents monetary compensations to end-users and negotiates with them about the DR services that they can provide for this compensation. Finally, each aggregator responds to the operator with a cost reduction offer. All the required information exchange takes place during the previous day, forming thus a day-ahead DR market.

Starting from the upper level, the operator has to find the monetary reward vector $\boldsymbol{\lambda}$ that minimizes its operational cost given by (3.3). The total gain of the proposed DR market is captured by Δc from (3.2), which is an increasing function of the provided reward $\hat{\lambda}$. Each aggregator receives a portion $\lambda_j \Delta c$ of this gain and provides part of it to the end-users in order to incentivize them to modify their consumption pattern. On the other hand the operator derives the remaining gain $(1 - \hat{\lambda})\Delta c$. However, the exact impact of reward vector $\boldsymbol{\lambda}$ on the operational cost of the operator is unknown, due to the involvement of the aggregators and the private disutility functions of the users. Thus, we propose a *repeated* auction mechanism with iterative elimination at each iteration step, where the aggregators compete with each other to sell their DR services.

At each iteration k, the operator announces the current total demand load $\mathbf{y}^k = \{y_t^k \geq 0 : t \in \mathcal{T}\}$ and a scalar total reward $\hat{\lambda}^k$. Then, it conducts a first price sealed-bid auction, where each aggregator announces to the operator the DR service that it can provide for this reward. The operator accepts the bid that guarantees the highest cost reduction, and this aggregator is removed from subsequent auctions. Then, reward per unit of cost reduction is increased by the operator according to $\hat{\lambda}^{k+1} = \hat{\lambda}^k + \xi$, and a new auction is conducted with the remaining aggregators.

By increasing the provided reward, the operator enables the aggregators to provide even higher compensation to the home users. Hence, the users get sufficient incentives to move even less elastic load out of peak-demand periods. This process is repeated until either there are no remaining aggregators, or the provided reward $\hat{\lambda}^k$ becomes larger than 1 which violates the constraint of the operator.

In order to calculate its DR bid, the aggregator receives the modified consumption patterns of

the users in response to the advertised compensation. Each aggregator would like to provide the minimum compensation to end-users, in order to reap maximum benefit for itself. However, such a strategy would reduce the quality of its DR services to the operator and consequently it would lower the chances that its bid wins over the bids of other aggregators. Thus, the competition of the aggregators is implicitly beneficial for the end-users.

Notice also that the seemingly plausible strategy of withdrawing from the first auction rounds does not guarantee increased benefit for the aggregator. On the one hand, $\hat{\lambda}^k$, the operator reward per unit of generation cost reduction, increases across iterations. On the other hand, each accepted bid leads to a more balanced total consumption pattern. Thus, later iterations provide only limited opportunities for further balancing, which in turn may lead to reduced profit for the aggregator despite the increasing reward.

At the lower level, the aggregator needs to know for each time slot t the analytical form of $d_{jt}(\boldsymbol{p}_j)$, which by (3.11) is a function of $\frac{\partial V_{it}(x_{it})}{\partial x_{it}}$, so as to calculate its optimal bid. Since the aggregator is unaware of the user disutility functions, such information may only be deduced by observing user responses to the announced compensation. For the single parameter disutility function (3.9) and under the assumption that the aggregator is aware of the generic form of the disutility function, it can deduce the inelasticity parameter through the following estimation phase. Each aggregator j selects an arbitrary $\ell \in \mathcal{T}$ and announces a unit compensation vector $\boldsymbol{p}_j^\ell$, where $p_{j\ell} = 1$ and $p_{jt} = 0, \forall t \in \mathcal{T} \setminus \{\ell\}$. Through the modified load pattern $\boldsymbol{x}_i(\boldsymbol{p}_j^\ell)$, the aggregator can deduce the exact form of the utility function of each user i, since it can calculate the corresponding v_i parameters from (3.13). Notice that this operation has to be performed only once.

At each iteration k, each aggregator j receives the total consumption vector $\boldsymbol{y}_t^k$ and reward $\hat{\lambda}^k$ from the operator. Thus, aggregator j can replace $d_{jt}(\boldsymbol{p}_j)$ into (3.4) and solve it similarly to the full information case. The total load of aggregators other than j can be calculated as $d_{-jt}^k = y_t^k - d_{jt}^k$.

The exact operation of the proposed day–ahead market is described in detail in Algorithm 3. For notational simplicity we omit the function arguments. Notice that lines $1-3$ describe a required initialization phase where the aggregators and the operator deduce the behaviour of users under flat pricing. In this direction, the operator announces zero reward to the aggregators for the following day, in order to deduce the reference consumption pattern of the residential users $\boldsymbol{x}_i^0$. Given the zero reward from the operator to the aggregators, the latter have no incentive to provide any DR services to the operator and hence announce zero compensation to the users. The end-users respond with the reference consumption pattern that they prefer in the absence of incentives. Lines $4-8$ correspond to the inelasticity estimation phase, where each aggregator calculates the inelasticity parameters of the demands (appliances) of its users. Lines $9-18$ constitute the main body of the proposed market, where the operator increases the provided reward, while the aggregators compete with each other through their bids.

Remark 3.5.1. *A nonprofit operator that only cares for the total benefit of the system would set $\hat{\lambda} = 1$. This would maximize the DR market gain Δc but it would also lead to operational cost c^0, which equals to that of flat pricing (i.e. no DR).*

Algorithm 3 The operation of the day-ahead DR market
***

Input: ξ: auction price increment step

1: $\lambda^0 \leftarrow 0$ // operator's zero reward
2: $\boldsymbol{p}_j \leftarrow \boldsymbol{0}$ // aggregators' zero compensation
3: $c^0 \leftarrow \sum_{t \in \mathcal{T}} c_t(y_t^0)$ // resulting load pattern- no DR
4: **for all** $j \in \mathcal{A}$ **do** // estimation of inelasticity param. v_i
5:     aggregator j announces unit reward vector $\boldsymbol{p}_j^\ell$
6:     user i responds by $\boldsymbol{x}_i$
7:     aggregator calculates v_i from (3.13)
8: **end for**
9: **repeat** // iteration k
10:     operator announces total demand vector $\boldsymbol{y}^{k-1}$
11:     $\hat{\lambda}^k \leftarrow \hat{\lambda}^{k-1} + \xi$ // operator increases reward by $\xi > 0$
12:     **for all** $j \in \mathcal{A}$ **do** // each agg calculates its response bid
13:         aggregator j calculates compensation $\boldsymbol{p}_j^k$ from (3.14)
14:         aggregator j responds by $\boldsymbol{d}_j^k$ to operator
15:     **end for**
16:     operator accepts the maximum bid
17:     $\mathcal{A} = \mathcal{A} \setminus \{j\}$ // accepted aggregator j is removed from negotiations
18: **until** $\hat{\lambda}^k \geq 1$ // termination condition
***

Remark 3.5.2. *Given that the flat price q_f is set so as to cover operator expenses, i.e. $q_f \sum_i W_i = c^0$ the proposed market has no deficit, since for any value of $\hat{\lambda}$ always holds $c(\hat{\lambda}) \leq c^0$.*

3.6 Numerical Evaluation

In order to evaluate the performance of the proposed market, we rely on a realistic DR dataset based on the per appliance demand generator of [23]. Our dataset captures the daily schedule of 33 types of appliances of diverse characteristics that may exist in a residence. We assume that the disutility of each user is given by the parametric disutility function (3.9) with v_i uniformly distributed in $[0, v_{max}]$, with $v = v_{max}/2$ denoting the mean inelasticity value.

We consider a market of a single operator, J aggregators and $N = 10^4$ households equally divided among the aggregators. In order to quantify the benefits of our DR scheme, we use the power generation cost function $c(y_t) = y_t^2$, which is a convex and increasing function of the corresponding total demand y_t at each timeslot t.

Fig. 3.3 and Fig. 3.4 depict the total consumption pattern across a day for the market approach of Section 3.5 for scenarios of low and high mean inelasticity v respectively. Fig. 3.3(a) and Fig. 3.4(a) correspond to the default case of flat pricing where no compensation is provided to the users. For the case of low inelasticity, Fig. 3.3(b) and Fig. 3.3(c) show the evolution of the load pattern across a day as the DR bids of the aggregators are accepted. The introduction of the aggregators' DR services smooths the load pattern significantly. In the end, the total demand is spread throughout the day. Instead, in the scenario of higher inelasticity depicted in Fig. 3.4(b) and Fig. 3.4(c), consumption is less evenly distributed across the day. Thus, inelasticity of demands, captured by inelasticity parameter

(a) Flat pricing   (b) First iteration   (c) Last iteration
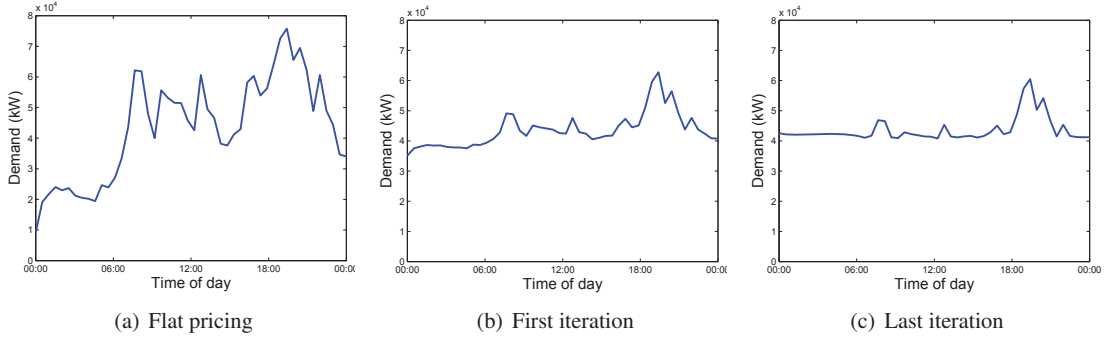
Figure 3.3: Demand distribution throughout the day for low mean inelasticity v.

(a) Flat pricing   (b) First iteration   (c) Last iteration
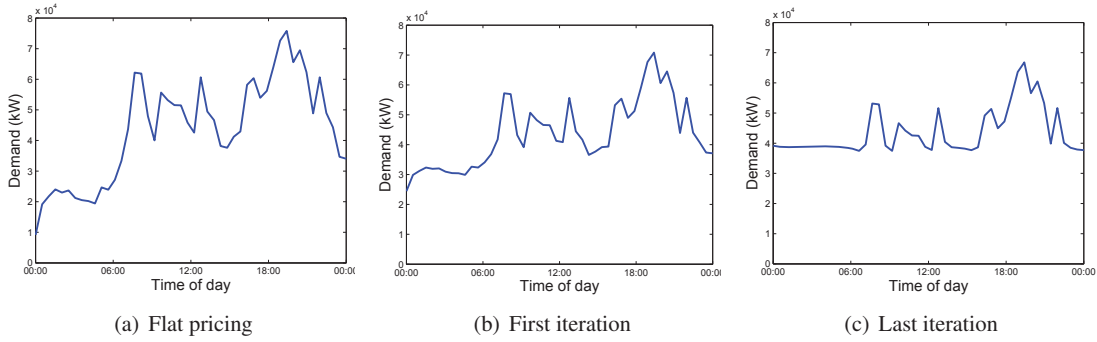
Figure 3.4: Demand distribution throughout the day for high mean inelasticity v.

v_i, plays a central role on whether the resulting consumption load pattern will be smooth across the day.

Next, we quantify the DR benefits for each of the participating entities, namely the net payoff derived by the end-users given by (3.6), the net benefit of the aggregators given by (3.4) and the total cost of the operator captured by (3.3). The actual reduction of the user's electricity bill is at least as much his net payoff, since the latter contains also the disutility term. Therefore, our findings in terms of net payoff can be directly translated to monetary savings in the electricity bill of the user. For comparison, remember that under flat pricing the cost of the operator is c^0, while all the aggregators and the users derive zero payoff.

First, we consider the impact of the reward strategy $\hat{\lambda}$ of the operator on the market for the case of a single aggregator ($J = 1$). In particular, we investigate how the total DR gain of our market, Δc, is allocated to the market participants. In Fig. 3.5(a) we observe that the operational cost of the operator is not monotonic in $\hat{\lambda}$, while electricity cost c is decreasing in $\hat{\lambda}$. We depict also the optimal solution $\hat{\lambda}^*$ that we derive from the benchmark case. In contrast, in Fig. 3.5(b) we see that the DR gain Δc is an increasing function of $\hat{\lambda}$ and the same holds for the total net profit of the aggregators and the net payoff of home users. The actual share of the arising DR benefits is depicted in Fig.3.5(c). Notice that for $\hat{\lambda} < 0.15$, the market derives zero DR gain. Hence, we depict the corresponding gains as zeros. In the particular setting, the users generally derive almost double the gains of the aggregators.

(a) Operator side

(b) Aggregator / Users side

(c) Share of the DR gains
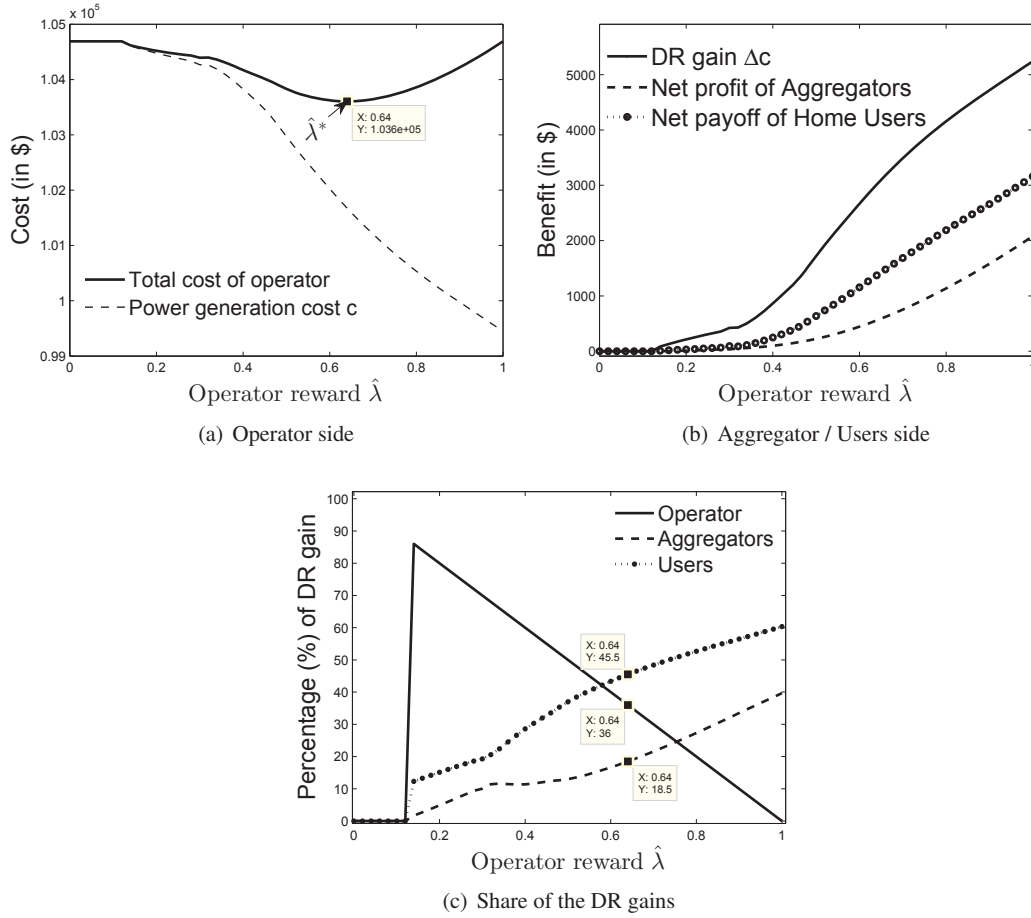
Figure 3.5: Impact of the operator reward strategy λ on the DR benefits of the market entities.

In addition, at $\hat{\lambda}^* = 0.64$, all participants derive a significant portion of the DR gains which is not true for extreme values of $\hat{\lambda}$.

The exact amount of DR benefits that each entity receives, depends on the efficiency of the applied DR scheme and the inelasticity of user demands. In Fig. 3.6 we compare the performance of the full information approach and the proposed market mechanism as the average inelasticity of demands increases. In the following plots, the rightmost point of the x axis corresponds to the case of extreme inelasticity which coincides with the solution of no compensation (i.e. flat pricing). We observe that our approach sacrifices some of the operators' gain for the sake of aggregators and users. Thus, although the operational cost of the operator is in general comparable (except for low inelasticity regime), our DR market provides a higher DR gain. This additional market gain is shared unevenly between the aggregator and the corresponding end-users. Compared to flat pricing though, the proposed DR market guarantees significant benefits for all the participating entities.

Next, we compare the performance of the proposed scheme of adaptive rewards $\hat{\lambda}^k$ with that achieved when a nonprofit operator provides all the DR benefits to the lower levels through a constant
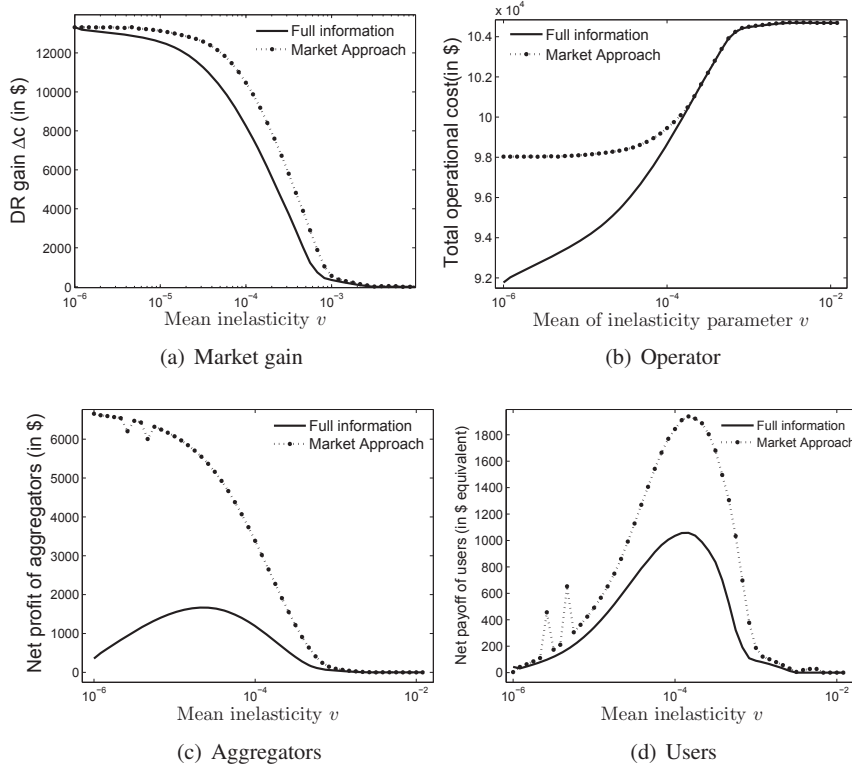
Figure 3.6: Impact of full information availability on the benefit derived by each market entity for different values of mean inelasticity v.

reward $\hat{\lambda} = 1$. For a scenario of $J = 5$ aggregators, we depict in Fig. 3.7 and Fig. 3.8 the impact of inelasticity on the benefits of each entity. From Fig. 3.7(a) we observe that our DR scheme provides a reduction in the total operational cost of up to 15% in comparison to the case of flat pricing. However, the exact cost reduction depends strongly on the inherent inelasticity of the user demands. The same holds for the power generation cost. Our scheme guarantees also that the aggregators and the users generally derive comparable financial benefits.

Interestingly, the net payoff of the users and the net profit of the aggregators are not monotonic in inelasticity parameter v. Instead the maximum is observed in the medium inelasticity regime. This can be easily justified. In the high inelasticity regime, the aggregators spend most of the reward provided by the operator in order to motivate end users to modify their demand pattern. This does not lead to increased net payoff for the users though, since the compensation provided is counteracted by the inconvenience caused. On the other end, low inelasticity leads to reduced negotiation power from the user side and consequently reduced income.

Similar results are derived from the scenario of a nonprofit operator, depicted in Fig. 3.7(b). However, here the total cost of the operator is constant, while small enough inelasticity enables the aggregators to provide low compensation to the end-users in order to acquire their DR services and hence exploit most of the DR benefits for themselves. By comparing the two approaches, we observe

(a) Our market mechanism of adaptive reward $\hat{\lambda}^k$ (b) The case of a nonprofit operator $\hat{\lambda}^k = 1, \forall k$
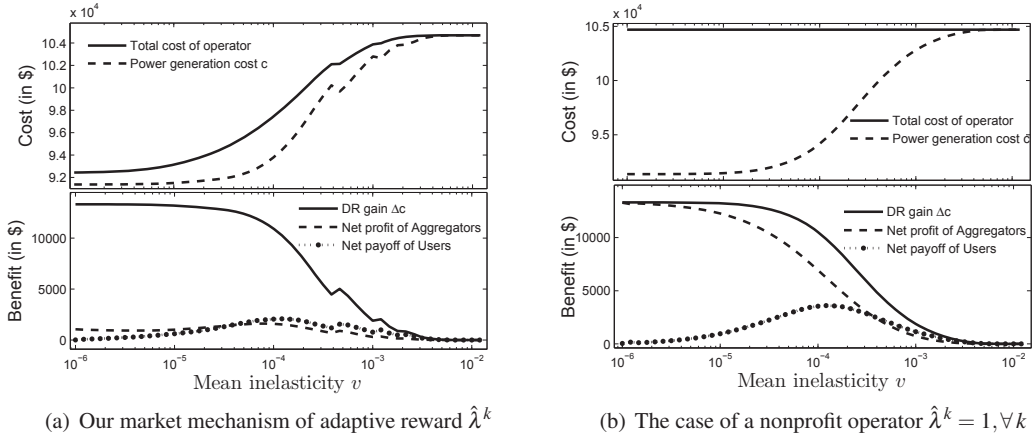
Figure 3.7: DR benefits for each market entity in a scenario of $J = 5$ aggregators.

(a) Our market mechanism of adaptive reward $\hat{\lambda}^k$ (b) The case of a nonprofit operator $\hat{\lambda}^k = 1, \forall k$
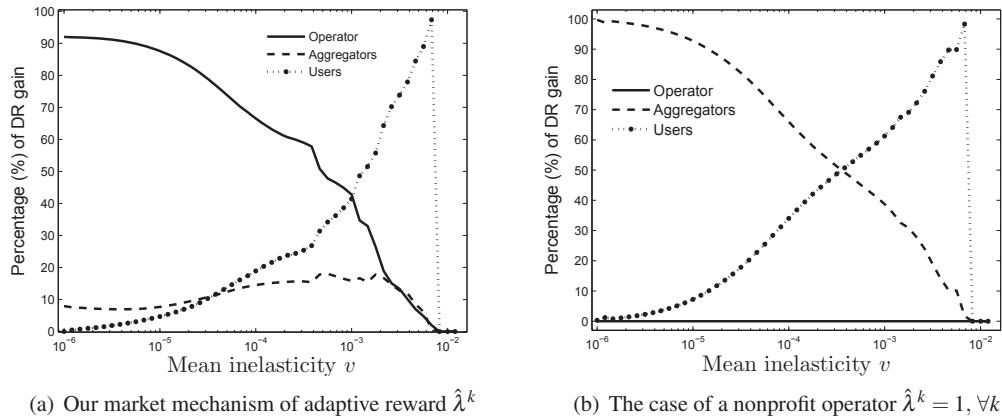
Figure 3.8: Percentage of DR gain derived by each entity in a scenario of $J = 5$ aggregators.

that although the benefits of the aggregators and the users increase in the case of a nonprofit operator, the electricity cost and the market gain remain almost the same. This indicates that our adaptive approach exploits most of the benefits arising from DR. However, as shown in Fig. 3.8 the portion of DR gain received by each entity differs significantly in the two approaches.

3.7 Conclusions

The main objective of this work is to analytically formulate the role of aggregators in future smart grid. We focused on a hierarchical market structure and investigated the interaction of end-users, the utility operator, and several competing aggregators. Under the assumption of price-taking and self-interested users, we characterized the optimal solution from the operator point of view that seeks to minimize its operational cost. We also proposed a day-ahead DR market that leads to significant benefits for all the involved entities. The proposed compensation strategy guarantees that compared

to flat pricing no entity has a budget deficit. In addition, the benefits of DR are maximized when the operator is nonprofit.

Based on realistic demand traces we demonstrated that access to user information enables the operator to exploit most of the DR benefits for itself, while the proposed non-cooperative market mechanism leads to a more fair allocation of the DR benefits. Interestingly, our numerical evaluation indicates that the utility of the end-users in not a monotonously increasing function of elasticity. Here, we assumed that the users truthfully respond to the announced compensation from the aggregators. However, since additional benefits may arise through strategic misreporting, the investigation of such user strategies and the derivation of mechanisms that guarantee truthfulness are interesting topics for future study.

Chapter 4

Dynamic Allocation and Migration of Mobile Tasks in Cloud Computing Systems

Previous chapters were devoted to our efforts on improving the operation of the power grid. Next, we investigate ways to make energy-constrained communication systems themselves more efficient. In this chapter, we consider the scenario of mobile devices that, due to limited processing and energy resources, outsource their computationally intensive tasks to the cloud. We propose a mobile cloud computing architecture that consists of a local cloud, attached to wireless access infrastructure (e.g. LTE base station), and a back-end cloud. Each task is hosted in a Virtual Machine (VM) and is associated with an evolving volume of data.

The mobile aims to minimize energy consumption due to execution of the task. The challenge for the cloud is to minimize the task execution and data download time to the user, whose location changes due to mobility. The key technique we propose is VM migration which enables efficient exploitation of the geographically dispersed cloud resources. In general, a VM should migrate to the server where it is expected to end faster, considering also the incurred migration delay. To account for all arising causes of latency, we propose a model for virtualized servers that captures interaction of collocated VMs and its impact on performance. This feature, along with accompanying data evolution and user mobility (which determines access to the local cloud) affects the task migration decisions significantly.

We propose three classes of online VM migration policies for the cloud, spanning fully uncoordinated ones, in which each user or server autonomously makes its migration decisions, up to cloud-wide coordinated ones, where migration decisions are made by the cloud provider. We identify when task offloading is beneficial for a mobile user in terms of energy consumption. Finally, we demonstrate how VM migrations can be exploited towards a more energy-efficient cloud.
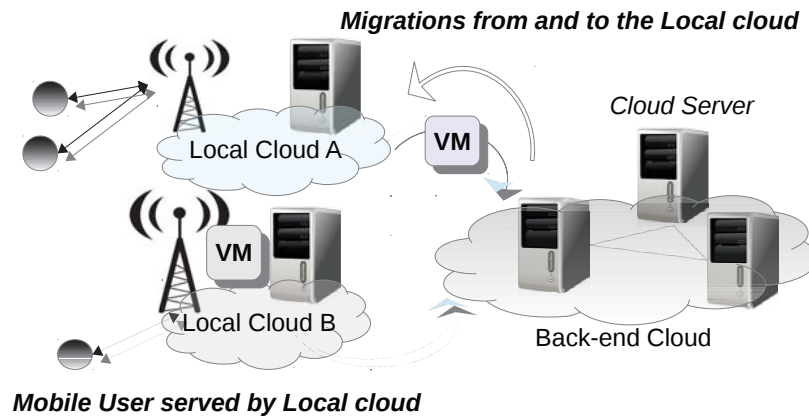
Figure 4.1: Mobile Cloud Computing architecture

4.1 Introduction to Mobile Cloud Computing

Cloud computing is one of today's most rapidly evolving technologies and it is increasingly adopted by large companies to host various service platforms (e.g. iCloud by Apple, GoogleApps by Google, EC2 by Amazon, etc). It facilitates rapid and flexible access to a shared pool of dynamically configured resources, notably, storage capacity and computational power. The collection of these pooled resources reside in geographically dispersed servers. All these would not have been possible without the recent advances in virtualization technology that enable the agile consolidation and parallel execution of diverse applications on the same physical machine (server), each hosted in a separate Virtual Machine (VM).

Parallel to that, end-user mobility has become an essential feature of contemporary wireless networks. Applications that run on mobile devices (e.g. speech recognition, environmental sensing, games) are now becoming more sophisticated than ever in terms of computing requirements and data generation. Proliferation of wireless access technologies like 3G, LTE and WiMax provide ample promise in taking cloud computing to the next level, where mobile devices can access the cloud to offload computationally intensive tasks. The later asset has given rise to what is known as *mobile cloud computing*. The problem we address in this work is how to best leverage cloud resources to minimize execution time of mobile tasks. This goal is realized through task migration policies that move the task/VM closer to the mobile user.

We consider the following architecture that brings into stage clouds together with mobile wireless devices. Mobile devices access cloud facilities through readily available hubs like 4G/LTE base stations or WiFi access points. There exists a back-end cloud facility whose servers are interconnected through wire-line links while the access link from the mobile device-client to the cloud is wireless (Fig. 4.1). In addition smaller sized cloud servers are attached to the points of wireless access, such as 4G/LTE base stations, forming local clouds in order to avoid the additional communication delay of the Internet. The local and the back-end cloud are interconnected and are both managed by a single provider.

The objective of the cloud provider is to efficiently handle application tasks that arise from cloud clients, in our case from energy-constrained mobile devices. This oftentimes translates to executing tasks as quickly as possible. Quicker task completion and data download time to the mobile user enhances user quality of experience and positions the provider better in the market. Besides, cloud resources become more often free and available for other tasks thus, increasing the available processing capacity. On the other hand, a mobile user is also concerned about the energy required to execute a task.

Efficient management of the cloud infrastructure is a challenging task due to the highly dynamic and unpredictable nature of the system. First, the pattern of instantaneous resource demand varies with time and location, since new task demands arise continually at various locations while others complete service.

Second, the resource supply is also subject to continuous changes. The available processing capacity of virtualized cloud servers changes due to the unpredictable interaction of tasks/VMs co-existing on the same physical server. This coexistence unavoidably leads to contention for available resources (e.g. CPU, caches, network, I/O) and an associated overhead that is often task-dependent and difficult to model or predict. Thus, the true available computing resources change significantly with time. Another source of variation of the processing capacity of cloud servers comes from the execution of cloud management tasks in the background.

Third, tasks are usually accompanied with a time-varying data volume, that may be increasing, decreasing or constant with time, corresponding to different types of applications that generate new information, compress it as they evolve, or do not modify data respectively. The time required to transfer a task to another server depends decisively on the pattern of data evolution.

The issues above motivate the need for *task migration*. In this direction the latest version of vSphere, the virtualization platform of VMware that has a market share of 60% [39], incorporated the capability of VM migration across distant servers [40]. In vMotion, the migration scheduler of vSphere, migrations are performed so as to keep average CPU and memory utilization balanced across servers. Instead, we propose that migrations should be performed according to the estimated performance. A task should migrate from its current server if the anticipated total execution time at a new server (including the migration and download delay) is less than the execution time at the current one.

User mobility introduces additional challenges, since at each time instant the task can be directly uploaded only to a certain subset of servers, notably those attached to the points of wireless access in range of the mobile user. The same holds for downloading the results once the task has finished execution. In general, a different cost can be assigned to each server, which captures the time delay of uploading/downloading data to/from this server. Thus, migrations should be such that the task follows the mobile user i.e. the server that executes the task should be "close" to the user in terms of delay.

For a simple motivating example, consider a CPU intensive task of 10^{10} CPU cycles and increasing volume of accompanying data. Initially the user uploads the task for execution to a back-end

cloud server of available CPU capacity 10^6 cycles/sec. Once the task has been completed by 50% the user moves in range of a local-cloud/BS of double the available CPU capacity, 2×10^6 cycles/sec. If the task stays on the initial server it will be completed after 5000 sec or 1.39 hours. Instead, given a local-backend interconnection link of 10Mbps and a current data volume of 10^{10} bits a migration to the local server will cost 1000 sec, and another 2500 sec for execution. Thus, the optimal strategy is to migrate the task to the local server immediately, leading to a completion time of 3500 sec or 0.97 hours, i.e. an improvement by $(1.39 - 0.97)/1.39 \times 100\% \simeq 30\%$. This migration is also beneficial regarding download time, since it avoids the additional communication delay over the Internet.

In this work we ask the question: what is the best way to perform migrations among the available cloud servers to minimize the average completion and data download time of mobile tasks? In order to adhere to a realistic scenario, we address the online version of the problem, in which information about the dynamics above is not available a priori, but rather it is presented online, just before the control decision is taken. For example, the rate of arrival of new tasks in each server is generally unknown. To counteract this, we propose that from time to time the cloud facility has to be reconfigured by properly reassigning some VMs to the cloud servers. We take a discrete time approach, where at each decision epoch some tasks are selected to migrate to another server. We propose three online task migration mechanisms that differ in complexity and required information exchange. The key underlying idea is that a migration should occur *only if it is beneficial for the processing time of the task, including the transfer delay for accompanying data*. In addition, we identify when task offloading is beneficial for a mobile user in terms of energy consumption and we also demonstrate how VM migrations can be exploited to reduce energy consumption of the cloud infrastructure itself.

The contributions of this work are as follows:

- Given that the problem of online scheduling with reassignment is NP-hard [13], we develop lightweight task migration mechanisms that capture the following cloud scenarios: (a) an extreme autonomic one, where each task selects its migration path individually, (b) an uncoordinated setup where migration decisions are taken by each cloud server separately based only on local information, (c) a fully coordinated setup where the cloud provider solves the global task scheduling/migration problem from the cloud provider point of view.

- We explicitly model the coexistence of several VMs on the same cloud server through an associated overhead on actual performance. This overhead depends on the number and types of coexisting tasks, is constantly measured and fed back to the entity that decides on task scheduling.

- We capture migration cost by incorporating the time required for communicating accompanying data to the new server in the decision making.

- We capture the impact of mobility on the migration strategy. The main idea is to migrate the task closer to the user as execution approaches its end.

- We demonstrate how VM migrations can be used for energy efficiency purposes.

The rest of this chapter is organized as follows. Initially, we provide an overview of related work in Section 4.2. In Section 4.3 we model interaction of tasks/VMs within the cloud. Section 4.4 describes the proposed migration mechanisms that capture the interaction of collocated tasks. Numerical results quantifying the performance of our schemes are presented in Section 4.5, while Section 4.6 concludes our study.

4.2   Related Work

Virtualization enables diverse tasks to run over a shared hardware platform. Each task is hosted on its own VM, which provides an isolated execution environment. However, operation over the same physical machine introduces significant contention for shared physical resources such as CPU, caches, disks or network. The problem of noisy-neighbours, where collocated tenants cause significant and unpredictable performance degradation, has been reported by several cloud customers [46, 47] and has even led companies to get off of the cloud [48]. The authors of [49] analyzed the performance of the Amazon Elastic Cloud Computing (EC2) facility and observed significant execution stalls due to processor sharing among the collocated VMs and drastically unstable TCP/UDP throughput. Similar performance limitations have been observed when computationally intensive scientific tasks were executed on existing cloud facilities [50].

Several works attempt to perform analytical [51] or experimental [52],[53] estimation of the multitenancy effect. The former though require a priori knowledge of the resource requirements(CPU, Memory, I/O) of each task, while the latter perform extensive profiling of different types of cloud tasks. However, [41] deduced from trace data from the Google cloud facility that significantly diverse tasks exist in the cloud and hence a characterization through profiling is impractical. Thus, we propose that multitenancy effect has to be predicted through online measurements as tasks are being executed. In this direction, [54] also investigates ways to provide predictable network performance to tenants.

Migration mechanisms in the context of cloud computing have received significant research interest mainly towards energy efficiency. However, migrations can be exploited also to improve cloud performance. Indicatively, [55] provides an experimental evaluation of the performance benefits of migrations. The most related work is [56] that applies max-weight inspired policies to maximize throughput through VM allocation/migration in dynamic cloud computing systems. In contrast to our work, the multitenancy effect, the evolving data footprint and the cost of migration that significantly affect the actual execution time are not considered. In [57], the authors propose a system that monitors resource usage and performs a migration whenever a Service Level Agreement (SLA) is violated for a sustained period. In all these works, the VM assignment/migrations are performed according to the required resources. Instead, we perform task migrations according to the actual performance of the task as this is observed online and by considering also the impact of multitenancy.

Generally, the impact of mobility has been considered only in the simpler scenario where a single task may migrate to a remote server, instead of being executed locally at the mobile terminal, a mech-

anism generally known as offloading. The resulting energy and time savings have been considered in [58], where a Markovian control framework is proposed. In a similar setting the authors of [59] propose the CloneCloud system that enables smartphones to offload their tasks to cloud computing facilities. Contrary to our model, these works do not take into account the interaction of multiple tasks and assume a static cloud environment that is not affected by the decisions made.

4.3 Mobile Cloud Computing Model

4.3.1 Cloud architecture

We consider a cloud architecture that consists of a set $\mathcal{K}$ of K cloud servers. A subset of these servers forms the back-end cloud, whereas the rest are attached to the wireless access infrastructure (Fig. 4.1). Such small scale local clouds provide processing resources that are closer to the mobile users, in order to avoid the additional communication delay of the Internet. Each server i is characterized by fixed processing capacity (speed) C_i flops/sec with the local cloud servers being of lower capacity in general. The available processing capacity may be time-varying due to cloud management background processes. A mobile task may be executed either at the local or at the back-end cloud.

At time t a mobile user is in the range of a wireless access technology, say a 4G base station (BS) j, and hence has access to its local cloud. Let $\mathcal{K}_j \subseteq \mathcal{K}$ denote the set of servers of local cloud j and $\gamma_{ij}(t)$ the capacity of the time-varying wireless link between user i and local cloud j at timeslot t. In addition, each local cloud is connected over the Internet with any back-end server k through an overlay link of delay L_{jk}. In order to simplify notation we may merge the two hop communication user–BS–server into a single link that captures the end-to-end path. Thus, the delay per unit of data between user i and server k through the corresponding BS j is $\frac{1}{r_{ik}} = \frac{1}{\gamma_{ij}(t)} + L_{jk}$.

The back-end infrastructure is connected through a fully mesh overlay network. We denote by $W_{\ell k}(t)$ the available bandwidth (in bits/sec) of the overlay link that connects back-end servers ℓ and k. We assume the generic case of time-varying server interconnection link capacity.

4.3.2 Application tasks

Application tasks arise continually, generated by mobile users at various locations of the cloud. Task i is characterized by its total processing burden B_i (in flops, or CPU cycles). Each task is also accompanied by an evolving volume of data in bits, $d_i(\cdot)$. This pattern may be increasing, decreasing or constant, modeling different types of applications that may generate new data or compress it as they evolve. We refer to this accompanying volume of data as the *data footprint* of the task.

A typical CPU-intensive task of decreasing footprint is video compression. Starting from an initial raw video of several gigabytes we end up with a compressed video of hundreds megabytes. As processing evolves, the part of the video that has been compressed becomes obsolete. On the other hand, most complex scientific calculations that are uploaded for execution at the cloud are characterized by increasing data footprint, since new data is continuously produced.

4.3.3 Task lifetime

The task lifetime consists of the following stages:

(i) **Task Upload**: Once a new task is generated by a user, the source code and any input data required for the initialization of the VM are uploaded to the cloud through the wireless link between the user and the corresponding point of wireless access. Then, the task is either executed at the local cloud, or its data are forwarded over the Internet to the back-end cloud, usually to the server that provides the minimum estimated data upload plus execution time.

(ii) **Execution / Migration**: the actual processing within the cloud. The execution of the task cannot be stalled, but the task may be transferred from its current server to a new one to continue execution there. This process is known as task migration and may occur multiple times during task execution.

(iii) **Download**: the mobile user retrieves the final results. We assume that once processing is completed, the final data are immediately downloaded by the user through its current BS. If the host server is not in range of the mobile user, data needs to be transferred to an accessible server.

We use the term *task lifetime T_i* to refer to the total time that task i spends in the cloud, including the time required for download. Upon initial assignment to a server, the task starts execution and accompanying data are generated. Each task i carries with it a so called progress indicator $x_i(t) \in [0,1]$ that evolves with time and denotes the percentage of completion. To understand the notation, for a task i that is executed at server j, and that presumably uses its entire capacity C_j, it would take B_i/C_j time to execute, and hence the progress indicator would evolve as $x_i(t) = \min\left\{1, \frac{C_j}{B_i}t\right\}$. The remaining processing requirement for the task is given by $b_i(t) = B_i(1 - x_i(t))$ while its data footprint evolution is given by $d_i(x_i(t))$.

4.3.4 Virtualization in cloud servers

A VM is a software segment that provides an isolated environment for the execution of a single task. In general execution of a task may require several communicating VMs. However, due to affinity constraints all the VMs comprising a task have to be scheduled on the same server. Thus, for simplicity we assume that each task is running on a single VM and use terms VM and task interchangeably. Upon migration of a task, a new virtual machine (VM) is created for this task on the destination server and the execution starts, while the initial VM is stopped and removed from the current server. A VM is also removed when the task finishes execution.

Although in theory an unlimited number of VMs can be collocated on the same server, *the co-existence of several VMs intuitively affects the performance experienced by each VM*. The incurred overhead leads to a graceful reduction in true available processing capacity, due to the inherent contention of multiple VMs for resources and the utilization of a part of CPU for performing other inter-VM management tasks. The term *multitenancy* is generally used for the coexistence of multiple VMs on the same physical machine.

We model this VM interdependence as follows. Let $\mathcal{A}_j(t)$ denote the set of active tasks of server

j at time t. For any server j with processing capacity C_j and $|\mathcal{A}_j| = n$, we assume that an amount of its service capacity $\varepsilon(n)$ is lost due to cross-VM management and VM contention. To model the impact of multi-tenancy, $\varepsilon(n)$ is taken to be increasing in the number of VMs n on the server. We assume that the remaining capacity is equally shared among collocated VMs. Hence, each task on server j obtains a processing share:

$$C_j^n = \frac{C_j}{n} - \varepsilon(n). \tag{4.1}$$

The overhead $\varepsilon(n)$ depends also on the *type* of coexisting tasks, in a way that is infeasible to quantify through profiling due to the numerous and significantly diverse tasks running on the cloud [41]. In order to circumvent this, the actual overhead is continually measured and fed back to the entity that makes the task migration decisions. Fix attention to computing the overhead $\varepsilon(n)$ due to n coexisting VMs on the same server. Let $\mathcal{S}_n$ denote the subset of servers that host n VMs. At each server $j \in \mathcal{S}_n$ and at each decision epoch τ, several sample values, say K, $\{C_{j\ell} : \ell = 1, \ldots, K\}$ are taken for the processing capacity that each hosted task enjoys. CPU consumption measurements are possible by using off-the-shelf cloud monitoring tools like Ganeti [42]. This leads to sample values $\varepsilon_{j\ell}(n) = \frac{C_j}{n} - C_{j\ell}, : \ell \in \{1 \ldots K\}$ of the multitenancy overhead at server j.

All the sample values collected by servers in $\mathcal{S}_n$ are passed to the hypervisor, the entity responsible for management of cloud resources. The hypervisor then aggregates these values to a sample mean estimate as follows:

$$\tilde{\varepsilon}(n) = \frac{1}{K|\mathcal{S}_n|} \sum_{j \in \mathcal{S}_n} \sum_{\ell=1}^{K} \varepsilon_{j\ell}(n) \tag{4.2}$$

This quantity serves as an estimate of the expected overhead for servers with n tasks, and it will be used in migration decisions. In particular, since the actual value of processing share that each VM enjoys is unknown (due to the unknown parameter $\varepsilon(n)$), all decision making regarding migrations is performed based on the estimated share for a server j with n coexisting tasks, given by $\tilde{C}_j^n = \frac{C_j}{n} - \tilde{\varepsilon}(n)$.

4.3.5 Migration

During its lifetime, a task may migrate to a new server. The task carries with it its progress indicator and its associated data volume. Consider task i migrating from server k to k' at time t_0. In order to resume execution at k' from the point it was interrupted, the entire volume of accompanying data needs to be moved to the new server. This data transfer will take place over the overlay link of capacity $W_{kk'}(t_0)$ and incurs a communication cost (delay) equal to $\frac{d_i(x_i(t_0))}{W_{kk'}(t_0)}$. Note that migrations from or to a local cloud experience the overlay link delay L. A task may migrate several times to different servers during the course of execution. Typically, each migration incurs additional delay, namely the setup time of the new VM. However, in contemporary cloud facilities this can be considered negligible.

The need for task migration arises from the continual evolution of the system. A migration should be performed, whenever it improves execution time. At any time we consider a possible migration, we need to compare two metrics:

(a) the anticipated execution time at the current server

(b) the expected completion time at the new server plus the time required to transfer the data volume from the current to the new server.

Suppose that task i resides and is executed at server k at decision time t_0. Its residual processing requirement is $b_i(t_0) = B_i(1 - x_i(t_0))$ and the amount of accompanying data is $d_i(x_i(t_0))$. Let there be n_k active tasks on server k. This means that the true available processing capacity for the task $C_k^{n_k}$ is given by (4.1). and the remaining processing time for the non-migration case (a) is:

$$D_i(k) = \frac{b_i(t_0)}{C_k^{n_k}} \tag{4.3}$$

Next, we consider whether migrating to server k' will lead to shorter task execution time (case (b)). Let there be $n_{k'}$ tasks on server k'. If task i moves there, it will receive a processing share of $\tilde{C}_{k'}^{n_{k'}+1} = \frac{C_{k'}}{n_{k'}+1} - \tilde{\varepsilon}(n_{k'}+1)$. Let $W_{kk'}(t_0)$ be the available bandwidth of the link between server k and k'. The estimated execution time at k' is:

$$\hat{D}_i(k,k') = \frac{d_i(x_i(t_0))}{W_{kk'}(t_0)} + \frac{b_i(t_0)}{\tilde{C}_{k'}^{n_{k'}+1}} \tag{4.4}$$

The hat notation will be used to refer to migration case (b), e.g. $\hat{D}_i(k,k')$ is the expected execution time of task i if it migrates from server k to k'. Notice that a migration is beneficial only if the new server can provide significantly improved performance so as to make up for migration delay.

4.3.6 Mobility

In our setting, the mobility pattern of a user i can be represented as a sequence of pairs $\{j_i(t), r_{ij}(t)\}$. For each timeslot t, the first element describes the current point of wireless access and the second one is the achievable link rate between user i and local cloud j. In order to simplify notation, we assume that each user generates one task and hence we use the same index i for both a user and his task.

As the user moves from one BS to another a different subset of servers is directly accessible, the one comprising the local cloud of the corresponding BS. In addition, the delay to each server varies with time due to the varying capacity of the wireless link. The state of communication links is not known a priori, but only at each timeslot just before a migration decision is made. Intuitively, as the task proceeds to completion, the migration strategy should favor the servers of local clouds within range of the mobile user.

The link capacity affects the duration of the download phase. Since we are interested in the total lifetime of a task, we extend our migration model to capture also the impact of mobility. A task is *completed* once its data have been transferred to the mobile user. For this purpose, we need to include in the previously defined metrics the additional term of the time required to move the final data from the server to the user. For the no migration case (a) the remaining lifetime of task i at the current server k becomes:

$$T_i(k) = D_i(k) + \frac{d_i(1)}{r_{ki}(t_0)}, \tag{4.5}$$

where $D_i(k)$ is given by (4.3) and $d_i(1)$ is the volume of the accompanying data once the execution is completed, i.e. for $x_i(t) = 1$. Notice that for the calculation we use the instantaneous value of link capacity, $r_{ki}(t_0)$.

In order to decide whether a migration to server k' (may be either local or back-end) is beneficial, we derive the following metric for migration case (b):

$$\hat{T}_i(k,k') = \hat{D}_i(k,k') + \frac{d_i(1)}{r_{k'i}(t_0)} \tag{4.6}$$

where $\hat{D}_i(k,k')$ is given by (4.4).

4.3.7 Energy consumption

Besides reducing execution time, task offloading can also provide significant energy savings to mobile devices. In order to execute task i at a remote server though, all the required data have to be uploaded to the cloud through the corresponding BS, say j. The upload phase introduces an energy cost that depends on the capacity $\gamma_{ij}(t)$ of the time-varying wireless access link at timeslot t and the data footprint $d_i(t)$ of the task. The exact energy cost of the upload phase is $\hat{E}_i(t) = f(\gamma_{ij}(t), d_i(t))$, where $f(\cdot)$ is a function characteristic of the transceiver of the mobile device. Generally though $f(\cdot)$ is decreasing in channel capacity and increasing in the data footprint.

Let E_i denote the energy required for local execution of task i. Then, task i should be offloaded at timeslot t if and only if $\hat{E}_i(t) < E_i$. Given that a user is aware of the data footprint evolution of his tasks and that the state of the wireless access link does not change within the time of interest, the optimal offloading strategy for each task can be easily calculated based only on local information.

Notice that energy-aware offloading from mobile users does not affect the strategy of the cloud provider. However, energy consumption of the cloud infrastructure itself and the resulting cost are major concerns of any cloud provider. Generally, the energy consumption of a cloud server is an increasing function of its load. In Fig. 4.2 we depict power draw as a function of load for a typical cloud rack server. We observe that it can be accurately approximated by a linear function of slope $0 < \alpha = \frac{\Delta P}{\Delta L} < 1$. In addition, power draw at idle state is 51 Watts, i.e 25% of that at maximum load. Thus, the specific server is characterised by the following power draw function:

$$p(L) = 0.155L + 50 \tag{4.7}$$

Obviously, the amount of energy consumed by a single server within a period of t hours is given by:

$$E(L,t) = p(L)t. \tag{4.8}$$
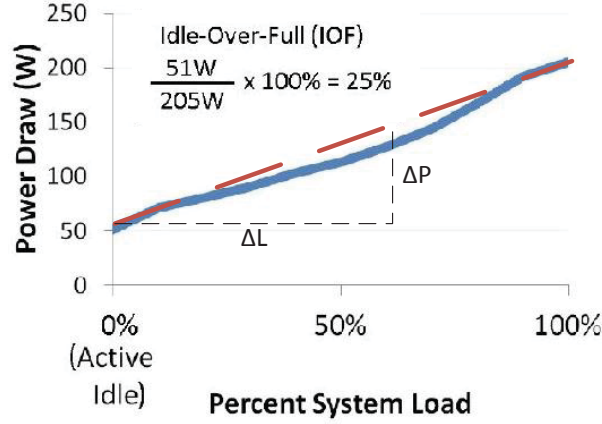
Figure 4.2: Power consumption of a typical cloud rack server as a function of its load (source: [60])

4.4 Algorithms for Efficient Task Migration in the Cloud

Our objective is to derive the migration policy μ that minimizes the average lifetime of the active tasks out of the set of available migrations $\mathcal{M}$ within a finite horizon H, i.e.

$$\min_{\mu \in \mathcal{M}} \quad \sum_{t=0}^{H} \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{A}_k(t)} T_i(\mu) \tag{4.9}$$

However, the cloud is a dynamically changing environment, and its performance is dependent on a number of uncontrollable and unpredictable parameters, such as task arrivals and completions, wireless link capacity variations and the effect of multi-tenancy. Thus, deriving the optimal migration strategy is NP-hard as a generalization of the online scheduling problem [13]. With this in mind, we propose three online migration mechanisms of different philosophy.

We follow a discrete time approach, based on the assumption that our view of the system status is updated once in each epoch period. A period of 5 minutes is typical in commercial virtualization platforms [40]. At the beginning of each epoch, control messages are circulated within the cloud, including the number of tasks running on each server and an estimate of the multi-tenancy overhead parameter $\tilde{\varepsilon}(n)$. Using this information the estimated processing share per task $\tilde{C}_k^{n_k}$ at any server k can be computed.

Initially, for ease of presentation we focus on what happens within the cloud premises, i.e. we ignore the need for downloading the final results and propose online migration algorithms that reduce the average task execution time. The algorithms are based on a greedy approach that provides the best available solution for the current state of the system. Next, we enhance the proposed algorithms to capture the impact of mobility.

4.4.1 Cloud–wide task Migration

Consider the *fully coordinated scenario*, where the cloud provider has to solve (4.9) by deciding which tasks to migrate. An important issue when scheduling multiple tasks on the cloud is that they compete for the same resources. Consider a task under tentative migration from its current server to a new one. The migration affects the performance of all tasks running at the current and the new server. A migration should be considered only if it is beneficial for *the system as a whole*. In this direction, we propose an online migration mechanism that considers the impact of a migration on all affected tasks.

Under processor sharing, each task receives a portion of processing capacity inversely proportional to the number of tasks running on the server. Thus, each migration improves the performance of collocated tasks in the host server, since it reduces competition. Fewer tasks have to share the available processing capacity, while the coexistence overhead also reduces. On the other hand, the arrival of a new task at the destination server would lead to more tasks sharing the same processing capacity, while the multi-tenancy overhead would also increase. Finally, the residual execution time of the task under migration may increase or decrease, depending on the load at the host and the destination server and the capacity of the interconnecting link.

At each decision epoch, all the active tasks located at any server are candidates for migration to any other server. We focus on task i currently hosted by server k and consider the impact of its migration to server k'. Initially, the remaining execution time of the involved tasks is calculated for the case of no migration. This is given by:

$$D_{tot}(k,k') = \sum_{l \in \mathcal{A}_k} D_l(k) + \sum_{l \in \mathcal{A}_{k'}} D_l(k'). \tag{4.10}$$

Both terms come from (4.3) and capture the execution time at each of the servers k,k' if no migration occurs.

Next, the remaining execution time is estimated for the case of migration to server k':

$$\hat{D}_{tot}(i,k,k') = \sum_{l \in \mathcal{A}_k \setminus \{i\}} D_l(k) + \hat{D}_i(k,k') + \sum_{l \in \mathcal{A}_{k'}} D_l(k'). \tag{4.11}$$

The first term corresponds to the improved execution time at the host server and is given by (4.3) for $n_k = n_k - 1$ and the second one is the expected execution time of the migrating task described by (4.4). The third term captures the degraded performance at the destination server k' and is given by (4.3) for $n_k = n_k + 1$. The difference in total execution time $D_{tot}(k,k') - \hat{D}_{tot}(i,k,k')$ may be either positive or negative. A positive value indicates that the migration of task i to server k' is beneficial for the system.

Since a migration to any other server is possible, this calculation is carried out for each possible destination server k'. Finally, out of all the possible migrations the one of maximum positive gain is performed. This process is repeated until no more migrations of positive gain can be found.

This online mechanism is executed in every epoch. Within an epoch new tasks arrive while others

complete execution. In general, our algorithm picks tasks from overloaded servers and moves them to underutilized ones. However, the selection of the task to migrate is performed according to the following guidelines jointly:

- prioritize migrations of tasks of increasing data volume pattern, since the migration cost of any such task increases as its processing evolves.

- migrate tasks of significant residual processing burden. The benefit of a migration is a function of the remaining processing time of the task. For example migrating a task that is close to completion, may not be beneficial even if the destination server is idle. On the other hand, a task of substantial remaining processing time can exploit the available capacity at the destination server more efficiently.

- preferably migrate tasks that experience significant multi-tenancy cost. Although this cost is generally increasing in the number of collocated tasks, its exact impact depends also on the type of collocated tasks.

Remark 4.4.1. *Our approach requires the processing burden of each task to be known (see (4.3),(4.4)). If such information is not available our model can be easily modified by replacing $b_i() = \bar{b}$, where $\bar{b}$ is the average burden of tasks. In this case the second guideline would not be applicable.*

4.4.2 Server-initiated task migration

The centralized nature and high complexity of the cloud-wide approach calls for distributed migration mechanisms that require less information to be circulated throughout the system. In this direction, we propose a server-initiated approach that enables each server to autonomously select which of its active tasks should migrate and where. Each server periodically checks whether the execution time of its active tasks can be improved through a migration to a new server. For this purpose, the anticipated gain for each possible migration to any new server needs to be estimated. The anticipated performance gain of migrating task i located at server k to server k' is defined a:s

$$D_{tot}(k) - \hat{D}_{tot}(i,k,k') = \sum_{l \in \mathcal{A}_k} D_l(k) - \sum_{l \in \mathcal{A}_k \setminus \{i\}} D_l(k) - \hat{D}_i(k,k')$$

In this case, we consider the total reduction of execution time of the tasks hosted by server k. The first term is the total execution time for the case of no migration, while the second term captures the execution time of the tasks remaining at k for $n_k = n_k - 1$ and the third refers to the migrating task i. Since migrations are initiated by the host server, the impact of the migration on the tasks located at the destination server is unknown and hence not considered. In the end, the migration of maximum gain is performed. This is repeated until no more migrations of positive estimated gain can be found.

The main characteristic of this algorithm is that it requires no synchronization among servers. Each one decides autonomously when to check for beneficial migrations. This exploration phase
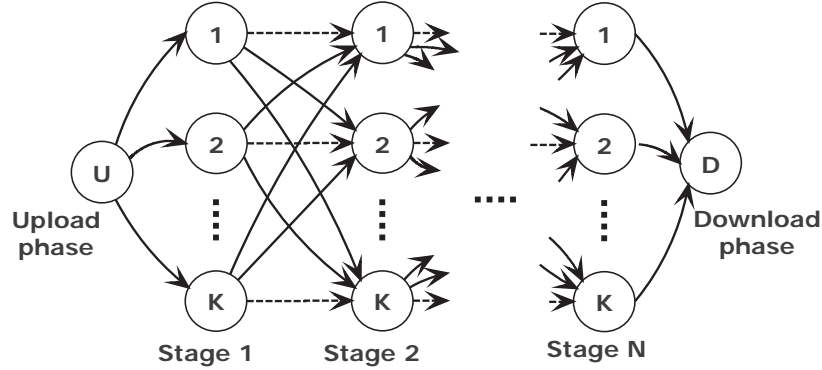
Figure 4.3: Transition graph for the single task migrating within the cloud.

could be triggered whenever a server considers it is overloaded, compared e.g. to the average server load.

4.4.3 Task-initiated migration

In contemporary cloud systems migration related decisions are made by the cloud provider. However, one could envision a scenario where each task/user autonomously decides its migration strategy towards minimizing its own execution time. Thus, we consider the case where a user may offload its task to one of multiple available cloud providers/servers, but is only aware of the advertised capacity of each.

We follow a discrete time approach and assume that within each epoch of duration τ the system state does not change. Due to multitenancy the actual capacity of each server k at timeslot t is a random variable $C_k(t)$ taking values in $[0, C_k]$. Since this randomness is uncontrollable, the actual capacity can be considered independent across servers, while for a given server it is generally correlated across time. The actual capacity is only revealed once a task migrates to a server. In addition, each migration from the current server introduces migration delay. The optimal migration strategy of the user is the one that minimizes its own execution time. This problem falls within the class of restless bandits with switching costs which has been shown to be PSPACE-Hard [43]. In our case, each arm corresponds to a server, the corresponding reward at timeslot t is $C_k(t) * \tau$, while the switching cost from arm k to k' corresponds to the migration cost $\frac{d_i(x_i(t))}{W_{kk'}(t)}$.

However, under the assumption that the actual capacity of each server is i.i.d. across time, the optimal migration strategy reduces to the following. At each timeslot t, task i may autonomously check whether a tentative migration from its current server k to server k' would decrease its estimated execution time , i.e. whether $\hat{D}_i(k,k') < D_i(k)$ and performs the migration of maximum gain.

Next, we consider the off-line deterministic case, where the exact capacities of servers and links at each stage are known a priori. The solution to this problem provides a lower bound of the execution time. The optimal migration strategy of a task is the sequence of servers that the task should traverse. The state space of the problem is depicted in Fig. 4.3.

Let each state denote the server that hosts the task of interest. The execution phase is divided into decision stages (epochs). In each epoch the user needs to decide whether the task should remain at the current server or a migration should be performed. We use dashed lines to denote the no–migration case that incurs zero cost and solid lines for the actual migrations. At each stage t, if task i remains at its current server k execution proceeds by $C_k(t) * \tau$, whereas in case of migration to a new server k' execution proceeds by $C_{k'}(t) * (\tau - \frac{d_i(x_i(t))}{W_{kk'}(t)})$. The optimal migration strategy corresponds to the shortest path and can be found through dynamic programming [44]. The search horizon is dictated by the total processing burden B_i, i.e. the search is terminated once the shortest path that meets the processing requirement is found.

Remark 4.4.2. *The proposed approaches require the same information per task but the first one requires a system wide view, while the second one cares only for the tasks running on the server and the last refers to a single task.*

4.4.4 The impact of mobility on migration decisions

Our analysis up to this point has been focused on reducing the execution time of tasks. In mobile scenarios though, the time required to download the results from the cloud is a significant portion of the task lifetime. Given also that the links between a server and a user are characterized by diverse and time-varying capacity, bringing mobility in the migration selection strategy is of utmost importance. In this direction, we modify the proposed algorithms so as to capture the incurred download cost, by using the estimated task lifetime given by (4.5) and (4.6) in place of (4.3) and (4.4) respectively. Hence, as a task moves towards completion (i.e. it has small remaining processing burden), it prefers servers that are in range of the user, wherever the latter might be. For ease of presentation we have not included the upload cost, since this has to be considered only once, namely in the initial assignment of a task to a server. Subsequent migration decisions are not affected by upload time.

4.4.5 VM migration mechanisms for energy efficiency

The main objective of the proposed algorithms was to improve cloud performance. Here, we explore an alternative use of VM migrations towards a more energy efficient cloud. In this direction, we devise VM migration strategies that minimize both energy consumption of the cloud infrastructure and the corresponding electricity cost.

We assume that energy consumption of each server is given by (4.7). In this case, it is straightforward to show that the most energy efficient strategy is to consolidate VMs in the fewest possible servers and turn off the idle ones. Interestingly, this strategy is in contrast to the objective of minimizing execution time, since extreme consolidation introduces also the problem of multitenancy. In order to deal with this tradeoff, we suggest that the cloud provider should consolidate VMs up to the point that no performance SLA is violated. In particular, we propose a two timescale approach. At a slow timescale, the operator makes a rough estimation on the number of servers that have to be activated

based on historical data. At a faster timescale, VM migrations are exploited to counteract the effects of multitenancy, according to one of the previously described algorithms.

Based on such an approach the required number of active servers can be determined. Next, we attempt to answer the question: which servers should be activated? Although all servers are equivalent in terms of performance, electricity cost is not the same due to variation of electricity prices from place to place. Given that cloud facilities consist of geographically dispersed datacenters all over the world, VM migrations can provide additional financial benefits by moving load to places of low cost. Given that electricity prices for the following day are generally known, this problem can be easily solved in an offline manner.

4.5   Numerical Evaluation

In order to compare the performance of the proposed schemes, we use an event driven simulator for a cloud system of 50 servers, each of capacity $C \in [0.1, 2]$ Tera–flops, interconnected through wireline communication links of mean available link capacity $W \in [1, 10]$ Mbps. The link and processing capacities are assumed to be i.i.d. across different slots.

New tasks arrive at each server i according to an inhomogeneous arrival process of parameter $\lambda_i(t)$ arrivals/sec, given by a gaussian distribution of mean 0.05 and standard deviation 0.015. The processing burden follows a heavy tailed pareto-like probability density function (pdf) with $P[B_i > x] = \min\{1, (1/x)^{1.25}\}$, where x is in teraflops, which is typical for CPU intensive tasks [45].

We consider tasks of increasing or decreasing data footprint evolving according to the linear model $d_i(x_i) = \min\{0, \alpha x_i + \beta\}$ with $\alpha \sim U(-10, 10)$, $\beta \sim U(0, 100)$ for the increasing and $\beta' \sim U(10, 1000)$ for the decreasing ones (in MBytes). The multi-tenancy cost due to VMs consolidated onto a single server is modeled as a gaussian random variable of mean $\mu = \frac{(k-1)}{2k}$. For comparison purposes, we depict also in each figure the performance of the no migration strategy and of a one-shot placement scheme, in order to quantify the importance of migrations. Placement follows the logic of our cloud wide approach, but now the server that will host the task is selected only once, when the task arrives at the system. The depicted values are averages over 100 instances.

Initially, we investigate the population of tasks N hosted by each server, which indicates the load balancing behaviour of each algorithm. Since N is a random variable, we depict in Fig. 4.4 its cumulative distribution function (cdf), i.e. the probability $P[N \leq k] \ \forall k \in \mathbb{N}$ for each algorithm. The slopes of the curves indicate how balanced the cloud is. In the cloud-wide approach a server hosts at most 3 VMs. Instead, the probability of having more than 3 VMs running on a server is 10% for the server-initiated approach and 25% for the task-initiated one. In the latter case the probability of having more than 20 tenants in a server is non-negligible ($\sim 4\%$); compared to the no migration case though the load is more balanced.

We depict in Fig. 4.5(a) the exact distribution of tasks on the servers through the probability mass function of N. We see that the cloud-wide migration strategy leads to a more balanced network, where in most of the cases the number of tasks N is from 1 to 3. Moving to the less coordinated strategies,
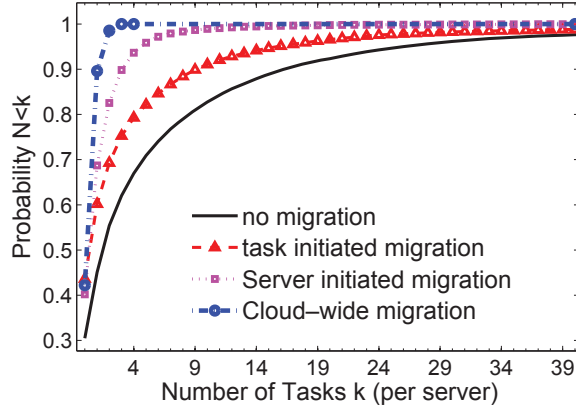
Figure 4.4: cdf of the number of tasks per server

(a) pmf of the number of tasks per server

(b) pmf for different link rates
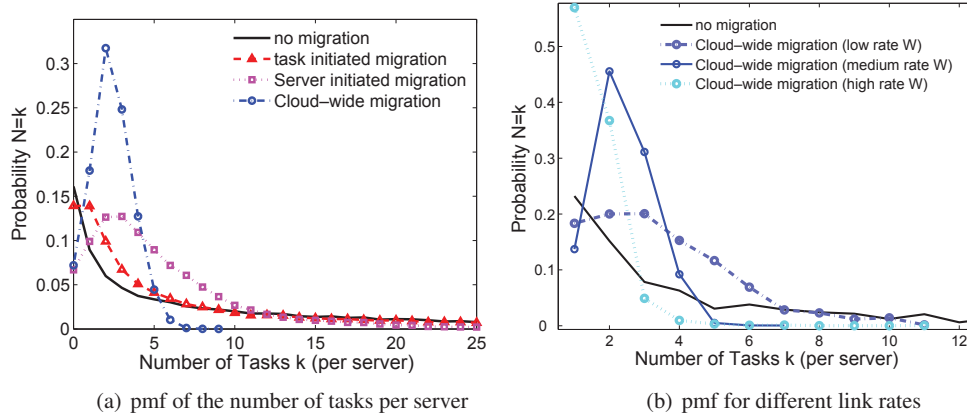
Figure 4.5: Load balancing behaviour of the proposed algorithms

the pmfs broaden. For example, in the user–initiated scheme the probability that a server is empty is significant, while even for $N > 20$ the probability is non-negligible.

Next, we consider how the distribution of tasks is affected by the link capacity of the cloud interconnection links. Thus, we depict in Fig. 4.5(b) the pmf of the cloud wide approach for three different scenarios, one where the cloud servers are interconnected by high/medium/low capacity links. As the link capacity decreases the pmf broadens indicating that it is not always optimal to perform a perfect load balancing. This is justified by the fact that in low link capacity regime, the migration cost becomes significant and hence it is preferable to be slightly unbalanced.

In Fig. 4.6(a) we quantify the impact of per server processing capacity on the average lifetime of the tasks. As expected the performance degrades as we move from the centralized approach that has system–wide information to the decentralized ones that are based only on local information. Increas-

(a) Available processing capacity

(b) Data footprint

(c) Available link capacity
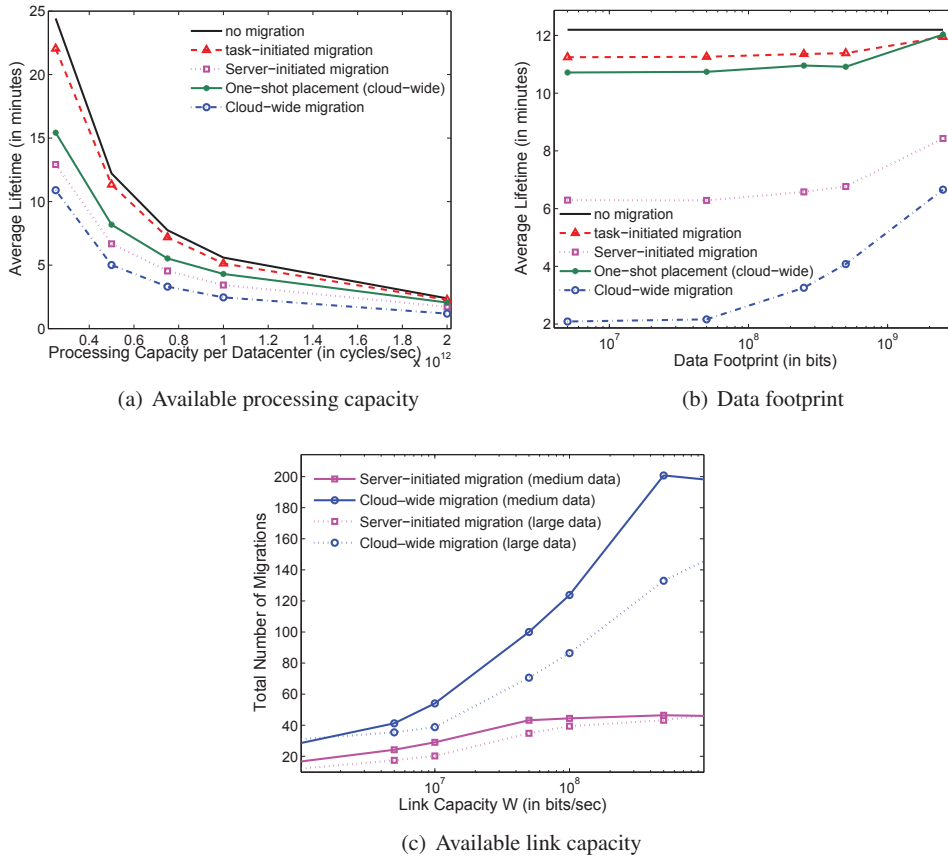
Figure 4.6: The impact of system/task parameters on cloud performance

ing server capacity causes the performance gap of the proposed schemes to diminish, indicating that careful migration decisions are most important when the cloud is overcommitted. Interestingly it is only the task-initiated approach that performs worse than the cloud wide placement approach. In the task-initiated approach we observed that although the average lifetime is not significantly better than the no migration case, the individual gain of the tasks that migrate is significant.

In Fig. 4.6(b) we depict the impact of data footprint on the average lifetime of the tasks. As the mean footprint increases the performance gap between the proposed algorithms decreases, since the "heavier" the tasks the greater is the migration cost. Finally, we consider how often migrations are performed by each algorithm in Fig. 4.6(c) for two different scenarios. For tasks of medium data footprint we observe that significantly more migrations are performed, compared to a system serving data-intensive tasks.

In order to stress the impact of mobility in migration decisions, we depict in Fig. 4.7 the lifetime of a task that is initially uploaded to a local cloud by a mobile user through its 3G access. We consider a task of increasing data footprint that can be executed either at the local (L) or the back-end cloud (B). We depict the task lifetime for the case of a) no migration, b) a strategy that does not consider
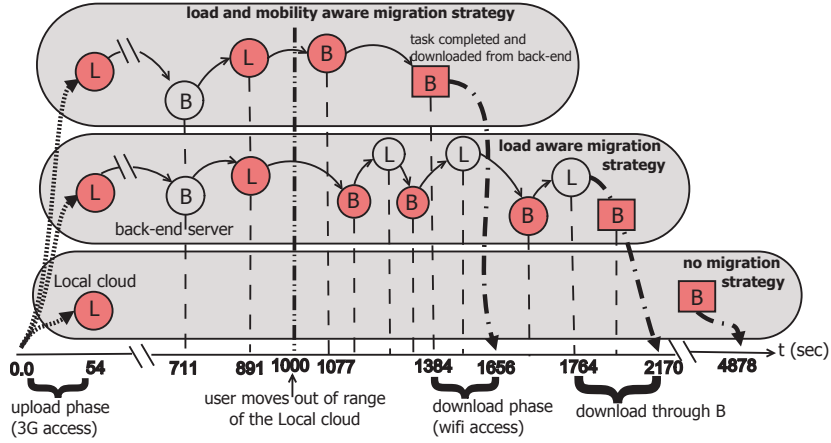
Figure 4.7: The lifetime of a mobile task under different migration strategies

migration cost and download time and c) the proposed task-initiated migration strategy that is mobility aware. We denote with shadow the closest to the user cloud facility in terms of communication delay. The no-migration strategy executes the task at the local cloud and hence exhibits the worst performance. Initially, the migration cost is negligible since the data footprint of the task is small. Thus, as long as the user is in range of the local cloud and migrations are costless, the other two migration strategies perform identically. Once the user moves out of range of the BS hosting the local cloud to a place that is closer to the back-end (WiFi access), the mobility-aware approach moves the task there. In contrast, the load-aware one performs several migrations to the least loaded server, in an attempt to exploit the best option in terms of available processing capacity, without considering though the increasing cost of each migration and the additional cost of downloading the final data from a distant server. Hence, it is outperformed by the mobility aware one.

4.6   Conclusion

In this chapter, we investigated the option of task offloading as a means of improving processing capabilities and lifetime of a mobile device. To this end, we developed migration policies that opportunistically exploit available processing capacity at cloud servers. Our techniques explicitly take into account user mobility by considering the download time of the data emanated from task execution. This essentially biases the migration decision in favor of servers within range of the mobile user.

In contrast to the migration policies currently applied in commercial virtualization platforms like vSphere [40], we demonstrate that migrations should be performed according to the estimated performance and not based on the estimated resource demands of each task/VM. Our performance analysis verifies the benefits arising from migration schemes that consider both migration cost and mobility of users in decision making.

Chapter 5

Energy Efficiency at the Wireless Device Access Level: The Impact of Energy Constraints on Medium Access

In this chapter, we move from the realm of Mobile Cloud Computing to the setting of energy-constrained medium access. Contemporary mobile devices are battery powered and operate on a tight energy budget. Besides, the scarcity of bandwidth resources leads to strong competition for the medium. In an attempt to capture these facts, we consider random medium access schemes for devices that support sleep modes, i.e. turning off electronic compartments for energy saving.

Due to hardware limitations, sleep mode transitions cannot occur at the medium access timescale, but only at a slower timescale. Each terminal can choose when to turn on/off and its probability to transmit on an arbitrary slot. Thus, we develop a two level model, consisting of a fast timescale for transmission scheduling and a slower timescale for the sleep mode transitions. We take a game theoretic approach to model the user interactions and show that the energy constraints modify the medium access problem significantly and lead to reduced price of anarchy. Our findings give valuable insights on the energy–throughput tradeoff for contention based systems.

5.1 Introduction to Energy Efficient MAC

As mobile communications become part of our everyday lives, new challenges for the system designers come to the foreground. First of all, the scarcity of bandwidth resources leads to extreme competition for the medium. Besides, the total energy dissipation by communication devices has been shown to amount to a significant portion of a nation's power profile, motivating efforts of per device energy economy. In an attempt to minimize their energy footprint and/or maximize the battery lifetime, existing wireless devices support radio *sleep modes*.

A generic wireless terminal consists of several circuit building blocks with the RF transceiver (radio) contributing significantly to the overall energy consumption. The RF transceiver itself con-

Table 5.1: Switching time and energy consumption of a CC2420 radio

Power mode	Switching time(ms)	Switching Energy (μJ)	Current Consumption (μA)
Tx	0	0	10000
idle	0.1	1.035	426
power down	1.2	42.3	40
deep sleep	2.4	85.7	0.02

sists of four subblocks. The transmit block that is responsible for modulation and up-conversion (i.e. transforms the baseband signal to RF), the receive block dedicated to the down-conversion and demodulation, the local oscillator that generates the required carrier frequency, and the power amplifier that amplifies the signal for transmission. Existing wireless devices support radio sleep modes that turn off specific subblocks, to minimize their energy consumption while inactive. For example, as shown in Table 5.1 the CC2420 transceiver ([61]) provides three different low power modes. In the deepest sleep mode, both the oscillator and the voltage regulator are turned off, providing hence the lowest current draw. However, this comes at the cost of the highest switching energy cost and the longest switching latency. On the other hand, the idle mode provides a quick and energy inexpensive transition back to the active state, but at the cost of higher current draw and consequently higher current consumption.

To address this tradeoff, the authors of [62] propose a scheme to dynamically adjust the power mode according to the traffic conditions in the network. They show that in a low traffic scenario, deep sleep should be preferred since most of the time the nodes tend to be inactive. In a high traffic setting though, a "lighter" sleep mode would be preferable, because frequent mode transitions incur high delay and energy costs, overshadowing the energy saving due to the low current draw. In a similar setting, [63] considers optimal scheduling of the sleep periods for the scenario of a mobile receiving data from a base station. The objective is to derive the sleeping strategy that balances the energy cost of frequent waking up to check for new packets and the retrieval latency.

In this direction, several energy aware MAC protocols have been proposed, either centralized or distributed ones, to resolve contention. However, most of them rely on the willingness of the nodes to comply with the protocol rules. Hence, they are vulnerable to selfish users that may deviate from the protocol in order to improve their own performance. Game theory comes as the ideal tool to model interactions among self–interested entities competing for common resources and it has also been considered recently for medium access.

In [64], the Nash Equilibrium Points (NEP) in a slotted ALOHA system of selfish nodes with specific quality-of-service requirements are studied. It has been observed that usually selfish behavior in medium access leads to suboptimal performance. For example, a prisoners dilemma phenomenon arises among selfish nodes using the generalized slotted Aloha protocols of [65]. A decrease in system throughput, especially when the workload increases due to the selfish behavior of nodes, has been observed in [66], [67].

In [70] a random medium access scheme is cast as a non-cooperative game where each node wants

to maximize an expression of its medium access probability and contention. The proposed scheme adapts to a continuous contention measure, which is estimated by observing consecutive idle slots between transmissions, instead of the binary feedback of collision detection. However, it requires that each node is aware of contention for the decision making and no energy constraints are assumed.

The interplay of medium access contention and energy consumption was considered recently in [68]. In particular, a scenario of users selecting their back-off control parameters based on the measured collision rate and their power consumption was considered. Each user attempts to balance the utility acquired by transmitting and the disutility caused to him due to the induced energy consumption. The existence, uniqueness and stability of the equilibrium points of the arising game were investigated. In a similar framework and in an attempt to mitigate the effects of selfishness, the authors of [69] study the problem of minimizing the energy consumption for given throughput demands for a contention MAC. They show that whenever the demands are feasible, there exist exactly two Nash equilibrium points and derive a greedy mechanism that always converges to the best one.

In this work we introduce an additional level of decision making capturing the ON-OFF strategy of the terminals over the classic ALOHA game. We model contention for the medium as a game, where users with specific energy constraints *select both the proportion of time that they sleep and their medium access probabilities*. To the best of our knowledge, this is the first work that addresses the interplay between contention and energy consumption *for systems that support sleep modes*. The contributions of this work can be summarized in the following:

- We characterize the throughput optimal strategy under energy constraints, which differs from traditional Aloha and serves as a reference point. We also provide a distributed counterpart strategy, which focuses on fairness.

- We formulate contention as a non-cooperative game and show that it has a unique NEP and bounded Price of Anarchy (PoA). Contrary to the game with no energy constraints, where the PoA is infinite, we find that energy constraints reduce the negative impact of selfish behaviour.

- Based on the rationality of the users we derive a modified strategy, which allows the users to observe competition. This policy is more efficient but has multiple NEPs.

The main characteristics of slotted ALOHA are also apparent in most contemporary contention-based systems, such as the IEEE 802.11. For example, all these systems exhibit a certain amount of inherent inefficiency; the total throughput in a common channel breaks down significantly, as the number of users and the message burstiness increase (see [71]). Despite its limited applicability and mainly due to its simplicity, Aloha remains a tractable insightful tool for studying such systems. Consequently, our results provide insights for other contention based systems.

5.2 System Model

We consider a communication scenario of $N = |\mathcal{N}|$ mobile terminals transmitting to a common destination (e.g. uplink to a Base Station). Time is slotted and within each timeslot each user may select
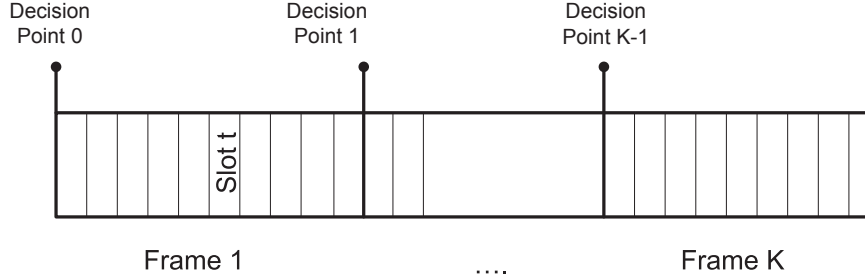
Figure 5.1: The structure of a superframe

either to transmit or to stay silent. Medium access is performed probabilistically, according to a slotted Aloha protocol, where a collision occurs whenever two or more terminals transmit concurrently. Each terminal has always packets at its buffer for transmission (i.e. saturated queue), but it has limited energy resources. Each device i is characterized by an energy budget $\tilde{e}_i$, representing either its available battery power or the maximum energy it is willing to expend. In order to save energy it can turn into a sleep mode, where most of the circuits are turned off. For analytical tractability we assume that each terminal may be in one out of two possible operation modes, ON or OFF.

In general, a mode transition incurs significant energy and time (delay) costs. Besides, due to hardware limitations the time required for a mode transition is of the order of msec, much larger than the duration of a timeslot. Consequently, transition at the timeslot level is neither feasible nor desirable. Hence, we introduce a new timescale (we call it frame) where the mode switching takes place. Several timeslots constitute a frame. The beginning of each frame is a decision point, where a node may change its operation mode. Within a frame, the nodes keep their mode fixed and may access the medium randomly with probability p. This is also assumed fixed on a per frame basis. For decision reasons and without affecting the operation of the system, we introduce yet another timescale, that of the super frame, where an arbitrary number of frames, say K, forms a superframe (Fig. 5.1)

A binary vector $\boldsymbol{q}_i(j) = \{0,1\}^K$, represents the ON or OFF state of user i on superframe j, and p_i its access probability (i.e. the access probability is selected only once per superframe). In practice the mobile terminals are autonomous and due to the limited knowledge available node level, synchronization of sleep modes is a difficult task. On the other hand, probabilistic ON-OFF has been deemed a feasible strategy ([72]). *Thus, we focus on a probabilistic version of the aforementioned problem.* Each user i is characterized by a probability of being ON, denoted with q_i and a medium access probability p_i. In matrix notation the strategy space can be written as $\mathcal{I} = \{\boldsymbol{p}, \boldsymbol{q}\}$, with $\boldsymbol{p} = [p_1, p_2, \ldots p_N]$ and $\boldsymbol{q} = [q_1, q_2, \ldots q_N]$.

If we consider the probability of user i successfully transmitting a packet in an arbitrary slot we can calculate throughput as

$$\bar{T}_i(\boldsymbol{p}, \boldsymbol{q}) = p_i q_i \prod_{j \in \mathcal{N} \setminus i} (1 - p_j q_j) \stackrel{p_i q_i \neq 1}{=} \frac{p_i q_i}{1 - p_i q_i} \prod_{j \in \mathcal{N}} (1 - p_j q_j) \tag{5.1}$$

66

Notice that the throughput of user i is an increasing function of p_i and q_i, but decreasing in the number of terminals N contending for the medium. The latter is in compliance with the classic ALOHA, but in practice also holds for CSMA/CA protocols.

The energy cost of user i is a random variable with a mean value of $\bar{E}_i(p_i, q_i) = q_i(c_1 + c_2 p_i)$, where c_1 is the energy consumption while ON and c_2 the additional cost imposed by the transmission. Obviously, in order to transmit, the node has to be ON. Here, we do not consider the energy consumption of the transition itself.

5.3 The Impact of Constrained Energy Resources on System Throughput

First, we would like to find the ON–OFF and the medium access probabilities that maximize the collision–free utilization of the medium, and consequently the throughput of the system. This can be formally expressed as the following optimization problem:

$$
\begin{aligned}
\underset{\mathcal{I}=\{\boldsymbol{p}, \boldsymbol{q}\}}{\text{maximize}} \quad & \sum_{i=1}^{N} \bar{T}_i(\boldsymbol{p}, \boldsymbol{q}) \\
\text{s.t.} \quad & \bar{E}_i(p_i, q_i) \leq \tilde{e}_i \quad \forall i \\
& \{p_i, q_i\} \in [0, 1]^2 \quad \forall i
\end{aligned}
\tag{5.2}
$$

The intuitive explanation of the energy constraint is the fact that each battery cycle provides $\tilde{e}_i$ resource and thus, the maximum energy constraint is directly mapped to a minimum recharge time for the mobile. Throughout this chapter, and without loss of generality, we assume that the users are ordered in decreasing energy budget, i.e. $\tilde{e}_1 \geq \tilde{e}_2 \geq \ldots \geq \tilde{e}_N$.

5.3.1 Throughput optimal scheduling in energy constrained ALOHA with sleep modes

In the classic ALOHA setting, where no energy constraints exist, the throughput optimal strategy would be the one that eliminates contention. Thus, if we could force only a single user, say user k, to access the medium with probability $p_k = 1$ in each frame, we would achieve the maximum total throughput. In our scenario though, due to the energy constraints, the users may not be able to stay continuously ON (i.e $q_k = 1$) or to transmit with $p_k = 1$. Then, what is the best way to exploit the available energy resources? For each user we need to find the portion of energy to spend for staying ON during the frames and the portion used for transmitting while being ON.

Lemma 5.3.1. *Out of all the throughput optimal strategies the most energy efficient ones are of the form $\mathcal{I}^* = \{\boldsymbol{1}, \boldsymbol{a}\}$ with $\boldsymbol{1} = [1, 1, \ldots 1]$ and $\boldsymbol{a} \in [0, 1]^N$. Thus, without loss of optimality we may restrict the strategy search space only to strategies where the nodes transmit continuously inside any ON frame.*

Proof. Let $\hat{\mathcal{I}}$ be a feasible throughput optimal strategy, where $\{\hat{p}_i, \hat{q}_i\}$ the strategy of user i. From eq. 5.1 we may see that throughput depends only on the pq products. Let $a_i = \hat{p}_i \hat{q}_i$. The strategy $\mathcal{I}^* = \{\mathbf{1}, \mathbf{a}\}$ with $\mathbf{a} = [\hat{p}_1 \hat{q}_1, \hat{p}_2 \hat{q}_2, \dots \hat{p}_i \hat{q}_i]$ achieves the optimal throughput, i.e. $\bar{T}(\mathcal{I}^*) = \bar{T}(\hat{\mathcal{I}})$.

Then, we prove that $\mathcal{I}^*$ is also feasible, as the most energy efficient strategy.

Any other feasible throughput optimal strategy $\hat{\mathcal{I}}$ can be written as an expression of $\mathbf{a}$ as $\{\hat{p}_i = \frac{a_i}{a_i + \delta_i}, \hat{q}_i = a_i + \delta_i\}$, with $0 < \delta_i \le \min\{1 - a_i, \frac{\tilde{e}_i - a_i(c_1 + c_2)}{c_1}\}$. Thus, regarding the energy efficiency we have:

$$
\begin{aligned}
\bar{E}(\hat{\mathcal{I}}) &= \sum_{i=1}^{N}(a_i + \delta_i)\left(c_1 + c_2\frac{a_i}{a_i + \delta_i}\right) \\
&= \sum_{i=1}^{N}\delta_i c_1 + a_i(c_1 + c_2) = \bar{E}(\mathcal{I}^*) + c_1\sum_{i=1}^{N}\delta_i \\
&> \bar{E}(\mathcal{I}^*).
\end{aligned}
$$

This completes our proof. $\qquad\square$

Based on Lemma 5.3.1, the optimization problem described by eq. 5.2 can be simplified to an expression that depends only on $\mathbf{q}$ leading to the objective function $\bar{T}(\mathbf{q}) = \sum_{i=1}^{N} q_i \prod_{j \in \mathcal{N}\setminus i}(1 - q_j)$ and constraints $0 \le q_i \le \tilde{q}_i = \min\left\{\frac{\tilde{e}_i}{c_1 + c_2}, 1\right\}$; this can be further simplified into a problem of binary integer programming.

Lemma 5.3.2. *The optimal solution is of the form* $\mathbf{q}^* = \mathbf{b}^* diag[\tilde{q}_1, \tilde{q}_2, \dots \tilde{q}_N]$ *where* $\mathbf{b}^*$ *is a binary row vector.*

Proof. The partial derivative of the objective function is given by:

$$
\begin{aligned}
\frac{\partial \bar{T}}{\partial q_k} &= \prod_{j \in \mathcal{N}\setminus k}(1 - q_j) - \sum_{j \in \mathcal{N}\setminus k} q_j \prod_{l \in \mathcal{N}\setminus\{k,j\}}(1 - q_l) \\
&= \prod_{j \in \mathcal{N}\setminus k}(1 - q_j) - \sum_{j \in \mathcal{N}\setminus k}\frac{q_j}{1 - q_j}\prod_{l \in \mathcal{N}\setminus k}(1 - q_l) \\
&= \left(1 - \sum_{j \in \mathcal{N}\setminus k}\frac{q_j}{1 - q_j}\right)\prod_{j \in \mathcal{N}\setminus k}(1 - q_j)
\end{aligned}
$$

We have thus shown that the sign of the partial derivative with respect to the kth element depends only in the parameter $\sum_{j \in \mathcal{N}\setminus k}\frac{q_j}{1 - q_j}$. This leads to the following decision making:

$$
q_k = \begin{cases} \tilde{q}_k, & \text{if } \displaystyle\sum_{j \in \mathcal{N}\setminus k}\frac{q_j}{1 - q_j} < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}
$$

$\qquad\square$

Lemma 5.3.3. *The optimal solution* $\boldsymbol{b}^*$ *is of the form* $\boldsymbol{b}^* = [1,1,\ldots 1,0,0,\ldots 0]$.

Proof. We will prove this by contradiction. Assume that $\tilde{\boldsymbol{b}} \neq \boldsymbol{b}^*$, a vector of k ones (i.e. any vector where the ones are not placed in the first k places), is the throughput optimal binary vector. We can construct a new vector $\hat{\boldsymbol{b}} = [1,1,\ldots 1,0,0,\ldots 0]$, which has only the first k users activated and gives identical throughput, by deriving the ON-OFF vector $\hat{\boldsymbol{q}}$ from the corresponding indices of $\tilde{\boldsymbol{b}}$. To construct such a vector we need to move the rightmost 1, say from position l of the initial vector to an earlier zero position say m with $m < l$. Based on Lemma 5.3.2, we can show that by fully activating or deactivating users (i.e. meeting the constraint with equality or setting access probability to zero), we receive a new schedule $\boldsymbol{b}^* = \hat{\boldsymbol{b}}$ and the corresponding $\boldsymbol{q}^*$ of increased throughput. This leads us to a contradiction regarding the optimality of $\tilde{\boldsymbol{b}}$. $\qquad\square$

Based on the aforementioned lemmas we may derive the centralized Algorithm 4 that yields the throughput optimal probabilistic strategy and is of linear, in the number of users N, complexity. The main idea behind this algorithm is that contention may or may not be beneficial, depending on the energy constraints of the users. Namely, an additional user is useful if and only if the energy resources of the already active users are not sufficiently large, leaving thus the medium underutilized. An additional user introduces a gain due to the exploitation of the empty frames, but also a loss, due to the collisions whenever he is concurrently active within a frame with someone else. If the average gain is greater than the induced loss, it is beneficial for the system to be enabled.

Algorithm 4 Optimal probabilistic frame scheduling

1: Order users in decreasing $\tilde{e}_i$. Without loss of generality,
   we reassign the indices such that $\tilde{q}_1 \geq \tilde{q}_2 \geq \ldots \geq \tilde{q}_N$
2: $\boldsymbol{q} \leftarrow \boldsymbol{0}$
3: $j \leftarrow 1$
4: **while** $j \leq N$ **and** $\displaystyle\sum_{i=1}^{j} \frac{q_i}{1-q_i} < 1$ **do**
5: $\quad q_j \leftarrow \tilde{q}_j$
6: $\quad j \leftarrow j+1$
7: **end while**

Theorem 5.3.1. *Algorithm 4 yields the throughput optimal probabilistic strategy.*

Proof. The optimality of 4 comes directly from Lemmas 5.3.1, 5.3.2 and 5.3.3.

$\qquad\square$

Note that this algorithm gives also the throughput optimal scheduling for the constrained version of the Aloha problem, where each terminal has an individual constraint on its medium access probability p_i. This is a simplified version of the problem considered in this work, where no sleep mode support exists, i.e. $q_i = 1$.

The algorithm above promotes the k less energy constrained users ($0 \leq k \leq N$ depends on actual constraints) and suppresses the rest. It will only serve as a performance benchmark, since it is a centralized algorithm that introduces important coordination and fairness issues. In particular, it requires

extensive coordination among the users and causes extremely unfair treatment of the users of low energy budget. Thus, in the following sections we develop algorithms that capture the autonomous nature of the users and fit to the dynamic distributed environments considered here.

5.3.2 A distributed fair algorithm

In order to capture the notion of proportional fairness we substitute the original objective function with the following: $U(\boldsymbol{p},\boldsymbol{q}) = \sum_{i=1}^{N} w_i \log \bar{T}_i$. The multiplicative factor w_i can be used to balance the throughput among the users of the system at will. For example, the value $w_i = \frac{\tilde{e}_i}{\sum_{k \in \mathcal{N}} \tilde{e}_k}$ would allow us to split the throughput proportionally to the energy budget of the users. By proper reformulation, the objective function can be rewritten as $U(\boldsymbol{p},\boldsymbol{q}) = \sum_{i=1}^{N} \log \left[(p_i q_i)^{w_i} (1 - p_i q_i)^{1-w_i} \right]$. This is a separable per user function that leads to a fully distributed implementation, requiring minimal information exchange. Actually the only information required is the value of the total energy available in the terminals, namely $\sum_{k \in \mathcal{N}} \tilde{e}_k$, information that can be easily acquired. The solution to this optimization problem, in accordance to Lemma 5.3.1, is of the form $\mathcal{I}^* = \{\boldsymbol{1}, \boldsymbol{a}\}$ with $\boldsymbol{a} = [\min\{w_1, \tilde{q}_1\}, \min\{w_2, \tilde{q}_2\}, \ldots, \min\{w_N, \tilde{q}_N\}]$.

5.3.3 A modified strategy

Up to now we assumed that each user makes a decision once for his strategy and applies it forever. As a result, user k whenever active, transmits with $p_k = 1$, independently of the number of active users within a frame. Thus, whenever two or more users select to transmit within a frame they receive zero payoff, but consume energy. Based on these observations, a rational player would be expected to backoff whenever a collision is detected. Although, the terminal is not allowed to switch off in a crowded frame, due to the switching time overhead incurred, it may reduce its access probability. This way, it would avoid spending energy on useless transmission that always lead on collisions and could utilize these savings for pursuing further contention–free frames. Building on this idea we propose the following modified strategy.

Any active user attempts a transmission within the first timeslot of the current frame. If the transmission succeeds he selects a medium access probability of $p_i = 1$ and captures the whole frame successfully. Otherwise he adjusts his strategy, and reduces his transmission probability to $\tilde{p}_i$. It can be shown that this strategy always yields better throughput than the original one. The expressions for throughput and energy consumption for this case are respectively given by:

$$\bar{T}_i = q_i \left\{ (1 - \tilde{p}_i) \prod_{j \in \mathcal{N} \setminus i} (1 - q_j) + \tilde{p}_i \prod_{j \in \mathcal{N} \setminus i} (1 - \tilde{p}_j q_j) \right\} \tag{5.4}$$

$$\bar{E}_i = q_i \left\{ c_1 + c_2 \left[\tilde{p}_i + (1 - \tilde{p}_i) \prod_{j \in \mathcal{N} \setminus i} (1 - q_j) \right] \right\} \tag{5.5}$$

Proof. In this approach user i exploits fully each frame where no contention exists and only partially the ones where competition has been identified. Thus,

$$\bar{T}_i = q_i \left\{ 1 \prod_{j \in \mathcal{N} \backslash i} (1 - q_j) + \tilde{p}_i \prod_{j \in \mathcal{N} \backslash i} (1 - \tilde{p}_j q_j) - \tilde{p}_i \prod_{j \in \mathcal{N} \backslash i} (1 - q_j) \right\} \qquad (5.6)$$

The first term corresponds to the no-contention frames while the second captures the total throughput derived from empty slots given that the user transmits with probability $\tilde{p}_i$. Given that the slots belonging to an empty frame have been counted twice, we have to remove the third term.

The resulting energy consumption of user i is either due to staying ON or a transmission:

$$\bar{E}_i = q_i c_1 + 1 q_i \prod_{j \in \mathcal{N} \backslash i} (1 - q_j) c_2 + \tilde{p}_i q_i \left[1 - \prod_{j \in \mathcal{N} \backslash i} (1 - q_j) \right] c_2 \qquad (5.7)$$

The first term is the energy consumed when the terminal is ON, the second corresponds to transmission cost in empty frames and the last is the energy consumed in frames where competition has been identified. $\qquad \square$

The complicated throughput and energy expressions for the modified scenario make the problem of optimal scheduling analytically intractable. Thus, in order to get insight on the performance of the modified strategy, we consider a simplified version of the problem by restricting the feasible user strategies. In particular, we assume that all the users of the system can be classified in one of the following three groups; the aggressive, the conservative and the passive ones. The former capture the medium whenever they are active (ON), the second transmit only whenever they sense an empty frame and the last do not participate at all.

We consider Algorithm 5 that performs a search over the solution subspace. This algorithm is of exponential complexity and will be used only for comparison purposes. It can be shown that in the optimal scheduling for our restricted solution space, at least one aggressive and one conservative user exist. In all the simulation cases that we have tried, this categorization did not to limit the system performance; nevertheless a formal proof that the optimal solution lies on the restricted subspace could not be derived.

5.4 Game Theoretic Approach

Previously we derived probabilistic medium access protocols that require coordination of actions of the users involved. However, in an autonomous setting as the one considered here, individuals may not comply with the rules imposed by the protocol. Actually, users may exhibit selfish behaviour and select the strategy that maximizes their own utility, namely their individual throughput, at the expense of others. Thus, in this section we model the user interaction/contention as a non-cooperative game. We derive models for both the initial and the modified scenario.

Algorithm 5 Modified optimal probabilistic frame scheduling

1: Search over $\mathcal{B} = \{\mathcal{A}, \mathcal{C}, \mathcal{P}\}$, i.e. the set of all the possible partitions of $\mathcal{N}$ of size 3, with $|\mathcal{A}| \geq 1$ and $|\mathcal{C}| \geq 1$ for the throughput optimal assignment:

2: **for all** $i \in \mathcal{A}$ (% *aggressive*) **do**

3: $\quad \{\tilde{p}_i, q_i\} = \left\{ 1, \min\left\{ \frac{\tilde{e}_i}{c_1 + c_2}, 1 \right\} \right\}$

4: **end for**

5: **for all** $k \in \mathcal{C}$ (% *conservative*) **do**

6: $\quad \{\tilde{p}_k, q_k\} = \left\{ 0, \min\left\{ \frac{\tilde{e}_i}{c_1 + c_2 \prod_{j \in \mathcal{N} \setminus k}(1 - q_j)}, 1 \right\} \right\}$

7: **end for**

8: **for all** $j \in \mathcal{P}$ (% *passive*) **do**

9: $\quad \{\tilde{p}_j, q_j\} = \{0, 0\}$

10: **end for**

A non-cooperative game is defined by a set of players, a set of strategies and a metric that indicates the preferences of the players over the set of strategies. In our case we have:

- **Players**: the N users

- **Strategies**: user's i set of feasible medium access and ON–OFF probabilities $\mathcal{I}_i = \{p_i, q_i : \bar{E}_i \leq \tilde{e}_i$ and $0 \leq p_i, q_i \leq 1\}$

- **User preferences**: represented by a utility function $U_i(\mathcal{I}_i)$; peer i prefers strategy $\hat{\mathcal{I}}_i$ to $\mathcal{I}_i$ iff $U_i(\hat{\mathcal{I}}_i) > U_i(\mathcal{I}_i)$.

5.4.1 The initial scenario as a non-cooperative game of perfect information

For the initial optimization problem (5.2) the utility function of user i is defined as $U_i(\mathcal{I}_i) = \bar{T}_i = p_i q_i \prod_{j \in \mathcal{N} \setminus i}(1 - p_j q_j)$. It can be easily verified that the throughput maximizing strategy of each individual is independent of the actions of the other users and comes as the solution of the following optimization problem:

$$
\begin{aligned}
&\underset{\mathcal{I}_i = \{p_i, q_i\}}{\text{maximize}} && p_i q_i \\
&\text{s.t.} && q_i(c_1 + c_2 p_i) \leq \tilde{e}_i \\
& && \{p_i, q_i\} \in [0, 1]^2
\end{aligned}
\tag{5.8}
$$

Lemma 5.4.1. *The throughput optimal strategy for user i is $\{p_i, q_i\} = \left\{ 1, \min\{ \frac{\tilde{e}_i}{c_1 + c_2}, 1\} \right\}$. The resulting game has a unique Nash Equilibrium Point, described by the strategy $\mathcal{I}^* = \{\mathbf{1}, \boldsymbol{q}^*\}$, with $q_i^* = \frac{\tilde{e}_i}{c_1 + c_2}$.*

Proof. Since the utility is an increasing function of both p_i and q_i the energy constraint should be satisfied with equality at the optimum.

The proof is similar in spirit to the one in Lemma 5.3.1. Let $\{\hat{p}_i, \hat{q}_i\}$ be the throughput optimal strategy of user i, with $\hat{p}_i < 1$. From the energy constraint we have $\hat{q}_i(c_1 + c_2 \hat{p}_i) = \tilde{e}_i$. There exists

the alternative strategy $\{\acute{p}_i, \acute{q}_i\} = \{1, \hat{p}_i \hat{q}_i\}$ of equivalent utility (individual throughput) but of lower energy consumption. In particular,

$$\acute{q}_i(c_1 + c_2 \acute{p}_i) = \acute{q}_i(c_1 + c_2) = \acute{q}_i(c_1 \acute{p}_i + c_2 \acute{p}_i) < \tilde{e}_i \tag{5.9}$$

Consequently, there exists $\delta > 0$ such that $(\acute{q}_i + \delta)(c_1 + c_2) = \tilde{e}_i$. This strategy $\{p_i^*, q_i^*\} = \{1, \acute{q}_i + \delta\}$ obviously yields increased utility for the user, which contradicts the user optimality of $\{\hat{p}_i, \hat{q}_i\}$ and completes our proof. The same result can be derived through the KKT conditions.

Given that the optimal strategy of a user is independent of the actions of the other users and is determined only by his own energy constraint, deriving the resulting NEP is straightforward. $\quad\square$

From the above, we may deduce that at the Nash equilibrium point we receive throughput, only when a single user is ON in a frame. Given the NEP of the game we may quantify the performance loss arising due to the selfishness of the individuals, by using so called *Price of Anarchy* (PoA) metric. This is the ratio of the value of the objective function at the global optimum to its value at the NEP and in our setting is given by:

$$\text{PoA} = \frac{\displaystyle\sum_{i \in \mathcal{S}} \frac{\tilde{e}_i}{c_1 + c_2} \prod_{j \in \mathcal{S}\backslash i} \left(1 - \frac{\tilde{e}_j}{c_1 + c_2}\right)}{\displaystyle\sum_{i \in \mathcal{N}} \frac{\tilde{e}_i}{c_1 + c_2} \prod_{j \in \mathcal{N}\backslash i} \left(1 - \frac{\tilde{e}_j}{c_1 + c_2}\right)} \geq 1, \tag{5.10}$$

where $\mathcal{S}$ is the set of enabled users at the global optimum.

Whereas in the classic Aloha games the PoA is unbounded, in our energy constrained Aloha setting the PoA is bounded, since the energy constraints impose a fictitious pricing scheme. The PoA grows unbounded only when at least two users have unconstrained energy resources. These users (e.g. power plugged stations) can capture the medium totally and consequently would involuntarily act as jammers for each other and for all the others, yielding hence zero system throughput.

5.4.2 The modified strategy as a non-cooperative game of perfect information

Here, we consider the game arising from the modified strategy. In this setting, we derive a non–cooperative game of perfect information, where at each iteration, given the parameters $\alpha = \displaystyle\prod_{j \in \mathcal{N}\backslash i} (1 - q_j)$ and $\beta = \displaystyle\prod_{j \in \mathcal{N}\backslash i} (1 - \tilde{p}_j q_j)$ a user selects its best response. By *best response* we mean that each user updates his decision variables, so as to maximize its utility function, in response to the others' actions.

Theorem 5.4.1. *The best response strategy of user i is given by:*

73

$$\tilde{p}_i = \begin{cases} 1, & if[c_1 + c_2\alpha]\beta > (c_1 + c_2)\alpha, \\ 0, & otherwise. \end{cases} \tag{5.11}$$

$$q_i = \frac{\tilde{e}_i}{c_1 + c_2[\tilde{p}_i + (1 - \tilde{p}_i)\alpha]} \tag{5.12}$$

The arising modified game has multiple NEPs.

Proof. Since the utility is an increasing function of $\tilde{p}_i$ and q_i, the constraint needs to be satisfied with equality. Thus, we may replace q_i from eq. 5.5 into the throughput expression, namely eq. 5.4. Then, the partial derivative of the objective function is given by the following expression:

$$\begin{aligned} \frac{\partial \bar{T}}{\partial \tilde{p}_i} &= \tilde{e}_i \frac{[\beta - \alpha][c_1 + c_2\alpha] - c_2\alpha[1 - \alpha]}{\{c_1 + c_2[\tilde{p}_i + (1 - \tilde{p}_i)\alpha]\}^2} \\ &= \tilde{e}_i \frac{[c_1 + c_2\alpha]\beta - (c_1 + c_2)\alpha}{\{c_1 + c_2[\tilde{p}_i + (1 - \tilde{p}_i)\alpha]\}^2} \end{aligned}$$

The sign of this expression depends only on $\alpha = \prod_{j \in \mathcal{N} \backslash i} (1 - q_j)$ and $\beta = \prod_{j \in \mathcal{N} \backslash i} (1 - \tilde{p}_j q_j)$. As a result given the actions of the other users, the objective function is either a strictly increasing or a strictly decreasing function of $\tilde{p}_i$. Thus, the best response strategy of user i is given by eq. 5.12. $\square$

5.5 Numerical Results

In order to quantify the throughput performance of the proposed schemes we perform extensive simulations of a small, easy to follow system. We consider a scenario of $N = 5$ terminals with energy constraints given by $\tilde{e} = [30, 25, 15, 10, 5]$ and $\{c_1, c_2\} = \{50, 70\}$ units. By abusing slightly the definition of PoA, we use the modified optimal as the performance benchmark and for each scheme we derive the performance ratio metric defined as:

$$\text{Performance\_ratio}_X = \frac{\text{Throughput of modified optimal}}{\text{Throughput of scheme X}} \tag{5.13}$$

Thus, all the figures depict the performance degradation in comparison to the modified optimal. Since the modified game has several NEPs we depict the PoA i.e. the ratio of the throughput at the optimum to the throughput at the worst NEP, the price of stability (PoS) i.e. the ratio of the throughput at the optimum to the throughput at the best NEP, and the mean performance at the NEPs. Regarding the initial setting, we depict the performance degradation of the original optimal, the fair approach and the initial game theoretic scheme.

Initially, we consider how the energy budget of the less energy constrained user affects the performance of the system as a whole. In Figure 5.2 we see that the additional power budget increases the performance degradation due to the additional collisions caused. The system stabilizes for $\tilde{e}_1 = c_1 + c_2$, where user 1 has sufficient energy budget to capture medium entirely on his own. We

74

Figure 5.2: The throughput performance of the system as an expression of the less energy constrained user
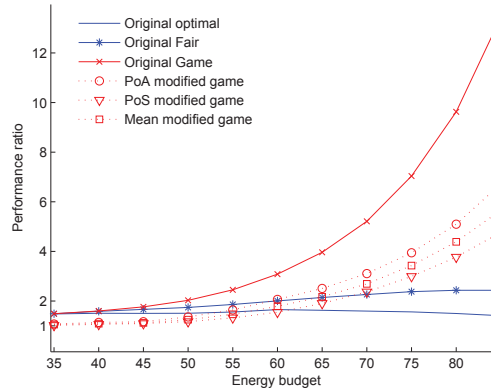
Figure 5.3: The throughput performance of the system as the total energy budget of the system increases

notice also that the modified strategy, of backing off whenever a collision is detected, provides significant performance benefits. The improvement is becoming more significant as the available energy budget increases.

In Figure 5.3 we depict the system performance as we relax the energy constraints of every terminal. We start from the energy budget vector of $\tilde{e} = [30, 25, 15, 10, 5]$ and increase each dimension by 5 within each iteration. The x axis refers to the $\tilde{e}_1$ value, whereas the whole vector is given by $\tilde{e} = [\tilde{e}_1, \tilde{e}_1 - 5, \tilde{e}_1 - 15, \tilde{e}_1 - 20, \tilde{e}_1 - 25]$. In this scenario, one may notice the benefits that arise from the coordination among the terminals. The coordinated approaches are robust again the increased availability of energy, whereas in the game theoretic approaches this leads to an increased number of collisions, due to the increased aggressiveness of the users.

Next, we consider the impact of the transmission cost c_2. We notice in Figure 5.4 that in a scenario of low energy constraints the increased transmission cost makes the users less aggressive,
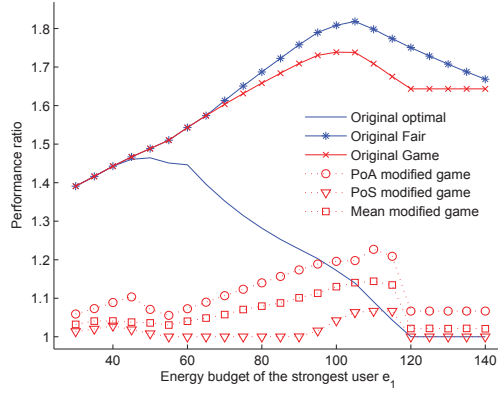
75
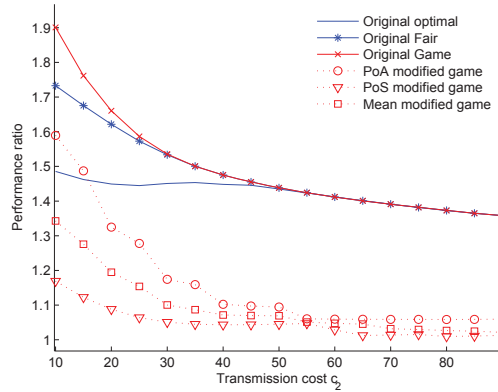
Figure 5.4: The throughput performance of the system as an expression of the transmission cost
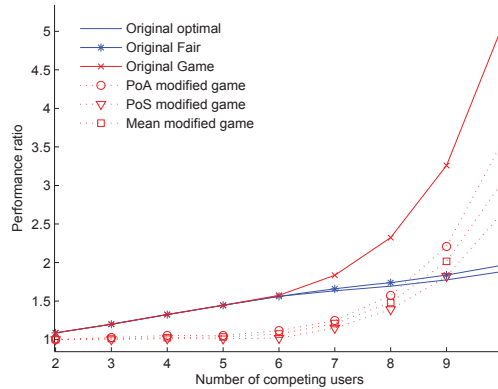
Figure 5.5: The throughput performance of the system an expression of the number N of competing terminals

leading thus to reduced collisions and consequently the performance gap among the coordinated and the game theoretic approaches diminishes. Nevertheless, the performance gap of the initial and the modified strategy remains more than 40% throughout.

Finally, we investigate the impact of the number of competing users on the performance at the resulting equilibria (Figure 5.5). Starting from a system of only two competing users with energy constraints $\tilde{e} = [5, 10]$ we keep adding in each iteration a single terminal of increasing energy budget. Thus, newcomer i is characterized by energy constraint $\tilde{e}_i = 5i$. As expected the introduction of additional users increases competition for the medium and leads to more collisions. Consequently, the performance of the game theoretic approaches deteriorates very fast, whereas the coordinated approaches exhibit significant robustness.

5.6 Conclusion

In this chapter we pursued a better understanding of the energy–throughput tradeoff for mobile devices that support sleep modes and operate according to contention medium access schemes. Initially, we derived the throughput optimal strategy of probabilistic medium access under energy constraints. Next, in order to capture the autonomous nature of mobile terminals we developed game theoretic models. Compared to the unconstrained case, we showed that energy constraints reduce contention and enable better exploitation of the medium.

We observed that due to lack of coordination, distance from the optimal strategy (i.e. PoA) tends to increase as energy availability increases. In order to counteract this, we developed a simple modified medium access scheme, where the contention state of the medium can be sampled within a frame by the competing stations. The modified scheme implicitly coordinates the actions of the terminals and leads to even better exploitation of the limited energy resources.

In this work, we have assumed non–cooperative games of perfect information. The scenario where each involved entity has only a subjective belief on its opponents' strategies is an interesting topic of future study. Besides, mechanism design could be considered as a means of driving the system to more efficient equilibrium points, e.g. by penalizing users that exhibit extremely aggressive behaviour.

Chapter 6

Energy Efficiency at the Wireless Network Level: Interference-Aware Relay Selection and Power Control Algorithms

In this chapter, we consider an interference-limited wireless network, where multiple energy-constrained mobile devices form source-destination pairs, communicate over the same channel and compete for a pool of relay nodes. In an attempt to maximize sum-rate of the system, we address the joint problem of relay assignment and power control.

Initially, we study the autonomous scenario, where each source greedily selects the strategy (transmission power and relay) that maximizes its individual rate, leading to a simple one-shot algorithm of linear complexity. Then, we propose a more sophisticated algorithm of polynomial complexity that is amenable to distributed implementation through appropriate message passing. We evaluate the sum-rate performance of the proposed algorithms and derive conditions for optimality.

We also provide guidelines on how our algorithms can be applied in 4G OFDMA systems. Our schemes incorporate two of the basic features of the LTE-Advanced broadband cellular system, namely *interference management* and *relaying*.

6.1   Introduction to Cooperative Communications

Cooperative communications exploit the broadcast nature of the wireless medium by using intermediate nodes as relays. Thus, a virtual Multiple Input Multiple Output (MIMO) system is formed, realizing the benefits of spatial diversity even when each node is equipped with a single transceiver [73].

In infrastructure-based wireless networks, strategically placed relays are part of the network infrastructure. Relays are generally used to extend coverage and enhance throughput with minimum

deployment cost. Although multihop communications require additional radio resources (frequency channels or time slots), relaying reduces the path loss significantly, by shortening the propagation path. This gain is maximized whenever a non line-of-sight (NLOS) path from a transmitter to the intended receiver is split, through an intermediate relay, into two line-of-sight (LOS) links. The relay nodes also create diverse paths that mitigate the effects of fading during the transmission of data from the source to destination. Finally, relaying may also increase capacity by enabling spatial reuse, allowing thus multiple transmissions to take place simultaneously in the same frequency/time slot throughout a cell, as shown in [74]. However, in such scenarios interference management is of crucial importance.

In this work, we consider the interference-limited environment that arises when multiple unicast communications take place over the same physical channel. Our network consists of communicating sources and destinations and idle nodes that may serve as relays. In this setting, we study relay selection and power control as the main mechanisms of achieving the maximum sum-rate performance for the system. We derive easy to implement heuristic algorithms that require minimal information exchange. The proposed algorithms are ideal for scenarios where the resource allocation decisions need to be made fast e.g. due to rapidly changing channel conditions. We also derive algorithms of polynomial complexity that coordinate the actions of the terminals and exhibit near optimal performance.

The difficulty of the problem under consideration lies on the following:

i. Interference couples relay selection and power control.

ii. Due to interference, the selected transmission power of any single device affects all the others

iii. The first and second hop transmission rates are coupled, since the achievable rate of the bottleneck link determines the maximum end-to-end rate.

In order to deal with these challenges, first we assume that each source is agnostic of relays, i.e. transmits without knowing whether a relay will be used to assist its transmission. Since rate is an increasing function of transmission power, everyone is expected to transmit at maximum power to achieve the maximum individual rate. However, later we relax this assumption, by introducing a protocol that allows sources and relays to coordinate their actions in an attempt to improve the total rate of the system.

The contributions of this work are the following:

1. We develop lightweight resource allocation algorithms (of at most polynomial complexity), amenable to distributed implementation and applicable to any relay assisted network (from ad-hoc to infrastructure-based ones) and any relaying strategy.

2. We derive conditions for the optimality of the proposed algorithms and characterize their impact on the sum-rate performance of the system.

3. We present a case study for the LTE-Advanced system, indicating the applicability of our proposed algorithms and the performance benefits derived.

This chapter is organized as follows. Section 6.2 provides an overview of existing works in the field. In Section 6.3 we present our system model and the assumptions made. In Section 6.4 we describe our relay selection and power control algorithms and derive conditions for optimality. Section 6.5 describes how our proposed algorithms can be applied in the LTE-Advanced framework. Numerical results quantifying the performance of our proposed schemes are presented in Section 6.6 for an interference-limited ad-hoc network and an LTE-Advanced scenario. Section 6.7 concludes our study.

6.2 Related Work

Several relaying strategies have been proposed with *Amplify and Forward* (AaF) and *Decode and Forward* (DaF) being the most common ones [75]. In the former, the relay acts as a repeater, amplifying the received signal (noise included) in the analog domain, whereas in DaF the relay decodes the received signal, re-encodes it and forwards it to the destination. Regardless of the strategy applied, the performance of cooperative communications highly depends on the allocation of network resources, namely the relay assignment and power control at transmitter side.

In this direction, the authors of [76] propose an iterative relay selection algorithm with a max-min fairness objective, where out of the relays that lead to improved minimum capacity, the best one is selected. After some iterations of reassignments, the proposed algorithm converges to the optimal assignment. In [77] the problems of relay selection and power allocation are modeled as auctions, where each user makes best response bids in an attempt to maximize its utility and the relay allocates its transmission power according to the bids. This leads to a distributed algorithm that converges to a Nash equilibrium point. The same resource allocation problem is modeled as a Stackelberg game in [78], where sources are the buyers and relays serve as sellers. A relative problem, that of scheduling users over multiple OFDM carriers in relay-enabled networks, is addressed in [79].

The significance of relay selection is indicated by the recent interest of the research community in relaying for next generation wireless systems. A practical system that benefits from the introduction of relay nodes is the 802.16j that was recently finalized in [80]. Besides, LTE-Advanced and IEEE 802.16m (under development [81, 82]) consider the use of relays so as to meet the 4G performance requirements described by IMT-Advanced, especially for users located close to the cell edge. All these systems are based on orthogonal frequency division multiple access (OFDMA) schemes for the downlink, mitigating thus the effects of intersymbol interference (ISI) and providing robustness to frequency selective fading.

A scenario of a single 802.16 cell consisting of a Base Station (BS), several infrastructure Relay Stations (RS) and Mobile Stations (MS) uniformly distributed within a cell is considered in [83]. The authors of this paper show that in the downlink, when the relays operate in transparent mode, i.e. just forward data but take no synchronization or control decisions, significant throughput improvement appears only for the half of the cell coverage area. In a similar setting [84] quantifies the tradeoff between coverage extension and capacity increase that relays can offer through spatial reuse.

Most of the works in the field of cooperative communications assume that the transmissions take place over orthogonal channels. In TDMA, OFDMA and CDMA for example, interference is caused only by transmitters using the same time slot, frequency channel and spreading code respectively. However, due to the scarcity of channel resources, in an attempt to improve spectrum efficiency, frequency reuse is very common in practice. For example, in cellular systems neighboring cells tend to use the same channels. Most existing works either consider this interference negligible or handle it as noise. Strong interference may appear though, whenever adjacent cells transmit over the same time/frequency to users located almost in the same location. The impact of interference becomes even more significant when the coverage areas of neighboring BSs overlap. In such cases proper interference management is of crucial importance, requires though extensive cooperation among the BSs for the relay selection and frequency allocation. The most straightforward way to tackle interference is transmission power control.

It was only recently that the standardization committees recognized inter-cell interference as one of the primary limiting factors of the performance of current cellular systems and set interference management as one of the major research directions for the next generation communication systems. In this direction, the authors of [85] propose a heuristic power allocation scheme, where each relay selects its power so as to achieve a minimum bit error requirement and minimize the interference caused. Work [86] investigates the performance of several emerging half-duplex relay strategies in interference-limited cellular systems. In [87] a message passing based algorithm for distributed power control and scheduling in a line network is proposed and is shown to be optimal for the K-hop interference model.

6.3 System Model

We consider a wireless network of N sources, N destinations and K intermediate nodes. The latter are not communicating within the time of interest and hence can serve as relays for the active communication pairs. All nodes are arbitrarily located in a plane. We denote with $\mathcal{S} = \{S_1, S_2, \ldots S_N\}$ the set of sources, $\mathcal{D} = \{D_1, D_2, \ldots D_N\}$ the set of destination nodes and $\mathcal{R} = \{R_1, R_2, \ldots R_K\}$ the set of potential relays. Sets $\mathcal{S}$, $\mathcal{R}$ and $\mathcal{D}$ are disjoint sets. We consider only point-to-point (unicast) communications with $\{S_i, D_i\}$ defining communication pair i. Each source has always queued packets to transmit to the corresponding receiver and may transmit them either directly to the destination or through a relay. All the transmissions take place in the same frequency channel and thus interference has to be taken into consideration.

Each node has a single transceiver and consequently, simultaneous transmission and reception is not feasible. Thus, communication is half-duplex i.e. each frame consists of two timeslots of fixed and equal duration of $T/2$ time units each. In the first slot, the sources transmit and the relays overhear the transmission. During the second one, each relay forwards the received signal to the proper destination. Here, for simplicity we assume that the two timeslots are of equal duration. Another option would be to set the relative duration of the timeslots, such that for each Source-Relay-
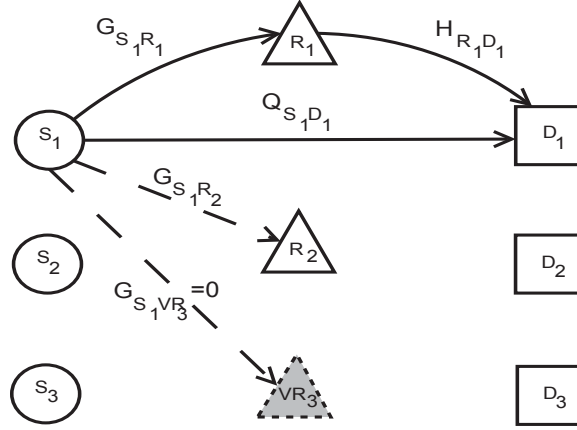
Figure 6.1: A network of $N = 3$ communication pairs and $K = 2$ relays

Destination link, equal amount of information is transferred through the two hops. However, such a scheme would require extensive coordination among the transmitters in order to synchronize their transmissions. An analysis of the scenario of unequal timeslots is presented in [88], but for the case of orthogonal channels, where no interference exists.

We use $\boldsymbol{G}$, $\boldsymbol{H}$ and $\boldsymbol{Q}$ to denote the source-relay, relay-destination and source-destination channel gain matrices respectively. For example, element $G_{S_i R_k}$ captures fading, path loss and antenna gains of the link between nodes S_i and R_k. We assume that the transmission frame length is small compared to the channel coherence time and as a result all channel gains can be considered fixed during the time of interest. We depict in Fig. 6.1 an indicative such network.

In this general setting, we would like to find the assignment $\boldsymbol{a} = [a_1, a_2, \ldots, a_N]^T$ of the relays to the sources and the transmission power of the sources $\boldsymbol{p}_S = [p_{S_1}, p_{S_2}, \ldots, p_{S_N}]^T$ and the relays $\boldsymbol{p}_R = [p_{R_1}, p_{R_2}, \ldots, p_{R_K}]^T$ that maximize the end-to-end rate of the system. We assume a maximum transmission power constraint for each transmitter. Without loss of generality, we assume that all nodes are characterized by the same maximum transmission power p_{max} and the assignments are described by:

$$a_i = \begin{cases} R_k, & \text{if relay } R_k \text{ is assigned to source } S_i \\ 0, & \text{no relay assigned (direct transmission).} \end{cases} \tag{6.1}$$

We mention here that we do not consider the scenario where multiple relays assist a single source. Nevertheless, we allow the same relay to be assigned to more than one sources. In this case some form of scheduling is required.

Our ultimate objective can be formally written as

$$\underset{\boldsymbol{p}_S, \boldsymbol{p}_R, \boldsymbol{a}}{\text{maximize}} \quad \sum_{i \in \mathcal{S}} r_{S_i}^{a_i}$$

$$\text{s.t.} \quad 0 \le p_{S_i} \le p_{max} \quad \forall S_i \in \mathcal{S} \tag{6.2}$$

$$0 \le p_{R_j} \le p_{max} \quad \forall R_j \in \mathcal{R}.$$

The expression of $r_{S_i}^{a_i}$, the end-to-end rate of source node S_i assisted by relay a_i, depends also on the strategy that the relay applies, with *Decode and Forward* (DaF), *Amplify and Forward* (AaF) and *Compress and Forward* (CaF) being the most common ones. Regardless of the strategy used, the rate is an expression of the following form:

$$r_{S_i}^{a_i} = f(\text{SINR}_{S_i D_i}, \text{SINR}_{S_i a_i}, \text{SINR}_{a_i D_i}). \tag{6.3}$$

We use the notation K_{ab} to denote a parameter K referring to link $a \to b$, where a is the transmitter and b the receiver. Thus, the signal to interference ratio (SINR) at relay a_i, when it decodes the transmission of source S_i is denoted as $\text{SINR}_{S_i a_i}$ and is given by:

$$\text{SINR}_{S_i a_i} = \frac{G_{S_i a_i} p_{S_i}}{\displaystyle\sum_{l \in \mathcal{S} \setminus S_i} G_{l a_i} p_l + \sigma_{a_i}^2}, \tag{6.4}$$

where $\sigma_{a_i}^2$ is the variance of the zero mean noise in the receiver of the relay. Obviously, if no relay is used, the achievable rate depends only on the SINR of the direct link, i.e. the first term of (6.3).

In this work, we mainly focus on a DaF scenario where only the sources transmit within the first timeslot. Then, each relay decodes the signal and re-encodes it. Within the second timeslot only the relays transmit, forwarding the signal to the respective destination. In the end, each destination has to retrieve the transmitted information out of the received signal. We assume that a destination decodes either the signal received from the respective relay or the direct signal, if no relay is used, leading to the following end-to-end rate expression:

$$\begin{aligned} r_{S_i}^{a_i} &= \max\left\{ r_{S_i}^0, \min\left\{ r_{S_i a_i}, r_{a_i D_i} \right\} \right\} \\ &= \frac{W}{2} \log_2\left(1 + \max\left\{ \text{SINR}_{S_i D_i}, \min\left\{ \text{SINR}_{S_i a_i}, \text{SINR}_{a_i D_i} \right\} \right\} \right), \end{aligned} \tag{6.5}$$

where W is the channel bandwidth. Here, we have assumed that the sources transmit only during the first timeslot, leaving hence the second one for the relay transmissions, which results to the $1/2$ factor. However, this assumption may be easily relaxed. In order to improve the achievable decoding rate at the destination, but at the cost of increased complexity, maximal ratio combining can be applied at the receiver similarly to [75, 76]. If no relay is used, the destination will decode the signal coming directly from the source, or the combination of signals received from the respective source and relay

otherwise. Then, the achievable rate would be:

$$r_{S_i}^{a_i} = \frac{W}{2} \log_2 \left(1 + \max \left\{ \text{SINR}_{S_iD_i}, \min \left\{ \text{SINR}_{S_ia_i}, \text{SINR}_{S_iD_i} + \text{SINR}_{a_iD_i} \right\} \right\} \right). \quad (6.6)$$

6.4 Relay Assignment and Power Control in Interference-limited Environments

In order to solve the optimization problem (6.2), we need to calculate the transmission powers at the sources and the relays, and the relay assignment that maximize the system-wide sum-rate performance. These two problems are strongly coupled, since the optimality of a relay assignment depends on the selected transmission powers and vice versa. Consequently, solving it even in a centralized way is particularly challenging. Initially, we decouple these two operations, by solving the two problems in an iterative way, i.e. for a given an transmission power allocation we derive the optimal relay assignment and then we derive optimal power allocation for the resulting assignment.

Given others' powers, the rate of a node is an increasing function of its transmission power. Hence, a natural starting point is to assume that all sources transmit at maximum power in an attempt to maximize their individual rate. We also consider fully cooperative relays, that transmit at maximum power in order to forward the signals to the destinations. Since we do not consider any power budget constraints, which would restrict the willingness of the relays to forward messages in order to save energy. Thus, the initial power allocation is $\boldsymbol{p}_S^0 = \boldsymbol{p}_R^0 = p_{max}$.

6.4.1 The relay selection problem

If we assume full Channel State Information (CSI) at the transmitter, the problem of finding the optimal relay assignment is of exponential in the number of sources complexity, namely $O\left((K+1)^N \right)$, since each source has $K+1$ choices, either to transmit through one of the K relays or directly to the destination. However, any schedule, where a relay is assigned to more than one sources cannot be sum-rate optimal.

Remark 6.4.1. *The sum-rate optimal assignment is a matching from the set of sources $\mathcal{S}$ to the set of relays $\mathcal{R}$.*

Proof. We will prove this by contradiction. Assume that the sum-rate $\tilde{r}$ achieved by an assignment where more than one sources, say $\{S_1, S_2, \ldots, S_l\}$ use the relay R_k is optimal. Since each relay has a single transceiver and operates over a specific channel a time sharing schedule has to be applied. Given the randomness of the channel gains the probability that any two of these sources achieve equal rates, i.e. $r_{S_i}^{R_k} = r_{S_j}^{R_k}$ is zero. As a result the rates are ordered, with say $r_{S_i}^{R_k}$ having the largest value. Thus, if instead of having the relay being shared by all these sources, we assign it to user S_i we get a $r^* > \tilde{r}$. Consequently no assignment, where a relay is assigned to more than one sources, can be sum-rate optimal . $\qquad\square$

85

This way the complexity of the problem is reduced, but remains exponential. In this work we aim to develop distributed protocols that can guarantee near optimal performance in at most polynomial time.

From the rate expressions above, we notice that even under full CSI at the transmitter, the sources are ignorant of the CSI of the second hop, and consequently cannot identify whether using a relay is beneficial. However, they can exclude some relays, which compared to direct transmission cannot offer any rate improvement. Generally, each source S_i is able to categorize relays in two sets, namely the bad ones $\mathcal{B}_{S_i} = \{R_k : \text{SINR}_{S_i D_i} \geq \text{SINR}_{S_i R_k}\}$ and the unknown ones $\mathcal{U}_{S_i} = \{R_k : \text{SINR}_{S_i D_i} < \text{SINR}_{S_i R_k}\}$. The first one includes all the relays, which cannot improve the performance of source S_i, whereas no decision can be made a priori for the relays in the latter one, since the achievable rate depends also on second hop parameters. As a result, each source has to make a guess on which relay may offer the maximum rate. In this direction, we devise the following alternatives.

One-shot greedy algorithm

Each source greedily selects the relay that is expected to maximize its own rate. That is, assuming full CSI at each source for its links to the relays, the source selects out of $\mathcal{U}_{S_i}$ the relay that has the best first hop performance, or none if $\mathcal{U}_{S_i} = \emptyset$. This relay selection can be formally described as:

$$
\begin{aligned}
a_i &= \arg\max_{k \in \{\mathcal{R} \cup 0\}} \hat{r}_{S_i}^k \\
&= \arg\max_{k \in \mathcal{R}} \{\text{SINR}_{S_i D_i}, \text{SINR}_{S_i k}\}
\end{aligned}
\tag{6.7}
$$

We use the hat to denote that this is an estimation of the actual achievable rate, based only on the first hop. Whenever a relay k is selected by two or more sources, since a relay within one timeslot cannot decode and forward more than one signal (see previous remark), it will forward the one that achieves the best rate, i.e. the strongest signal.

This myopic relay selection approach is of linear complexity and is expected to yield suboptimal assignments, whenever the performance of the second hop is the limiting factor. Due to lack of coordination this scheme may also yield suboptimal assignments even if all first hop channels are worse than the corresponding second hop ones. Since each source S_i either uses its best candidate relay out of $\mathcal{U}_{S_i}$ or none (direct transmission to the destination), whenever a relay is selected by two or more sources, useful relays may remain unassigned and the diversity gain is not fully exploited.

Remark 6.4.2. *Whenever i) the achievable rates in the first hop links are smaller than the corresponding second hop ones (i.e. the first hop is the bottleneck) for all the communication pairs, and ii) the greedy algorithm returns an assignment where no relay is selected by more than one source, this is the optimal assignment. In this case the optimal assignment is a maximum matching of size $|S|$ from the set of sources $\mathcal{S}$ to the set $\mathcal{R} \cup \mathcal{D}$.*

Other possible strategies that we do not consider in this work are either the unassigned sources
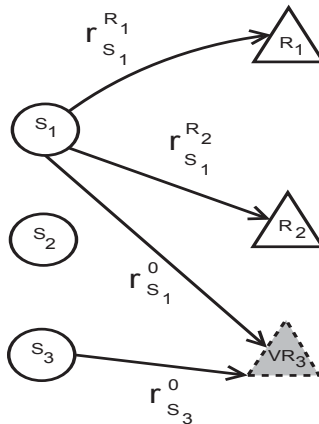
Figure 6.2: The bipartite graph that we use to model the relay assignment as a MWM problem

not to transmit at all or the unassigned sources by using a marking mechanism to select the best out of the remaining relays through subsequent rounds. Alternatively, relays could decide on their own which signal to forward.

In the proposed approach each source acts for itself and no coordination of actions exists. Next, we propose an alternative algorithm that enforces the cooperation of the nodes through appropriate message exchanges.

Bipartite Maximum Weighted Matching (MWM) approach

The relay selection problem can be mapped into the problem of finding the maximum weighted matching in a properly constructed complete bipartite graph $\mathcal{G} = \{\mathcal{S}, \mathcal{R}, \mathcal{E}\}$, as the one shown in Fig. 6.2, where the weights of the edges are given by $w_{S_i R_k} = r_{S_i}^{R_k}$. If the interference in the receiver of each link is known, then these weights can be easily calculated.

As previously mentioned, we assume that each node selects its transmission power without considering the interference it causes. Then, interference in the first hop can be easily estimated through appropriate pilot transmissions. However, the interference in the second hop cannot be known a priori, since it depends on which relays will be selected to forward data. Thus, interference at each destination cannot be estimated prior to the relay assignment. Besides, without an estimation of interference we cannot calculate the achievable rate in the second hop. To overcome this, we adopt a conservative approach and assume that in the optimal assignment all the relays are used. Notice that this approach may lead to an overestimation of actual interference.

Under this assumption, the distributed algorithm of [89] can be applied to find the maximum weighted matching. The relays and the sources exchange messages and finally after $O\left(\max(N,K)\right)$ iterations we converge to the optimal assignment. However, this algorithm has some limitations that should be taken into account. First of all, it works only on balanced bipartite graphs. Thus, in order to turn our graph into a balanced one, we need to introduce $|N - K|$ virtual nodes, either on the left side (virtual sources) with zero weights, if $N < K$, or on the right side (virtual relays) of the graph with

$w_{S_i R_k} = r^0_{S_i}$ otherwise. Secondly, it converges only when the maximum weighted matching is unique. To guarantee this, we have to add infinitesimal random values ε_i (disturbances) into the weight of each link, making thus any previously equally weighted matchings, ordered.

Next, we derive conditions for the optimality of the MWM algorithm.

Remark 6.4.3. *With high probability, i.e. with probability going to* 1 *as the number of sources over the number of relays goes to infinity ($\frac{N}{K} \to \infty$), our MWM algorithm finds the optimal assignment.*

Proof. Whenever in the optimal assignment every relay is assigned to a source, all the relays will eventually transmit. Consequently, the interference estimation that we made for the second hop will be accurate and the MWM will return the optimal assignment. As $\frac{N}{K} \to \infty$ every source can find a relay to improve its rate performance leading thus to a perfect matching. □

Starting from the initial power allocation $\boldsymbol{p}^0_S$, $\boldsymbol{p}^0_R$, these algorithms can find an assignment of the available relays to the sources. Given this, we may then modify the transmission powers of the sources and the relays in an attempt to maximize the sum-rate. In the following section we propose two alternatives for the power control part.

6.4.2 The power control problem

Given relay assignment $\boldsymbol{a}$, we have to find the optimal transmission powers for the sources and the relays. In this direction, we propose heuristic power control that attempts to equalize the rates of the first and the second hop and one based on high-SINR approximation.

Rate equalization algorithm (Req)

In our setting, we have two apparently independent steps of power control, one in the first and one in the second hop. Nevertheless, they are coupled, since the rate of both hops have to be equal. If $r_{S_i a_i} > r_{a_i D_i}$, the relay cannot forward the data to the destination at the rate they are transmitted by the source. Thus, we say that for this transmission pair the 1st hop is the bottleneck link. On the other hand, if $r_{S_i a_i} < r_{a_i D_i}$, the transmission rate of the relay cannot be fully utilized, leading to a bottleneck in the second hop.

As a result, for the given relay assignment and starting from initial power vectors $\boldsymbol{p}^0_S$, $\boldsymbol{p}^0_R$, sources and relays should iteratively update their transmission powers to match the rate of the other hop. However, since increasing the transmission power, also increases interference, something that may lead us to worse sum-rate performance, the rate equalization process should be applied only in the non-bottleneck links, leading the transmitters to reduce their transmission power. That is, for each source-relay assignment where $r_{S_i a_i} > r_{a_i D_i}$ source S_i will reduce its transmission power to match the rate of the second hop. Otherwise, the corresponding relay will have to reduce its transmission power. The proposed heuristic for each communication pair is formally described in Algorithm 6.

If we apply this iteratively and since power adaptation is always a power reduction we get a contraction mapping and convergence to a stationary point is guaranteed.

Algorithm 6 Rate equalization step for communication pair i

---

if $r_{S_ia_i} \geq r_{a_iD_i}$ **then**
    $p_{S_i}^{t+1} = y$ such that $r_{S_ia_i}(y) = r_{a_iD_i}(\boldsymbol{p}_R^t)$
else
    $p_{a_i}^{t+1} = z$ such that $r_{a_iD_i}(z) = r_{S_ia_i}(\boldsymbol{p}_S^t)$
end if

Joint source and relay power control algorithm (JsrPC)

The problem of sum-rate maximization through power control has been extensively studied for single hop networks. Although it has not been solved yet due to its nonconvex nature, several approximations have been proposed. The authors of [90] proposed the approximation $\log(1+\mathrm{SINR}) \approx \mathrm{SINR}$ for the low SINR regime. On the other hand, in [91] a distributed algorithm that converges geometrically fast was proposed for the power control in the high SINR regime, based on the approximation $\log(1+\mathrm{SINR}) \approx \log\mathrm{SINR}$.

In our two-hop scenario the rates of the first and the second hop have to be equal. Thus, we may modify the aforementioned algorithms by incorporating the additional constraint $r_{S_ia_i} = r_{a_iD_i}$ for each communication pair i. To achieve this we may apply unconstrained power control on either hop, say the second one, and then a constrained one on the other. If we apply the unconstrained power control in the relay hop in timeslot t, the transmission powers of the sources for the high SINR regime will be given by:

$$
p_{S_i}^{t+1} = min\left\{ \left(\sum_{k\in\mathcal{S}\backslash S_i} \frac{G_{S_ia_k}}{\sum_{l\in\mathcal{S}\backslash S_k} G_{S_la_k}p_{S_l}^t + \sigma_{a_k}^2} \right)^{-1}, y \text{ such that } r_{S_ia_i}(y) = r_{a_iD_i}(\boldsymbol{p}_R^t), \; p_{max} \right\} \quad (6.8)
$$

Here, the first term corresponds to the actual control of transmission power as described in [91]. The second term corresponds to the constraint introduced by the fact that increasing the rate of the source-relay link beyond the rate of the relay-destination link is meaningless. This is actually an incorporation of the Req algorithm (see Algorithm 6). The third term corresponds to the physical limitation of maximum transmission power p_{max} that we assumed for every node of the system.

Next, we derive conditions for the optimality of the JsrPC algorithm.

Remark 6.4.4. *In the high SINR regime (i.e. SINR $\gg$ 1) the JsrPC algorithm yields the optimal power allocation whenever the bottleneck is in the same hop for all the communication pairs.*

Any combination of the aforementioned algorithms along with some control messages can lead to the design of distributed protocols of at most polynomial complexity.
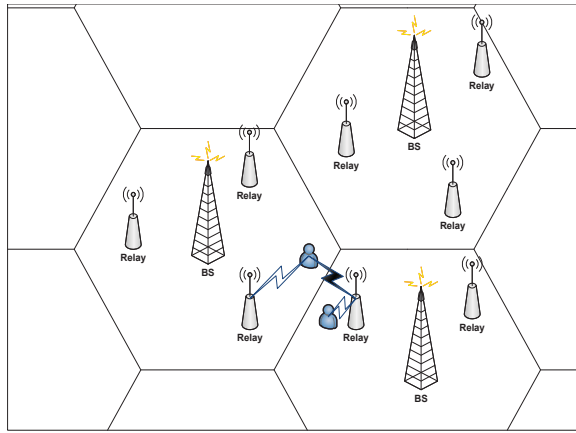
Figure 6.3: The cellular structure of an LTE-Advanced system

6.5 Relay Selection and Power Control in the context of LTE-Advanced

The cellular downlink (uplink) communication scenario, where all the logical transmitters (receivers) reside in the same physical entity, namely the Base Station, comes as a special case of the previously described system model. In this section we focus on the downlink of the upcoming LTE-Advanced system. Such a system is depicted in Fig. 6.3.

Cellular relay-assisted networks can be thought of as a multilevel hierarchical tree structure. The BS lies at the highest level, being the root, the RSs are organized in the intermediate levels and the MSs are the leaves. Thus, each downlink communication can be represented as a top-down path starting from the root, whereas the uplink takes place the reverse way. Here, we consider a two-hop downlink scenario consisting only of a single level of relays, which is also the case for LTE-Advanced. Nevertheless, the analysis performed here can be easily extended to the uplink.

In contrast to the generic scenario described earlier, here relays are infrastructure nodes, strategically placed, either to serve cell edge users or in places where coverage holes appear, and much fewer in number than the MSs. Besides, since communications within a cell take place over orthogonal subcarriers, only inter-cell interference is apparent. Thus, now each BS has to decide how to assign the relays to the MSs, and how to allocate the subcarriers to the communication pairs.

From the point of view of the BS, it is easily deduced that in order to maximize the downlink throughput of the cell, it will allocate all the available subcarriers. Thus, for the first timeslot the BS has to assign every single subcarrier to a receiver, which may be either an RS or an MS. If we assume that the subcarrier allocation is performed in an end-to-end path basis, i.e. the subcarriers used in the first timeslot for communication with relay R_k, are used in the second timeslot only by the same relay for transmission to the corresponding MS, the problem of relay selection and power control for each subcarrier can be mapped to our original problem.

If we focus on a single subcarrier, the sources of interference are the transmitters of the adjacent cells that use the same subcarrier, namely BSs during the first hop and either BSs and / or relays
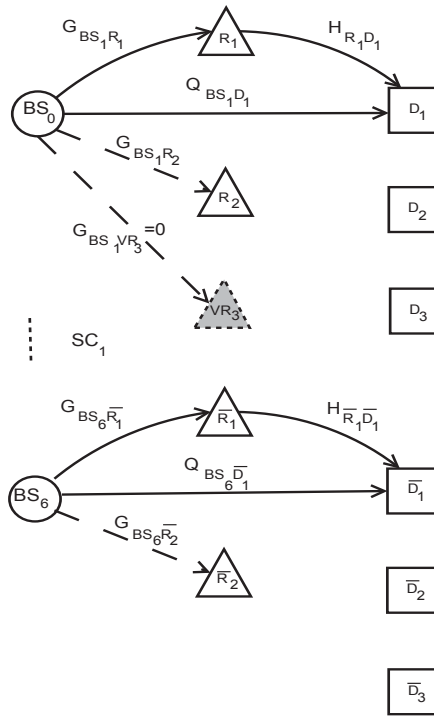
Figure 6.4: A cellular system of 7 cells, with the central one having $N = 3$ mobiles and $K = 2$ relays

during the second hop. Besides, in this setting no contention for the available relays exists among the BSs, since each one can only use the relays within its cell. In other words for each BS_i there exists a set of relays $\mathcal{R}_i$ that can be used, but all these are disjoint sets. A representation of this structure for a single subcarrier, the SC1, is shown in Fig. 6.4. From this figure it is obvious that for each subcarrier and each BS a proper bipartite graph like the one depicted in Fig. 6.2 can be created . Then, the relay selection algorithms of the previous section can be directly applied in this graph. The proposed power control algorithms are also directly applicable to this new setting, with the only difference being the sources of interference. However, in the cellular networks all the resource allocation decisions are made by the BS. Thus, the distributed versions developed earlier are needless. Instead centralized approaches should be applied like the well known Hungarian algorithm for MWM [92].

We saw earlier that in order to maximize the benefits of cooperative communications, the source (the BS in our case) has to be aware of the physical layer conditions for the BS-RS, BS-MS and RS-MS links. In LTE-Advanced such information can be acquired through the CSI-reference signals (CSI-RS), which estimate the condition of a channel and assist the beamforming and scheduling decisions. CSI-RSs are transmitted in every kth subframe, with k being configurable. In our scenario the BS needs also to know the interference caused by the adjacent cells. This can be realized through the Relative Narrowband TX Power (RNTP) indicator, which is transmitted by each BS to its neighbors through the X2 interface.

6.6 Numerical Results

In this section we present some simulation results for the generic scenario of an interference-limited ad-hoc network. We also simulate a cellular network, similar to the upcoming LTE-Advanced, in order to get some insight on the relative performance of the proposed schemes.

6.6.1 The generic scenario

For the following simulations we assume a specific number of nodes, uniformly distributed over a given $q \times q$ square area. We use the channel gain function

$$G_{S_i R_k}(d_{S_i R_k}) = [\max(d_{S_i R_k}, d_0)]^{-\beta}, \qquad (6.9)$$

where $d_{S_i R_k}$ denotes the distance of S_i and R_k, and d_0 the radius around the transmitter where unit gain is assumed to hold. In the following experiments we assume $d_0 = 1$, a path loss coefficient of $\beta = 2$, maximum power (p_{max}) and bandwidth (W) equal to 1 and white Gaussian noise (AWGN) of zero mean and variance 0.01.

In an attempt to quantify the performance loss due to the decoupling of relay selection and power control, we first consider a simple motivating example of two communication pairs and one relay, all randomly located within a square. For such a small network we can also derive the joint relay selection and power control optimal solution. In Fig. 6.5(a) we depict the impact of side length q on the sum-rate performance of the system. As side length q increases, the average distance of any two nodes increases and transmissions experience higher path loss. As expected the polynomial approach performs significantly better than the lower-complexity greedy approach. Generally though, all proposed schemes outperform the option of direct transmission.

In such a simple network, relay assignment is quite trivial, leading thus to identical sum-rate performance for the greedy and the MWM approach. Furthermore, applying the Req algorithm has no significant impact, since no interference exists in the second timeslot. On the other hand, applying the JsrPC algorithm, which combines the MWM relay selection with power control the performance is improved towards the optimum solution, especially for small distances where the gain from the interference mitigation is more significant.

However, in the previous setting, at least for small distances the interference is significant, making the assumption of high SINR regime not valid. Thus, we consider a high SINR scenario where the two communication pairs lie on the vertices of a square and the relay is randomly placed within the square. In Fig. 6.5(b) we depict the impact of side length q, i.e. the distance separating the two communication pairs. As expected here the proposed algorithms perform significantly better, exhibiting near optimal performance. We mention that all the values depicted are mean values of at least 1000 simulation runs.

In order to demonstrate scalability of the proposed algorithms, we consider larger scale random topologies consisting of $N = 15$ sources. We investigate the impact of the number of relays K on performance. We expect that up to a point as the number of available relays increases, better sum-rate
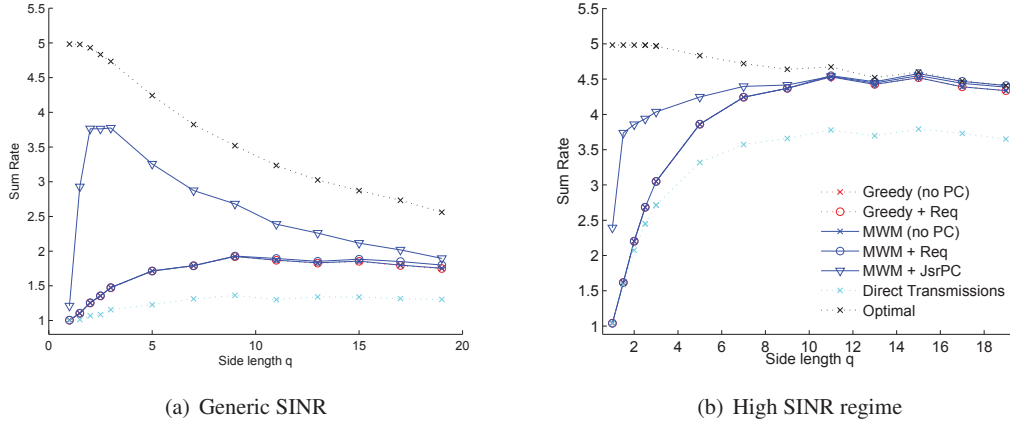
(a) Generic SINR

(b) High SINR regime

Figure 6.5: Sum-rate performance (for different SINR regimes) vs. Side length q of the square area

(a) Square dimension $q = 2$

(b) Square dimension $q = 10$
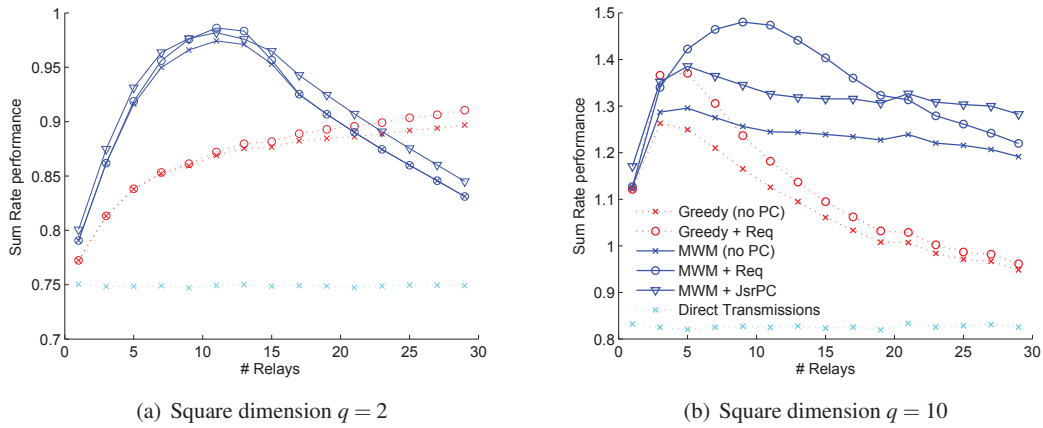
Figure 6.6: Sum-rate performance for 15 sources inside a square of side length q vs. Number of Relays K

performance can be achieved. However, this is not always the case for our algorithms.

The greedy approach makes a myopic decision based only on the first hop. On the other hand, MWM relay selection overestimates interference in the second hop, when the optimal assignment leaves some relays unused. Thus, we expect worse performance as the number of relays K gets larger than N, something evident in Fig. 6.6(a). Here, the small side length q maximizes the impact of interference overestimation, because the redundant terms have large values (due to high channel gains). Nevertheless, compared to direct–transmissions, it is evident that our relay selection algorithms exploit the benefits of cooperative diversity. Furthermore, the proposed power control schemes improve the achievable sum-rate in general. However, since all transmissions take place in the same channel, and we are considering a dense network, we lie in the extremely low SINR regime, where the JsrPC provides only marginal benefits.

In Fig. 6.6(b) we increase the side length q, which also causes path-loss to increase. Here, we

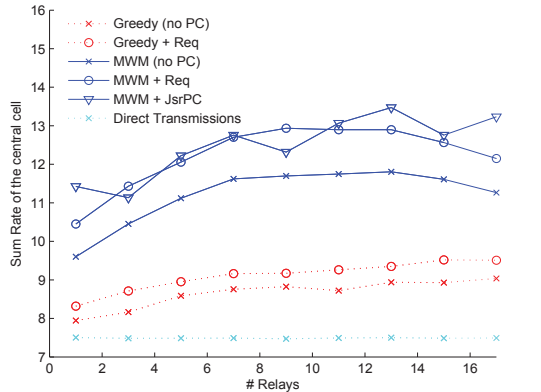Figure 6.7: The impact of number of relays K on sum-rate performance of 20 mobiles located in a cell

observe that the performance benefit from relaying becomes more significant. Besides, this leads us to an environment of less interference, where the power control algorithms perform significantly better. Through proper relay selection and power control the achievable sum-rate gain approaches 100%. Finally, here increasing the number of available relays does not degrade performance losses, since due to increased pathloss most of the relays are beneficial.

6.6.2 The LTE-Advanced scenario

In this section, we study the applicability of the proposed algorithms on a LTE-Advanced scenario. We consider a system of 7 cells, one located in the center and its six direct neighbors, as the one depivted in Fig. 6.3. Each cell has a radius of 3km, the RSs are deterministically placed in a distance of 1.5km from the BS, so that cell edge users benefit the most and 20 MSs are randomly placed within each cell.

Given that OFDM is used in the downlink of LTE-Advanced, interference is only apparent from neighboring cells. We consider a system of 32 data subcarriers sharing a total bandwidth of 5 Mhz. Here, we assume LOS paths for the BS-RS, and the RS-MS communications. This is a logical assumption, since usually infrastructure relays are placed on the rooftops of high buildings. On the other hand, we consider NLOS path loss for the BS-MS links. We also assume lognormal shadowing. The NLOS model that we use is given by $36.5 + 23.5\log_{10} d + \chi_{\text{NLOS}}$, where $\chi_{\text{NLOS}} \sim \mathcal{N}(0,8)$, $\chi_{\text{LOS}} \sim \mathcal{N}(0,3.4)$ in dB, and d denoting the distance between the transmitter and the receiver in meters.

Under this model we study the sum-rate performance of the central cell. The neighboring cells not only serve as sources of interference, but also participate in the power control updates. Fig. 6.7 shows the performance of our proposed algorithms as a function of the number of available relays in the cell. The first thing one may notice is that in the cellular setting the proposed algorithms perform much better. This can be justified by the fact that in LTE-Advanced interference experienced by the

receivers is in general some orders of magnitude smaller than the actual signal. Besides, the scenario considered here incorporates the benefits arising from the transformation of a NLOS link into two LOS ones through relaying. Thus, the sum-rate performance of the cell is improved significantly. Concluding we could say that all the remarks made for the interference limited scenario also hold here, but now the environment is less demanding (lower interference, higher transmission powers etc.).

6.7   Conclusion

In this chapter, we investigated the interaction of relay selection and power control in interference-limited networks. We developed easy to implement distributed algorithms of at most polynomial complexity that are applicable to any type of relay assisted wireless systems and offer significant improvement in the sum-rate performance. We also demonstrated that our algorithms meet the objectives of interference management and relaying exploitation of 4G wireless systems.

Throughout this work we have assumed that the system parameters do not change within the time of interest. It would be interesting though to derive online versions of the proposed schemes that capture the dynamic scenario of nodes entering or leaving the system or time-varying channels. In this case, instead of solving the problem from scratch, online matching schemes like the one proposed in [93, 94] can be used.

Chapter 7

Conclusions and Future Work

7.1 Summary of Contributions

This thesis explores the novel challenges that arise from the limited availability of energy resources at device, system and community level. On the one hand, we considered ways to improve the energy efficiency of the power grid itself. On the other hand, we devised resource allocation schemes that optimally exploit the limited energy resources of mobile devices and we investigated the impact of energy constraints on performance of networked systems.

7.1.1 Smart grid

Initially, we explored the potential of dynamic pricing towards a less costly and more stable power grid. We proposed a day-ahead DR mechanism that captures both the price setting strategy of the operator and the response of home users. Our negotiation-based scheme enables an accurate prediction of the demand distribution for the following day and hence utility operator can efficiently apply economic dispatch even in presence of DR. We showed that properly designed pricing mechanisms align the seemingly contradictory interests of the cost minimizing operator and the self-interested home-users. Our realistic numerical evaluation provides ample evidence that existing works tend to overestimate DR benefits and provide only limited incentives to end-users. This partially justifies the limited penetration of DR schemes in the residential sector.

Due to the large number of home users and the negligible impact of each of them on the DR market, we introduced a new market entity, the aggregator, that represents a set of home users in the negotiations with the utility operator. In this direction, we devised a hierarchical DR market model for the smart grid where aggregators compensate end-users to modify their demand pattern, and compete with each other to resell DR services to the utility operator. Interestingly, our numerical evaluation showed that DR benefits are maximized when the utility is non-profit, which is consistent with the current trend in electricity market restructuring. Interestingly though, in such a hierarchical market the utility of the end-users is not a monotonously increasing function of elasticity, which indicates that truthful reporting may not be the optimal strategy of the users.

7.1.2 VM migration schemes for the cloud

Next, in an attempt to improve the energy efficiency of mobile devices per se, we investigated the Mobile Cloud Computing (MCC) scenario where energy-constrained devices offload their computationally intensive tasks to remote cloud servers. We considered the objectives of minimizing execution time and energy consumption. We also identified the lack of QoS guarantees as the missing tile of cloud computing. However, providing QoS guarantees is particularly challenging in the dynamic MCC environment, due to the time-varying bandwidth of the access links, the ever changing available processing capacity at each cloud server and the time-varying data volume of each virtual machine. Thus, we proposed novel cloud architectures and migration mechanisms that effectively bring the computing power of the cloud closer to the mobile user.

In contrast to the migration policies currently applied in commercial virtualization platforms, like vSphere [40], we showed that migrations should be performed according to the estimated performance and not based on the estimated resource demands of each task/VM. The underlying idea of the proposed schemes is that a VM should migrate from its current server if the anticipated total execution time at a new server (including the migration and download delay) is less than the execution time at the current one. An overview of cloud scenarios in which VM migration schemes can be applied is presented in [38].

7.1.3 Energy efficient wireless access

We also investigated the energy-throughput tradeoff for mobile devices that support sleep modes and compete for access to the shared medium. We characterized the throughput optimal strategy under energy constraints and we developed game theoretic models that capture the autonomous nature of the terminals. As expected, the lack of coordination causes some performance degradation. Our analysis shows that energy constraints modify the medium access problem significantly. Interestingly, the limited energy resources indirectly reduce contention and lead to bounded price of anarchy.

Next, we developed a modified medium access scheme, where the contention state of the medium can be sampled within a frame by the competing stations. We identified that in this case each mobile user faces an interesting dilemma: whenever the frame is sensed busy, is it beneficial to backoff in order to save energy or should it transmit? The proposed scheme leads to more efficient exploitation of the medium, by implicitly coordinating transmissions within a frame, without any direct information exchange among the competing terminals.

7.1.4 Interference management for energy efficient wireless access

Finally, we considered the interaction of relay selection and power control in interference-limited environments. Starting from an uncoordinated approach, we showed that system performance is mainly determined by the competition of users for the same relays and by the interference they cause to each other. Thus, we developed easy to implement distributed algorithms of at most polynomial complexity that are applicable to any type of relay assisted wireless systems. The key idea is to

coordinate the selection of relays and adjust transmission power through minimum message passing.

For given subcarrier allocation, we also outlined how the proposed algorithms can be applied in OFDMA systems like LTE-Advanced. Our algorithms meet two major objectives of 4G systems, namely relaying and interference management. Through extensive simulations, we demonstrated that the proposed schemes guarantee significant improvement in terms of sum-rate performance and by adjusting transmission power also lead to tangible energy savings for the mobile devices.

7.2   Usefulness of Results in Practical Systems

In this section, we attempt to translate the major findings of this thesis into practical guidelines for the design of energy-efficient systems.

7.2.1   Electricity market restructuring

In this thesis we proposed a novel DR market, where all the required communication between the utility operator and the home users takes place before the actual day of operation. Our day-ahead approach provides numerous advantages. First of all, guarantees that the home users are aware of the exact pricing pattern for the following day and hence can schedule their demands accordingly. Second, enables the operator to accurately predict the total consumption for each period of the following day. Finally, guarantees that there is always sufficient time so that the negotiations among generators, the utility operator and the home users converge to a stable operating point. All these would not be possible if a real-time market approach was adopted instead.

Besides, we suggest that DR should be performed in an automated manner with minimum human intervention, since the lack of knowledge and time from the users' side to respond to dynamic prices hinder large-scale penetration of residential DR programs. Finally, existing DR schemes expect users to bear the investment cost of smart meter installation, while the utilities reap most of the DR benefits. This is mainly due to the fact that each individual amounts to a small portion of the total demand and hence has limited negotiation power. Our work provides tangible evidence that organization of home users into communities under the umbrella of an aggregator will drastically change the DR landscape.

7.2.2   Cloud providers

In the pursue of energy efficient cloud computing, most cloud providers mainly focus on the deployment phase. In this thesis, we suggest that significant improvements can be also derived through novel migration mechanisms that require no changes of the existing infrastructure. Live VM migrations enable providers to adjust the load at each server. Given that energy consumption is an increasing function of the load of a server, we suggest that VM migrations can be used to reduce total energy consumption of a cloud facility by consolidating tasks on fewer physical servers and turning off unused ones.

Apart from energy consumption per se, a major objective of a cloud provider is the minimization of the corresponding electricity cost. In this direction, VM scheduling enables the cloud provider to exploit temporal variation of electricity prices by scheduling non-critical tasks in low-cost periods. In addition, given that servers are geographically dispersed over the world, exploiting spatial variation of electricity prices is an additional option. At each time instance, VM migrations enable execution of tasks wherever cost of electricity is the lowest.

Contemporary virtualization platforms such as vSphere of VMware introduced recently the capability of VM migrations across distant servers. Currently migrations are performed so as to keep average CPU and memory utilization balanced across servers. Since energy consumption of cloud infrastructure corresponds to a significant portion of a cloud provider's operating costs, we expect that migration mechanisms that consolidate load in fewer physical machines will be implemented soon. However, any such migration scheme should also consider performance degradation caused due to multitenancy.

7.2.3 Mobile network operators

In this thesis, we underlined the importance of relaying and interference management in 4G wireless communication systems. Relays enhance throughput of cell edge users, while power control is a promising mechanism for handling interference and calibrating energy consumption. Although both are basic features of 4G systems, their exact implementation is not defined in the corresponding standards, but is vendor-specific. Thus, mobile network operators (MNO) should preferably invest on 4G implementations that incorporate such mechanisms and achieve better exploitation of the available resources.

In addition, we suggest that sleep modes can provide significant energy benefits, especially in scenarios of contention-based medium access. A mobile device by turning off its transceiver can increase its lifetime. A side effect of sleep modes is that contention for the medium is implicitly reduced and hence also throughput increases. Although here we focused on the user side, similar energy saving techniques can be applied at the base stations. In particular some base stations can be turned off in low traffic periods, hence leading to reduced operating costs for MNOs.

7.3 Future Challenges and Open Research Problems

We conclude this thesis with a discussion of open research problems in the area of smart grid and energy-constrained systems.

7.3.1 Smart grid

In the smart grid context, we have developed day-ahead DR market models that rely on iterative negotiation mechanisms and hence require communication among the participating entities. In addition, we assumed that the actual preference of the user can be deduced in an automated manner. However,

as we showed the benefits of DR depend strongly on the accuracy of this estimation. For this purpose, novel regression methods that accurately estimate the utility function of a user through historical data need to be developed in order to fully exploit the potential of demand-response.

Virtual aggregators and incentives

The potential benefits from the introduction of aggregators has to be further explored. Since each aggregator represents a significant amount of total demand in the DR market, it can act strategically as price anticipator by considering the impact of its actions on the price setting strategy of the operator. Hence, the increased negotiation power of the aggregators can be exploited to maximize the DR benefits for the home users. Instead of being represented through a third-party, home users could also form self-organized DR communities, i.e. virtual aggregators. For this purpose, tools from game theory can be used to model coalition formation among users and to analyze the competition of several virtual aggregators in selling their DR services to the operator. Here, we have assumed that the users truthfully respond to the announced compensation from the aggregators. However, since additional benefits may arise through strategic misreporting, the investigation of such user strategies and the derivation of mechanisms that guarantee truthfulness are interesting topics for future study.

Distribution automation

In this thesis the transmission and distribution constraints of the system are not taken into account. However, such operational constraints are critical for the grid survivability. The power grid is a dynamic system mainly due to the continual variability of demands, but also due to the massive distributed and variable energy generation from renewables. In this direction, the potential of distribution automation has to be explored and particularly towards the proper integration of renewable energy sources and energy storage devices. Distribution automation consists in frequency and voltage control schemes that take into account the constraints of the distribution network through power flow equations. Frequency control keeps the active power generated in the grid equal to the active power consumed, while voltage control at the distribution feeders minimizes distribution losses.

Optimal storage placement and dimensioning

An interesting topic of future study is that of optimal placement of renewables and storage facilities so as to balance demand and supply and minimize distribution losses. This problem becomes even more challenging when optimal dimensioning of energy storage facilities is also considered. In general, energy has to be stored and consumed locally, close to where it is generated so as to minimize distribution losses. On the other hand, central large batteries provide more opportunities to exploit fluctuations of price, demand and renewable generation. Deriving the optimal solution to this tradeoff could eventually turn residential renewable systems into a reality. A preliminary study of optimal storage placement within the distribution network can be found in [95].

7.3.2 Energy efficient cloud computing

In the context of cloud computing the derivation of optimal VM migration schemes towards a more energy efficient, and hence less costly, cloud is an active area of research. Such an issue becomes especially critical given that cloud facilities are large-scale datacenter networks, which consist of multiple dispersed server facilities. In this thesis, we showed that VM migrations can effectively reduce energy consumption of the cloud ecosystem.

Energy-driven VM migration schemes

Energy consumption of a cloud facility depends on the number of active server machines. Further, for each active machine, energy consumption is an increasing function of the server load. Task migration can contribute to significant savings in energy consumption by (*i*)reducing the number of active servers through appropriate concentration of the tasks on fewer physical machines with the aid of virtualization (consolidation), and (*ii*) reducing the energy consumption of individual servers by moving the processes from heavily loaded to less loaded servers (load balancing).

The challenge lies in that a migration has contradicting consequences due to the reasons above: by reducing the number of active servers as reason (*i*) dictates, the load is concentrated on fewer servers which are thus loaded more. On the other hand, by performing load balancing according to reason (*ii*), the heavy load of some servers is alleviated, but at the same time more servers would need to be activated to accommodate it. The final choice will depend on the relative amounts of energy consumption of servers when they are idle and loaded, as well as on the precise dependence of energy consumption on physical machine load. An associated challenge would therefore be to derive analytical models for energy consumption. Given also that consolidation introduces unpredictable performance due to multitenancy, providing QoS guarantees becomes particularly challenging in this context.

Integration of renewable sources

In addition, the prevalent policy in datacenter server farms is to build a collocated renewable energy source (e.g. wind turbine or photovoltaic) with the objective to minimize the dependence of the datacenter on the main power grid. The output of these renewable sources is highly time-varying and unpredictable. On the other hand, migration of tasks among different datacenters allows us to control the instantaneous power demand, which is a function of load. Given also that electricity prices are time-varying and different from place to place, the arising challenge is to select the VM migration strategy that matches the dynamic power supply of renewable sources and dynamic datacenter power demand at a minimum cost, without violating any user QoS service level agreement (SLA) though.

7.3.3 Energy efficient wireless access networks

Although important steps forward in energy efficiency of the cloud itself have been made, the wireless access network has been generally neglected. In this thesis, we mainly focused on the mobile

device side and we investigated how sleep modes enable better exploitation of the limited available energy. However, a recent survey [96] indicates that the potential energy efficiency gains at the Mobile Network Operators (MNO) side are even more significant, ranging from 10% to 25%.

Energy efficient infrastructure

The first step required for more efficient mobile networks is the identification of the least efficient network components. For this purpose, tools that measure and monitor energy consumption of existing network infrastructure need to be developed. Based on this analysis, an MNO has to explore the available options such as to install renewable energy sources, upgrade energy-intensive network equipment, improve the design of its base stations so as to rely less on air conditioning (e.g. by using techniques like free air cooling). Selecting the optimal one requires a cost-benefit analysis of the available solutions.

Energy efficient control

The main issue of the aforementioned approaches is that they require significant investment. Instead, less radical approaches could provide comparable benefits. Indicatively, dynamic resource allocation algorithms that adapt to the load conditions require only minimal changes of existing infrastructure and hence can be easier adopted by telecom operators. Representative such methods are transmission power control, turning off base stations in low traffic periods and reassociation of mobile users, and scheduling schemes such as bandwidth and capacity adaptation [97], that make use of the minimum amount of resources required to meet the arising traffic.

Bibliography

[1] World Bank, Electric Power Consumption. `https://www.google.gr/publicdata/explore?ds=d5bncppjof8f9_&ctype=l&met_y=eg_use_elec_kh`.

[2] Navigant Research, Report on Residential Demand Response. `http://www.navigantresearch.com/research/residential-demand-response`.

[3] G.Y. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen, and S. Xu. Energy-efficient wireless communications: tutorial, survey, and open issues. *IEEE Wireless Communications*, 18(6), 28–35, 2011.

[4] Network of Excellence in Internet Science. D8.1. Overview of ICT energy consumption. `http://www.internet-science.eu/sites/internet-science.eu/files/biblio/EINS_D8\%201_final.pdf`.

[5] Intel eyes mobile device market, plans new processors, smartchips. `http://www.thehindubusinessline.com/industry-and-economy/info-tech/intel-eyes-mobile-device-market-plans-new-processors-smartchips/article4806223.ece`.

[6] Joint Research Center: Institute for Energy and Transport. Data centres energy efficiency awards 2013. `http://iet.jrc.ec.europa.eu/energyefficiency/sites/energyefficiency/files/presentation_dc_coc_awards_2013_1.pdf`.

[7] Centre for Energy-Efficient Telecommunications (CEET). The power of wireless cloud: An analysis of the energy consumption of wireless cloud. `http://www.ceet.unimelb.edu.au/pdfs/ceet_white_paper_wireless_cloud.pdf`.

[8] H. Allcott. Rethinking real-time electricity pricing. *Resource and Energy Economics*, 33(4):820 – 842, 2011.

[9] O. Ardakanian, S. Keshav, and C. Rosenberg. Markovian models for home electricity consumption. In *Proceedings of the 2nd ACM SIGCOMM workshop on Green networking*, GreenNets '11, pages 31–36, New York, NY, USA, 2011. ACM.

[10] T. Bapat, N. Sengupta, S.K. Ghai, V. Arya, Y.B. Shrinivasan, and D. Seetharam. User-sensitive scheduling of home appliances. In *Proceedings of the 2nd ACM SIGCOMM workshop on Green networking*, pages 43–48. ACM, 2011.

[11] L. Chen, N. Li, L. Jiang, and S.H. Low. Optimal demand response: problem formulation and deterministic case. *Control and Optimization Theory for Electric Smart Grids*, 3(15):63–86, 2012.

[12] ISO New England. Electricity prices/demand data, 2011.

[13] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.

[14] L. Gkatzikis, T. Salonidis, N. Hegde, and L. Massoulie. Electricity markets meet the home through demand response. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 5846 –5851, dec. 2012.

[15] B. Hajek. Performance of global load balancing by local adjustment. *IEEE Transactions on Information Theory*, 36(6):1398 –1414, Nov 1990.

[16] I. Koutsopoulos and L. Tassiulas. Control and optimization meet the smart power grid-scheduling of power demands for optimal energy management. *Arxiv preprint arXiv:1008.3614*, 2010.

[17] J. P. Lawrence and K. Steiglitz. Randomized pattern search. *IEEE Transactions on Computers*, C-21(4):382 –385, April 1972.

[18] N. Li, L. Chen, and S.H. Low. Optimal demand response based on utility maximization in power networks. In *Power and Energy Society General Meeting, 2011 IEEE*, pages 1 –8, July 2011.

[19] A. Mohsenian-Rad, V.W.S. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid. *IEEE Transactions on Smart Grid*, 1(3):320 –331, Dec. 2010.

[20] A.-H. Mohsenian-Rad and A. Leon-Garcia. Optimal residential load control with price prediction in real-time electricity pricing environments. *IEEE Transactions on Smart Grid*, 1(2):120 –133, Sept. 2010.

[21] A.-H. Mohsenian-Rad, V.W.S. Wong, J. Jatskevich, and R. Schober. Optimal and autonomous incentive-based energy consumption scheduling algorithm for smart grid. In *Innovative Smart Grid Technologies (ISGT), 2010*, pages 1 –6, Jan. 2010.

[22] R. Schober P. Samadi and V. W. S. Wong. Optimal energy consumption scheduling using mechanism design for the future smart grid. In *First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct. 2011.

[23] I. Richardson, M. Thomson, D. Infield, and C. Clifford. Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings*, 42(10):1878–1887, 2010.

[24] M. Roozbehani, M.A. Dahleh, and S.K. Mitter. Volatility of power grids under real-time pricing. *Arxiv preprint arXiv:1106.1401*, 2011.

[25] M. Roozbehani, M. Rinehart, M.A. Dahleh, S.K. Mitter, D. Obradovic, and H. Mangesius. Analysis of competitive electricity markets under a new model of real-time retail pricing. In *8th International Conference on the European Energy Market (EEM)*, pages 250 –255, May 2011.

[26] P. Samadi, A. Mohsenian-Rad, R. Schober, V.W.S. Wong, and J. Jatskevich. Optimal real-time pricing algorithm based on utility maximization for smart grid. In *First IEEE International Conference on Smart Grid Communications*, pages 415 –420, Oct. 2010.

[27] H.R. Varian. *Intermediate microeconomics: a modern approach*. WW Norton & Co., 2006.

[28] "Commercially available aggregator programs in California" `http://www.pge.com/mybusiness/energysavingsrebates/demandresponse/`.

[29] "Energy Pool: European Demand Response Operator," `www.energy-pool.eu`.

[30] A. Papavasiliou, H. Hindi, and D. Greene, "Market-based control mechanisms for electric power demand response," in *Proc. IEEE Conf. on Decision and Control (CDC)*, Dec. 2010, pp. 1891 –1898.

[31] H. Kim and M. Thottan, "A two-stage market model for microgrid power transactions via aggregators," *Bell Labs Technical Journal*, vol. 16, no. 3, pp. 101–107, 2011.

[32] I. Koutsopoulos and L. Tassiulas, "Optimal control policies for power demand scheduling in the smart grid," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 6, pp. 1049 –1060, July 2012.

[33] A. Migdalas, P. Pardalos, and P. Värbrand, *Multilevel optimization: algorithms and applications*. Springer, 1997, vol. 20.

[34] N. Gatsis and G. Giannakis, "Cooperative multi-residence demand response scheduling," in *Proc. Conf. on Information Sciences and Systems (CISS), 2011*, March 2011, pp. 1 –6.

[35] L. Huang, J. Walrand, and K. Ramchandran, "Optimal smart grid tariffs," in *Information Theory and Applications Workshop (ITA)*, Feb. 2012.

[36] H. Hindi, D. Greene, and C. Laventall, "Coordinating regulation and demand response in electric power grids using multirate model predictive control," in *Proc. IEEE PES Conf. on Innovative Smart Grid Tech.*, Jan. 2011, pp. 1 –8.

[37] P. Srikantha, C. Rosenberg, and S. Keshav, "An analysis of peak demand reductions due to elasticity of domestic appliances," in *Proc. E-Energy Conf.* New York, NY, USA: ACM, 2012, pp. 28:1–28:10.

[38] L. Gkatzikis and I. Koutsopoulos, "Migrate or Not? Exploiting Dynamic Task Migration in Mobile Cloud Computing Systems," in *IEEE Wireless Communications Magazine: Special Issue on Mobile cloud computing vol.20, no.3, June 2013.*

[39] *"Virtualization industry quarterly survey."* `http://www.v-index.com/`.

[40] *"vsphere resource management guide."* `http://www.vmware.com/support/pubs/vsphere-esxi-vcenter-server-pubs.html`.

[41] *A. K. Mishra, J. L. Hellerstein, W. Cirne, and C. R. Das, "Towards characterizing cloud backend workloads: insights from google compute clusters,"* ACM SIGMETRICS Performance Evaluation Review on Industrial Research*, vol. 37, pp. 34–41, March 2010.*

[42] *"Ganeti : Cluster-based virtualization management software."* `http://code.google.com/p/ganeti/`.

[43] *C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control,"* Math. Oper. Res*, vol. 24, pp. 293–305, 1999.*

[44] *D. Bertsekas, "Dynamic programming and optimal control," 1995.*

[45] *M. Harchol-Balter and A. B. Downey, "Exploiting process lifetime distributions for dynamic load balancing,"* ACM Transactions on Computer Systems*, vol. 15, pp. 253–285, August 1997.*

[46] "Rethink it: Getting rid of noisy neighbours." `http://blogs.vmware.com/rethinkit/2010/09/getting-rid-of-noisy-cloud-neighbors.html`.

[47] "Has amazon ec2 become over subscribed?." `http://alan.blog-city.com/has_amazon_ec2_become_over_subscribed.htm`.

[48] "Mixpanel: Why we moved off the cloud." `http://code.mixpanel.com/2011/10/27/why-we-moved-off-the-cloud/`.

[49] G. Wang and T. Ng, "The impact of virtualization on network performance of amazon ec2 data center," in Proceedings IEEE INFOCOM 2010, pp. 1 –9, March 2010.

[50] A. Iosup, S. Ostermann, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "Performance analysis of cloud computing services for many-tasks scientific computing," IEEE Trans. Parallel Distrib. Syst., vol. 22, pp. 931–945, June 2011.

[51] S.-H. Lim, J.-S. Huh, Y. Kim, G. M. Shipman, and C. R. Das, "D-factor: a quantitative model of application slow-down in multi-resource shared systems," in Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems, SIGMETRICS '12, (New York, NY, USA), pp. 271–282, ACM, 2012.

[52] X. Pu, L. Liu, Y. Mei, S. Sivathanu, Y. Koh, and C. Pu, "Understanding performance interference of i/o workload in virtualized cloud environments," in IEEE 3rd International Conference on Cloud Computing (CLOUD) 2010, pp. 51 –58, July 2010.

[53] Y. Koh, R. Knauerhase, P. Brett, M. Bowman, Z. Wen, and C. Pu, "An analysis of performance interference effects in virtual environments," in IEEE International Symposium on Performance Analysis of Systems Software (ISPASS) 2007., pp. 200 –209, April 2007.

[54] V. Jeyakumar, M. Alizadeh, D. Mazières, B. Prabhakar, and C. Kim, "Eyeq: practical network performance isolation for the multi-tenant cloud," in Proceedings of the 4th USENIX conference on Hot Topics in Cloud Computing, HotCloud'12, (Berkeley, CA, USA), pp. 8–8, USENIX Association, 2012.

[55] K. Srinivasan, S. Yuuw, and T. J. Adelmeyer, "Dynamic vm migration: assessing its risks and rewards using a benchmark," in Proceeding of the second joint WOSP/SIPEW international conference on Performance engineering, ICPE '11, (New York, NY, USA), pp. 317–322, ACM, 2011.

[56] S. Maguluri, R. Srikant, and L. Ying, "Stochastic models of load balancing and scheduling in cloud computing clusters," in INFOCOM, 2012 Proceedings IEEE, pp. 702 –710, March 2012.

[57] T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, "Black-box and gray-box strategies for virtual machine migration," in Proceedings of the 4th USENIX conference on Networked systems design & implementation, NSDI'07, pp. 17–17, 2007.

[58] S. Gitzenis and N. Bambos, "Joint task migration and power management in wireless computing," IEEE Transactions on Mobile Computing, vol. 8, pp. 1189 –1204, Sept. 2009.

[59] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "Clonecloud: elastic execution between mobile device and cloud," in Proceedings of the sixth conference on Computer systems, EuroSys '11, (New York, NY, USA), pp. 301–314, ACM, 2011.

[60] *The power limit of servers,* `http://www.datacenterdynamics.com/focus/` `archive/2013/05/power-limit-servers`.

[61] *Texas Instruments, "cc2420 radio transceiver."* `http://focus.ti.com/lit/ds/` `symlink/cc2420.pdf`.

[62] *R. Jurdak, A. Ruzzelli, and G. O'Hare, "Radio sleep mode optimization in wireless sensor networks,"* IEEE Transactions on Mobile Computing, *vol. 9, pp. 955 –968, July 2010.*

[63] *A. Azad, S. Alouf, E. Altman, V. Borkar, and G. Paschos, "Optimal control of sleep periods for wireless terminals,"* IEEE Journal on Selected Areas in Communications, *vol. 29, pp. 1605 –1617, September 2011.*

[64] *Y. Jin and G. Kesidis, "Equilibria of a noncooperative game for heterogeneous users of an aloha network,"* IEEE Communications Letters, *vol. 6, pp. 282 –284, July 2002.*

[65] *R. Ma, V. Misra, and D. Rubenstein, "An analysis of generalized slotted-aloha protocols,"* IEEE/ACM Transactions on Networking, *vol. 17, pp. 936 –949, June 2009.*

[66] *E. Altman, R. E. Azouzi, and T. Jiménez, "Slotted aloha as a game with partial information,"* Computer Networks, *vol. 45, no. 6, pp. 701–713, 2004.*

[67] *R. El-Azouzi, T. Jiménez, E. S. Sabir, S. Benarfa, and E. H. Bouyakhf, "Cooperative and non-cooperative control for slotted aloha with random power level selections algorithms," in* Value-Tools '07, *pp. 1–10, 2007.*

[68] *Y. Jin and G. Kesidis, "Distributed contention window control for selfish users in ieee 802.11 wireless lans,"* IEEE Journal on Selected Areas in Communications,, *vol. 25, pp. 1113 –1123, August 2007.*

[69] *I. Menache and N. Shimkin, "Efficient rate-constrained nash equilibrium in collision channels with state information," in* IEEE INFOCOM 2008, *2008.*

[70] *L. Chen, S. Low, and J. Doyle, "Random access game and medium access control design,"* IEEE/ACM Transactions on Networking, *vol. 18, pp. 1303 –1316, Aug. 2010.*

[71] *G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function,"* IEEE Journal on Selected Areas in Communications, *vol. 18, pp. 535 –547, March 2000.*

[72] *O. Dousse, P. Mannersalo, and P. Thiran, "Latency of wireless sensor networks with uncoordinated power saving mechanisms," in* Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing, *MobiHoc '04, (New York, NY, USA), pp. 109–120, ACM, 2004.*

[73] *T. Cover and A. Gamal, "Capacity theorems for the relay channel,"* IEEE Transactions on Information Theory,, *vol. 25, pp. 572 – 584, Sep 1979.*

[74] *K. Jacobson and W. Krzymien, "System design and throughput analysis for multihop relaying in cellular systems,"* IEEE Transactions on Vehicular Technology,, *vol. 58, pp. 4514 –4528, Oct. 2009.*

[75] *J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior,"* IEEE Transactions on Information Theory,, *vol. 50, no. 12, pp. 3062–3080, 2004.*

[76] Y. Shi, S. Sharma, Y. T. Hou, and S. Kompella, "Optimal relay assignment for cooperative communications," in MobiHoc '08: Proc. of the 9th ACM international symposium on Mobile ad hoc networking and computing, (NY, USA), pp. 3–12, ACM, 2008.

[77] J. Huang, Z. Han, M. Chiang, and H. Poor, "Auction-based resource allocation for cooperative communications," IEEE Journal on Selected Areas in Communications, vol. 26, pp. 1226 –1237, September 2008.

[78] B. Wang, Z. Han, and K. J. R. Liu, "Distributed relay selection and power control for multiuser cooperative communication networks using stackelberg game," IEEE Transactions on Mobile Computing, vol. 8, no. 7, pp. 975–990, 2009.

[79] K. Sundaresan and S. Rangarajan, "On exploiting diversity and spatial reuse in relay-enabled wireless networks," in MobiHoc '08: Proc. of the 9th ACM international symposium on Mobile ad hoc networking and computing, (NY, USA), pp. 13–22, ACM, 2008.

[80] "IEEE Standard for local and metropolitan area networks part 16: Air interface for broadband wireless access systems amendment 1: Multiple relay specification," IEEE Std 802.16j-2009 (Amendment to IEEE Std 802.16-2009), pp. c1–290, 12 2009.

[81] 3GPP, "LTE-Advanced." http://www.3gpp.org/article/lte-advanced.

[82] IEEE, "802.16m." http://www.ieee802.org/16/tgm/.

[83] V. Genc, S. Murphy, and J. Murphy, "Performance analysis of transparent relays in 802.16j MMR networks," in 6th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks and Workshops, 2008. WiOPT 2008., pp. 273 –281, April 2008.

[84] G. Paschos, P. Mannersalo, and T. Bohnert, "Cell capacity for ieee 802.16 coverage extension," in 5th IEEE Consumer Communications and Networking Conference, 2008. CCNC 2008., pp. 933 –937, Jan. 2008.

[85] K. Jacobson and W. A. Krzymie, "Minimizing co-channel interference in wireless relay networks," in International Conference on Wireless Communications, (Calgary, AB, Canada), pp. pp. 339–346, 2005.

[86] S. W. Peters, A. Y. Panah, K. T. Truong, and R. W. Heath, "Relay architectures for 3gpp lte-advanced," EURASIP Journal on Wireless Communications and Networking, vol. 2009, pp. 1–14, 2009.

[87] A. Reddy, S. Shakkottai, and L. Ying, "Distributed power control in wireless ad hoc networks using message passing: Throughput optimality and network utility maximization," in 42nd Annual Conference on Information Sciences and Systems, 2008. CISS 2008., pp. 770 –775, march 2008.

[88] A. Host-Madsen and J. Zhang, "Capacity bounds and power allocation for wireless relay channels," IEEE Transactions on Information Theory,, vol. 51, no. 6, pp. 2020–2040, 2005.

[89] M. Bayati, D. Shah, and M. Sharma, "Max-product for maximum weight matching: Convergence, correctness, and lp duality," IEEE Transactions on Information Theory,, vol. 54, pp. 1241–1251, March 2008.

[90] *R. L. Cruz and A. V. Santhanam, "Optimal routing, link scheduling and power control in multi-hop wireless networks," in* Proc. IEEE INFOCOM, *(San Francisco, USA), 2003.*

[91] *C. W. Tan, M. Chiang, and R. Srikant, "Fast algorithms and performance bounds for sum-rate maximization in wireless networks," in* Proc. IEEE INFOCOM, *(Rio de Janeiro, Brazil), 20-25 Apr 2009.*

[92] *H. W. Kuhn, "The hungarian method for the assignment problem,"* Naval Research Logistics Quarterly, *vol. 2, no. 1-2, pp. 83–97, 1955.*

[93] *S. Khuller, S. G. Mitchell, and V. V. Vazirani, "On-line algorithms for weighted bipartite matching and stable marriages," in* ICALP, *pp. 728–738, 1991.*

[94] *S. Deb, D. Shah, and S. Shakkottai, "Fast matching algorithms for repetitive optimization: An application to switch scheduling," in* 40th Annual Conference on Information Sciences and Systems, *pp. 1266 –1271, March 2006.*

[95] *L. Gkatzikis, G. Iosifidis, I. Koutsopoulos and L. Tassiulas, "Collaborative Placement and Use of Storage Resources in the Smart Grid"* IEEE under review in IEEE International Conference on Smart Grid Communications (SmartGridComm), *2013*

[96] *GSMA, Mobile Energy Efficiency Benchmarking.* `http://www.gsma.com/publicpolicy/mobile-energy-efficiency/mobile-energy-efficiency-benchmarking`.

[97] *A. Ambrosy, O.Blume, H. Klessig and W. Wajda, "Energy saving potential of integrated hardware and resource management solutions for wireless base stations," submitted in* IEEE International Conference on Smart Grid Communications (SmartGridComm), *2013*