Multiple Access Computation Offloading

MULTIPLE ACCESS COMPUTATION OFFLOADING

BY

MAHSA SALMANI, M.A.Sc. (Electrical and Computer Engineering),

McMaster University,

Hamilton, Ontario, Canada

A THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2019) McMaster University

(Department of Electrical and Computer Engineering) Hamilton, Ontario, Canada

TITLE:	Multiple Access Computation Offloading
AUTHOR:	Mahsa Salmani M.A.Sc. (Electrical and Computer Engineering), McMaster University, Hamilton, Ontario, Canada
SUPERVISOR:	Prof. Timothy N. Davidson
NUMBER OF PAGES:	xx, 184

To my love

Foad

&

to my beloved parents

Zahra and Mohammadhossein

Lay Abstract

The rapid increase in the number of smart devices in wireless communication networks, and the expansion in the range of computationally-intensive and latency-sensitive applications that those devices are required to support, have highlighted their resource limitations in terms of energy, power, central processing unit (CPU), and memory. Mobile edge computing is a framework that provides shared computational resources at the access points of wireless networks and gives such devices the opportunity to offload (a portion of) their applications to be executed at the access points. In order to fully exploit such an opportunity when multiple devices seek to offload their applications, the available communication and computation resources must be efficiently allocated amongst those devices. The ultimate goal of this thesis is to obtain the optimal communication resource allocation in a K-user offloading system while different constraints on the devices and on the applications are satisfied. To that end, this thesis shows that the minimum energy consumption is obtained when the system exploits the full capabilities of the channel, the maximum allowable latency of each user, and the differences between the latency constraints of each user. Accordingly, this thesis proposes an optimized signalling structure and, based on that structure, low-complexity algorithms that achieve an energy-optimal resource allocation in a K-user offloading system.

Abstract

The limited energy and computational resources in small-scale smart devices impede the expansion of the range of applications that those devices can support, especially to applications with tight latency constraints. Mobile edge computing is a promising framework that provides shared computational resources in the access points in the network and provides devices in that network with the opportunity to offload (a portion of) their computational tasks to the access points. To effectively capture that opportunity in an offloading system with multiple devices, the available communication and computation resources must be efficiently allocated. The main focus of this thesis is on the optimal allocation of communication resources in a K-user offloading system. The resource allocation problem that is considered in this thesis captures minimizing the total energy consumption of users while the requirements of the users, and their computational tasks, are met. That problem is addressed for two of the most widely-considered classes of computational tasks in the literature, namely, indivisible tasks (binary offloading) and divisible tasks (partial offloading).

This thesis begins with an exploration of the impact of the choice of multiple access scheme that is employed by the system on the total energy consumption of the users. In particular, the problem of minimizing the total energy consumption of a two-user binary offloading system is tackled under various multiple access schemes, namely

time division multiple access (TDMA), sequential decoding without time sharing, independent decoding, and multiple access schemes that can exploit the full capabilities of the channel, which are referred to as full multiple access schemes (FullMA) in this thesis. Using a decomposition-based approach, closed-form solutions to the resource allocation problem are obtained. Those expressions show that by exploiting the full capabilities of the channel, a FullMA scheme can significantly reduce the total energy consumption of the users as compared to the other schemes. The closed-form expressions also show that when the channel gains of the two users are equal, the TDMA scheme can achieve the optimal energy consumption. For the case of partial offloading, an analogous analysis leads to a reduced-dimension design problem and an extension to the optimally result for TDMA.

In the next step of the development, the insights obtained from the decomposition-based analysis of the two-user case are used to tackle the communication resource allocation problem for a K-user offloading system in which the users are assumed to be served over a single time slot. Based on their performance in the two-user case, FullMA and TDMA schemes are considered. The mixed-integer optimization problem that arises in the binary offloading case is addressed by employing a decomposition approach with a closed-form expression obtained for the optimal resource allocation for given offloading decisions, and a tailored pruned greedy search algorithm developed herein for the offloading decisions. By exploiting the maximum allowable latency of each individual user, the proposed algorithm is able to significantly reduce the energy consumption of the users in comparison to the existing algorithms in the literature that assume equal latency constraints for all users. Furthermore, with the closed-form optimal solution to the resource allocation problem obtained for given

offloading decisions, the proposed algorithm has a significantly lower computational cost compared to the existing algorithms. In the partial offloading case, a quasi-closed-form solution is obtained for the resource allocation problem.

Finally, a time-slotted signalling structure is proposed as an optimal transmission structure for a generic K-user offloading system. Furthermore, an optimal time-slotted structure that requires only K time slots is developed for a K-user offloading system that employs a FullMA scheme. The proposed time-slotted structure not only exploits the maximum latency constraint of each user, it also exploits the differences between the latency constraints of the users by taking advantage of the interference reduction that arises when a user finishes offloading. The proposed time-slotted FullMA signalling structure significantly reduces the energy consumption of the users compared to some existing methods that employ the TDMA scheme, and compared to those with FullMA, but sub-optimal single-time-slot signalling structures. Moreover, the computational cost of the proposed time-slotted algorithm is significantly lower than that of the existing algorithms in the literature.

Acknowledgements

First and foremost I would express my deepest appreciation and gratitude to my supervisor, Prof. Timothy N. Davidson, for all his technical guidance and generous support throughout my Ph.D. study. It would have been impossible to complete this work without his motivation, thoughtful insights and continuous support. I would also like to thank him for all his tireless efforts and endless patience in revising and improving my academic papers. Looking back to the past few years, I am deeply delighted that I had the opportunity to work under his supervision which helped me to grow both technically and personally.

I would also like to express my gratitude to my supervisory committee, Dr. Jian Kang Zhang and Dr. Dongmei Zhao, for their invaluable comments and technical support during my Ph.D. study.

I would like to thank all my beloved friends and colleagues at McMaster University. In particular, I express my gratitude to Atta Ziaee, Darya Emami, Ehsan Taghavi, Mina Attari, Pooyan Mehrvarzi, Sahand Sepehrvand, and Yasamin Fazliani whose presence has always been priceless and made my Ph.D. journey more enjoyable and fruitful. I would also like to take this opportunity to thank the administrative team of McMaster ECE department, especially Cheryl Gies, Kelly Lyth, and Tracey Coop. My special thanks go to Joe Peric, not only for all his technical help, but for his

friendly presence and support.

I am truly grateful to my beloved parents, Zahra and Mohammadhossein, for their immeasurable love and unconditional support. They have always encouraged me to explore what I have dreamed. And here goes my special thanks to my little sister, Parisa, the presence of whom, from the very first seconds, has warmed my heart.

Last but by no means least, I would like to express my deepest gratitude to my forever best friend and love, Foad, who has always been beside me in all stages of my Ph.D. study. I would like to thank him for all his endless love, encouragements and support throughout my research.

Abbreviations

CPU	Central Processing Units
EL-FullMA	Equal-Latency Full Multiple Access
FDMA	Frequency Division Multiple Access
FullMA	Full Multiple Access
ID	Independent Decoding
IEEE	Institute of Electrical and Electronics Engineers
MCC	Mobile Cloud Computing
MEC	Mobile Edge Computing
OFDMA	Orthogonal Frequency Division Multiple Access
SDwts	Sequential Decoding without time sharing
STS-FullMA	Single-Time-Slotted Full Multiple Access
TDMA	Time Division Multiple Access
TS-FullMA	Time-Slotted Full Multiple Access

Contents

List of Figures

Chapter 1

Introduction

The increases in the number of smart devices and in the provision of wireless communication networks have enabled the realization of near ubiquitous access to a range of computational applications and services that previously could only be envisioned (e.g., Weiser, 1991). Although those developments are promising, they have only fuelled the desire to expand the range of applications that the devices are able to support, including games, voice and video processing, online social networks, autonomous driving, and augmented reality. The limitations in the computational, storage and energy resources in smart devices have become significant impediments to progress towards providing those applications (Cisco, 2017). Although there have been improvements in the central processing units (CPU), storage and battery lifetime of mobile devices, they are not currently capable of meeting the computational and energy requirements of computationally and energy intensive applications. To address those issues, two broad approaches have been developed to provide the opportunity for the devices to "offload" their computational tasks.

1.1 Mobile Cloud Computing

The first of those approaches is Mobile Cloud Computing (MCC) (e.g., Dinh *et al.*, 2013; Fernando *et al.*, 2013; Khan *et al.*, 2014; Rahimi *et al.*, 2014). MCC provides powerful shared computational and storage resources in the core of the network. Those resources enable the devices to execute more complex applications using different computation offloading techniques. However, the increasing number of mobile devices in communication networks pose several challenges regarding the latency and security requirements that MCC must satisfy. In particular, the limited bandwidth that is required to be shared among the devices in the network, along with the long propagation distance between the mobile devices and the cloud centre, impedes the use of MCC to support the applications with low-latency and real-time execution requirements (Dinh *et al.*, 2013; Mao *et al.*, 2017b; Abbas *et al.*, 2018).

1.2 Mobile Edge Computing

In order to tackle the challenges that MCC has faced, a new paradigm named Mobile Edge Computing (MEC) has been developed. In MEC the computation and memory resources are provided at the access points and base stations of the network, rather than in a central cloud in the core of network (Patel *et al.*, 2014). In that way, the computation and storage functions required by the users are pushed to the "edges" of the communication network (Mach and Becvar, 2017; Mao *et al.*, 2017b; Abbas *et al.*, 2018). By employing a direct connection between the mobile devices and the computation resources, MEC can reduce the propagation time and networking delays

as compared to MCC. Accordingly, MEC has the potential to support computation-intensive applications with low-latency and context-awareness requirements, such as augmented reality (Mao *et al.*, 2017b; Abbas *et al.*, 2018).

1.3 Computation Offloading

By providing additional computational resources, the MEC framework gives mobile devices the opportunity to offload (a fraction of) their computational tasks to the shared computational resources so that the (fraction of the) tasks can be executed in those resources rather than the mobile devices (locally). Some early prototypes of mobile computational offloading systems include MAUI (Cuervo *et al.*, 2010), CloneCloud (Chun *et al.*, 2011), and ThinkAir (Kosta *et al.*, 2012).

One of the main challenges in computation offloading is the decision on whether or not each device's computational task should be offloaded (Mach and Becvar, 2017; Mao *et al.*, 2017b). Prior to all the factors in the communication network that can affect the offloading decisions, such as the availability of communication and computation resources and the required latencies of the users, it is the nature of the computational tasks that determines the form that the available offloading opportunities take. There are two main classes of the computational tasks that are considered in the literature in this area (Khan *et al.*, 2014; Khan, 2015; Mach and Becvar, 2017; Akherfi *et al.*, 2018), namely, indivisible tasks and divisible tasks. Indivisible tasks are tasks in which the computational components are so tightly coupled that the task cannot be partitioned. Accordingly, indivisible tasks are either totally offloaded or completed locally in the devices. On the other hand, if the computational task has independent or loosely coupled components, the mobile device can benefit from the

potential parallelism between the access point and the mobile device by offloading a portion of its computational tasks while the remainder is executed locally in the device. Accordingly, this thesis will focus on two broad classes of offloading systems:

- Binary offloading: In this case the computational tasks are (considered to be) indivisible, and hence each user must either offload the whole task or execute it locally.

- Partial offloading: In this case the computational tasks can be partitioned into different parts and each part is either offloaded or executed locally. In partial offloading case, the mobile devices can benefit from the potential parallelism between the processors in the mobile devices and in the access point. This thesis considers "data-partitionable" indivisible computational tasks (Wang *et al.*, 2016), in which a simple-to-describe operation is applied, independently, to different blocks of data.

1.3.1 Resource Allocation in Computation Offloading

When there are multiple devices that are seeking to offload their computational tasks, the offloading system must address a number of challenges, including the energy that each user would expend to offload (a portion of) its computational task to the access point (Barbera *et al.*, 2013), the latency requirements of the tasks (Lei *et al.*, 2013), contention for the limited communication resources of the networks (Mao *et al.*, 2017b), and, in some cases, contention for the limited shared computation resources at the access point (Liu *et al.*, 2013; Salmani and Davidson, 2017b). To address those challenges, the problem of deciding which of the devices should offload their tasks to the access point, and, in the case of partial offloading, which fractions of

the tasks should be offloaded, can be formulated as a joint optimization problem over the available computation and communication resources (e.g., Sardellitti *et al.*, 2015; Chen *et al.*, 2015; Muñoz *et al.*, 2015, 2014; Wang *et al.*, 2017a, 2018a; Mao *et al.*, 2017a). That formulation typically captures the energy that would be required to complete the computational task locally on the mobile device and the latency incurred in doing so, and the energies and latencies associated with transmitting the required information to the shared computational resources, completing the task there and returning the results to the mobile device.

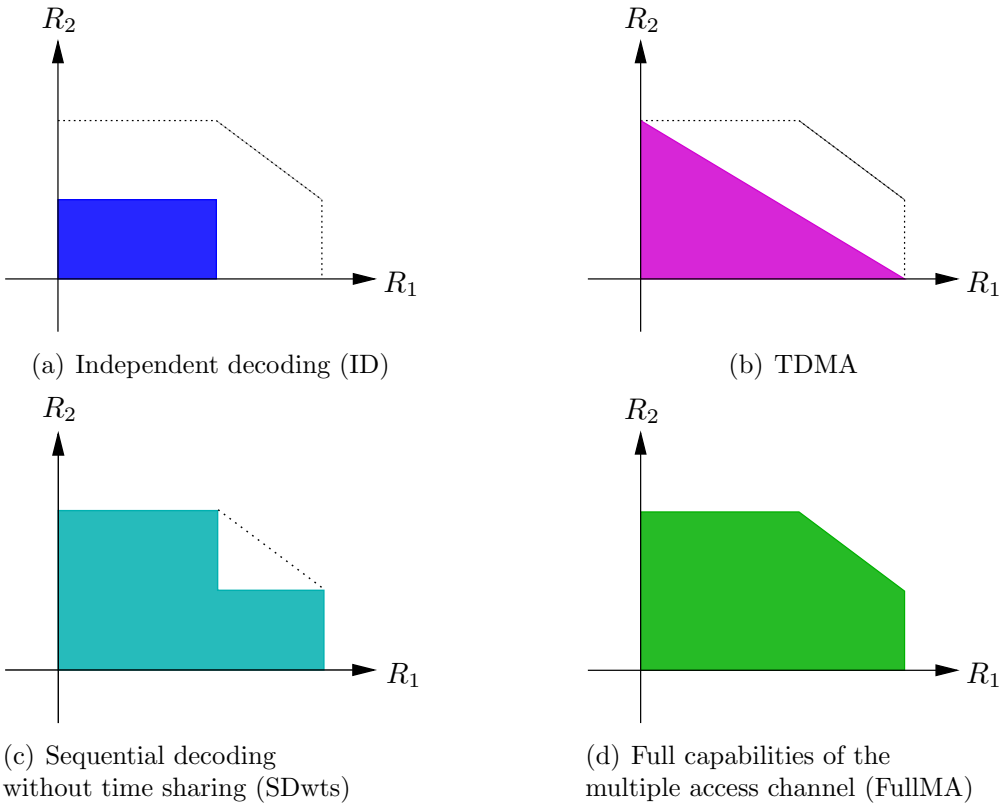
The allocation of the available communication resources to the offloading devices, with the aim of reducing the total user energy consumption, is fundamentally dependent on the multiple access scheme that is employed by the multi-user offloading system. The multiple access scheme determines the transmission strategy of the offloading devices, and hence it determines the set of transmission rates at which the devices can communicate reliably with the access point. Accordingly, the achievable rate region of the multiple access scheme imposes constraints on the time and the energy that each of the devices is required to expend to offload their tasks to the access point. Most of the existing literature in this area considers orthogonal multiple access schemes, such as time division multiple access (TDMA) and frequency division multiple access (FDMA) (e.g., Sardellitti *et al.*, 2015; Chen *et al.*, 2016a,b; You *et al.*, 2017; Mao *et al.*, 2017a). There are other works that assume that the access point performs independent decoding (e.g., Sardellitti *et al.*, 2015) or ordered minimum mean square error successive interference cancellation (e.g., Wang *et al.*, 2017a). For a simple two-user system that employs a single time slot, the achievable rate regions are illustrated in Fig. 2.1, where the shaded regions are the achievable rate regions

for (a) independent decoding (ID), (b) TDMA, and (c) sequential decoding without time sharing (SDwts). Part (d) illustrates the capacity region.

It can be seen from Fig. 2.1 that assigning orthogonal channels to the offloading devices, or employing independent decoding or a successive decoding scheme without time sharing, can indeed limit the range of the achievable rates by which the mobile devices can transmit to the access point. As a result, they cannot, in general, obtain the optimal energy consumption of the offloading system. Motivated by that observation, a core contribution of this thesis is a characterization of the performance of offloading systems that employ a multiple access scheme that exploits the full capabilities of the channel. In particular, this thesis addresses the problem of minimizing the energy consumption of an offloading system in which a capacity achieving multiple access scheme (which is called a full multiple access (FullMA) scheme in this thesis) is employed (Cover and Thomas, 2012).[1] Furthermore, the conditions under which the sub-optimal multiple access schemes, such as TDMA and independent decoding, can achieve the optimal energy consumption are determined.

The problem of minimizing the energy consumption of an offloading system in a single-user scenario has been tackled by Kumar and Lu (2010); Zhang *et al.* (2013); Sardellitti *et al.* (2015); Zhang *et al.* (2012), and Mahmoodi *et al.* (2019). Kumar and Lu (2010) study the general conditions under which a single user can save energy by fully offloading its computation workload in a simplified communication network. Zhang *et al.* (2013) propose an offloading strategy to minimize the energy consumption of the single user. In particular, Zhang *et al.* (2013) minimize the energy consumption

[1]That is, a scheme that enables operation at rates that approach the boundary of the capacity region. Examples include Gaussian signalling with joint decoding, and Gaussian signalling with optimally-ordered sequential decoding and time sharing (Cover and Thomas, 2012; El-Gamal and Cover, 1980). In this thesis, FullMA will denote any such "capacity-approaching" multiple access scheme.



(a) Independent decoding (ID)

(b) TDMA

(c) Sequential decoding without time sharing (SDwts)

(d) Full capabilities of the multiple access channel (FullMA)

Figure 1.1: The set of achievable rates for a two-user system that employs a single time slot under different multiple access schemes.

of local execution in the user by optimally setting the clock frequency of the user. Partial computation offloading in a single-user offloading system has also been studied by Zhang *et al.* (2012) and Mahmoodi *et al.* (2019). Zhang *et al.* (2012) assume a sequential dependency among the computation components of the user's task and propose an efficient partitioning algorithm to minimize the energy consumption in the user, while Mahmoodi *et al.* (2019) develop an offloading decision for any arbitrary component dependencies.

Minimizing the energy consumption of an offloading system in which multiple users seeking to offload their computational tasks is more complicated than the single-user case, due to the fact that the available communication and computation resources must be efficiently allocated to the users. In that case, the design of the communication network (e.g., the multiple access scheme that is employed by the offloading system), plays important roles in the energy consumption of the offloading system. There are several work in the literature addressing the resource allocation problem in an offloading system with multiple users (e.g., Chen *et al.*, 2016c; Sardellitti *et al.*, 2015; Chen *et al.*, 2018; Wang *et al.*, 2018a; Mao *et al.*, 2017a; You *et al.*, 2017; Wang *et al.*, 2018b). Sardellitti *et al.* (2015) and Wang *et al.* (2018a) formulate the energy minimization problem in an offloading system with the TDMA scheme for binary and partial offloading cases, respectively. Chen *et al.* (2018) investigate the joint communication and computation resource allocation in an offloading system with multiple users, each having multiple tasks, when the available bandwidth is shared among the users in an orthogonal way. The frequency division multiple access scheme is considered as the multiple access scheme of a multi-user partial offloading system by Mao

et al. (2017a). Chen *et al.* (2016c) develop a resource allocation algorithm for a multi-user partial offloading system in which the receiver treats the interference among the users as noise and employs independent decoding. You *et al.* (2017) address the energy minimization problem of a multi-user partial offloading system when TDMA and OFDMA schemes are employed by the system, but they constrain the users to have the same latency. As seen above, most of the existing work in the literature consider orthogonal (in time or in frequency) channels for the users in an offloading system. The non-orthogonal multiple access (NOMA) scheme has been considered by Wang *et al.* (2018b). They tackle the communication resource allocation in a multi-user offloading system when all the users are constrained to operate under a pre-defined latency constraint, which is the shortest latency constraint among the users.

1.4 Thesis Contributions

The main goal of this thesis is to tackle the problem of minimizing the total energy consumption of the users in an offloading system when a capacity achieving multiple access (FullMA) scheme is employed as the transmission scheme. To the best of the author's knowledge, the approaches proposed in this thesis are the first ones to address the optimal resource allocation problem under a full multiple access scheme. In order to obtain the energy-optimal communication resource allocation in a multi-user offloading system, a "time-slotted" signalling structure is introduced, and the optimal time-slotted signalling structure for any full multiple access scheme is proposed.

This thesis has been assembled in a "sandwich-style" format, with each technical chapter representing a submitted journal article. As explained below, these articles represent the independent work of the author of this thesis, Mahsa Salmani. In each

of the submitted journal articles, variants of which are included in this thesis as Chapters 2, 3, and 4, the problem formulations were conceived, the mathematical analysis was performed, and the numerical experiments were carried out by Mahsa Salmani under the supervision of Prof. Timothy N. Davidson. The first draft of each of the articles was written by Mahsa Salmani, and was revised in conjunction with Prof. Davidson.

1.4.1 Chapter 2: Multiple Access Computational Offloading: Communication Resource Allocation in the Two-User Case

This chapter seeks insight into the impact of the choice of the multiple access scheme by developing solutions to the mobile energy minimization problem in the two-user case when plentiful shared computational resources are available. In that setting, the allocation of communication resources is constrained by the latency constraints of the applications, the computational capabilities and the transmission power constraints of the devices, and the achievable rate region of the chosen multiple access scheme. For both indivisible tasks and the "data-partitionable" divisible tasks (Wang *et al.*, 2016), this chapter provides closed-form and quasi-closed-form solutions, respectively, for systems that can exploit the full capabilities of the multiple access channel, and for systems based on time division multiple access (TDMA). For indivisible tasks, quasi-closed-form solutions for systems that employ sequential decoding without time sharing or independent decoding are also provided. Analysis of the closed-form and quasi-closed-form solutions show that when the channel gains are equal and the transmission power budgets are larger than a threshold, TDMA (and

the other suboptimal multiple access schemes that have been considered) can achieve an optimal solution. However, when the channel gains of each user are significantly different and the latency constraints are tight, systems that take advantage of the full capabilities of the multiple access channel can substantially reduce the energy required to offload.

Preliminary versions of portions of this work were published in a number of conference proceedings. Salmani and Davidson (2016) outline some of the core analytical work for the case of binary offloading with a FullMA scheme. Salmani and Davidson (2017c) describe the first results for the case of partial offloading in a two-user system that employs a FullMA scheme. The first observations on the energy efficiency of a time-slotted structure are stated by Salmani and Davidson (2017a), in which energy minimization in the binary offloading case under suboptimal multiple access schemes, namely TDMA, sequential decoding without time sharing, and independent decoding is studied. The material in this chapter has been posted on arxiv (available at https://arxiv.org/abs/1805.04981v2), and a condensed version has been submitted to the *IEEE Transactions on Signal Processing* as "Multiple Access Computational Offloading: Communication Resource Allocation in the Two-User Case" with authors Mahsa Salmani and Timothy N. Davidson.

1.4.2 Chapter 3: Uplink Resource Allocation for Multiple Access Computational Offloading

Having obtained closed-form and quasi-closed form solutions for the optimal resource allocation problem in a two-user offloading system in Chapter 2, the main focus of this chapter of the thesis is to find the minimum total user energy consumption of

an offloading system with K users that operate in a single time slot. As mentioned above, in most of the previous work on K-user offloading systems, the multiple access schemes employed by the system are restricted, e.g., the channels between the users and the access point are sometimes assumed to be orthogonal in time or frequency (You *et al.*, 2017; Chen *et al.*, 2018), or it is assumed that the access point performs independent decoding (Sardellitti *et al.*, 2015). In other cases, it is assumed that all the users have the same latency (Wang *et al.*, 2018b).

In Chapter 3 of this thesis, efficient algorithms for optimizing the communication resource allocations for the multiple access scheme that exploits the full capabilities of the multiple access channel and the maximum allowable latency of each user are provided. As in Chapter 2, this chapter considers energy minimization problems for both indivisible computational tasks and data-partitionable divisible tasks (Wang *et al.*, 2016). For the indivisible case, the combinatorial structure of the binary offloading problem of deciding which users will offload their tasks and which will complete them locally suggests a natural decomposition into an outer search strategy for the offloading decisions and the inner optimization of the communication resources for given offloading decisions. That inner subproblem is referred to herein as the "complete offloading" problem. For the case of partial offloading of data-partitionable tasks, the fraction of each task to be offloaded will be optimized jointly with the communication resource allocation. The proposed strategy for solving the resource allocation problems is based on the insights developed for the two-user case in Chapter 2, which suggest the application of algebraic decompositions of the problem. The polymatroid structure of the capacity region of the multiple access channel (see Tse and Hanly, 1998) is exploited to obtain closed-form optimal solutions for the powers in terms of

the transmission rates in both the complete offloading and partial offloading cases. In the complete offloading case, closed-form optimal solutions for the transmission rates of the users are derived, and hence the optimal energy consumption is obtained. Those solutions also form the core of a tailored greedy search algorithm that is developed to find good solutions to the binary offloading problem. In the partial offloading case, in addition to closed-form optimal solutions for the transmission powers, closed-form optimal solutions for the fraction of bits offloaded by each user are also obtained. Accordingly, the energy minimization problem is reduced to a K-variable optimization problem in terms of the transmission rates, for which a simple coordinate descent algorithm is guaranteed to find a stationary solution.

A portion of this chapter is published as a conference paper (Salmani and Davidson, 2018a), in which the energy minimization problem in a K-user binary offloading system under a FullMA scheme is addressed. The material in this chapter has been posted on arxiv (available at https://arxiv.org/abs/1809.07453v2), and a condensed version has been submitted to the *EURASIP J. Signal Process.* as "Uplink Resource Allocation for Multiple Access Computational Offloading" with authors Mahsa Salmani and Timothy N. Davidson.

1.4.3 Chapter 4: Energy-optimal Time-slotted Multiple Access Computational Offloading

In Chapter 3, an efficient resource allocation algorithm for a K-user offloading system in which all the users are operating over a single time slot has been obtained. Although the algorithm developed in that case exploits the maximum allowable latency of each of the users, the single-time-slot structure meant that it could not

consider the differences between the latency constraints of the users. In particular, it does not exploit the reduction in interference that occurs when a user completes its transmission.

This chapter seeks to find an energy-optimal communication resource allocation for a K-user offloading system when not only the maximum allowable latency of each user, but also the differences between those latencies, are taken into account. First, a time-slotted signalling structure is introduced that exploits the differences between the users' latencies by taking advantage of the interference reduction that arises when a device completes its offloading. Then, an optimized time-slotted structure for a multiple access scheme that exploits the full capabilities of the channel (FullMA) is obtained. The optimized signalling structure enables the offloading system to substantially reduce the dimension of the resource allocation problem, and it leads to efficient algorithms to tackle that problem for both indivisible tasks and data-partitionable divisible tasks. The numerical experiments provided in this chapter illustrate that the proposed time-slotted FullMA signalling structure significantly reduces the energy consumption of the devices compared to some existing methods that employ orthogonal multiple access schemes, such as TDMA, and compared to those with FullMA, but sub-optimal single-time-slot signalling structures.

A portion of this chapter, in which the resource allocation problem in a K-user system in the case of binary offloading is studied, is published as a conference paper (Salmani and Davidson, 2019a). The material of this chapter has been submitted to the *IEEE Transactions on Signal Processing* as "Energy-Optimal Time-Slotted Multiple Access Computation Offloading" with authors Mahsa Salmani and Timothy N. Davidson.

1.4.4 Chapter 5: Conclusion and Future Work

This chapter summarizes the main contributions and the conclusions of the thesis, and outlines some of the potential directions for future work.

Chapter 2

Multiple Access Computational Offloading: Communication Resource Allocation in the Two-User Case

2.1 Abstract

By offering shared computational facilities to which mobile devices can offload their computational tasks, the mobile edge computing framework is expanding the scope of applications that can be provided on resource-constrained devices. When multiple devices seek to use such a facility simultaneously, both the available computational resources and the available communication resources need to be appropriately allocated. In this chapter, we seek insight into the impact of the choice of the multiple

access scheme by developing solutions to the mobile energy minimization problem in the two-user case with plentiful shared computational resources. In that setting, the allocation of communication resources is constrained by the latency constraints of the applications, the computational capabilities and the transmission power constraints of the devices, and the achievable rate region of the chosen multiple access scheme. For both indivisible tasks and the limiting case of tasks that can be infinitesimally partitioned, we provide a closed-form and quasi-closed-form solution, respectively, for systems that can exploit the full capabilities of the multiple access channel, and for systems based on time-division multiple access (TDMA). For indivisible tasks, we also provide quasi-closed-form solutions for systems that employ sequential decoding without time sharing or independent decoding. Analyses of our results show that when the channel gains are equal and the transmission power budgets are larger than a threshold, TDMA (and the suboptimal multiple access schemes that we have considered) can achieve an optimal solution. However, when the channel gains of each user are significantly different and the latency constraints are tight, systems that take advantage of the full capabilities of the multiple access channel can substantially reduce the energy required to offload.

2.2 Introduction

The widespread adoption of mobile computing devices and wireless communication networks has enabled the development of applications and services that previously could only be envisioned (e.g., Weiser, 1991). The success of these developments is fuelling the ambition for future applications and services, but as that ambition has grown, the modest computational, storage and energy resources of the mobile devices

have become significant constraints. The mobile cloud, mobile edge, and fog computing frameworks seek to address those constraints by providing shared computational resources to which mobile devices can offload their computational tasks, or a portion thereof (e.g., Satyanarayanan *et al.*, 2009; Kumar and Lu, 2010; Liu *et al.*, 2013; Miettinen and Nurminen, 2010; Kumar *et al.*, 2013; Fernando *et al.*, 2013; Barbera *et al.*, 2013; Mao *et al.*, 2017b). Offloading offers the potential for the mobile device to obtain the results of computationally-intensive or memory-intensive tasks more quickly than would be possible using local computation, and it also offers the potential to better manage the battery life of the device. Some early prototypes of mobile computational offloading systems include MAUI (Cuervo *et al.*, 2010), CloneCloud (Chun *et al.*, 2011), and ThinkAir (Kosta *et al.*, 2012).

The offloading opportunities provided by the mobile cloud computing framework need to be balanced against the energy required to communicate reliably with the networking infrastructure that connects to the shared computing resources (Barbera *et al.*, 2013), the latency of that communication, the contention for the limited communication resources of the network (Lei *et al.*, 2013), and the contention for the limited computational resources of the cloud (Liu *et al.*, 2013). Indeed, the problem of deciding when and how to exploit the resources provided by the mobile cloud computing framework can be formulated as a joint optimization problem over the available computational and communication resources (e.g., Sardellitti *et al.*, 2015; Chen *et al.*, 2015; Muñoz *et al.*, 2015, 2014; Salmani and Davidson, 2016; Wang *et al.*, 2017a, 2018a). That formulation typically captures the energy that would be required to complete the computational task locally (on the mobile device) and the latency incurred in doing so, and the energies and latencies associated with transmitting the

required information to the shared computational resources, completing the task there and returning the results to the mobile device. When the components of the task at hand are tightly coupled, the task is often considered to be indivisible, and hence the decision of whether or not to offload the task is a binary decision. When the task can be partitioned into separate components, the system can take advantage of the implicit parallelism between the mobile user and the access point. As a result, the formulation of the offloading problem may include decisions on which components to offload, or, in the limit, what fraction of the task to offload. When there are multiple devices that are seeking access to the computing resources, the architecture of the envisioned system will determine whether the offloading decisions are to be made centrally, or in a distributed fashion. As this discussion suggests, in the general case, the problem of deciding whether or not to offload (a fraction of) a computational task can be a computationally demanding problem in and of itself. Therefore, the development of insight into the structure of good solutions has the potential to guide the development of practical algorithms.

In this chapter we seek to develop insight into the impact of the choice of the multiple access scheme in a multiuser offloading system in which the allocation of resources is performed centrally. In previous work on such systems, the multiple access scheme has been chosen *a priori*. For example, the users' channels may be assumed to be orthogonal (as they are in time division multiple access, TDMA, and frequency division multiple access, FDMA) (Sardellitti *et al.*, 2015; Chen *et al.*, 2016a,b; You *et al.*, 2017), or it may be assumed that the access point performs independent decoding (Sardellitti *et al.*, 2015) or ordered minimum mean square error successive interference cancellation (Wang *et al.*, 2017a). These choices can limit the set of rates

at which the users can communicate reliably with the access point, and hence they can constrain the potential of computational offloading. In the main contribution of this chapter, we do not place constraints on the multiple access scheme and that enables us to take the advantage of the full capabilities of the wireless channel.

In order to develop insight into the impact of the choice of the multiple access scheme and the corresponding allocation of communication resources, we will consider a two-user scenario with a single access point that is equipped with plentiful computational resources. Each user's task has a separate latency constraint, and we will consider both indivisible tasks and the limiting case of infinitesimally divisible tasks. The goal is to make the offloading decisions (or choose the offloaded fractions) and to allocate the communication resources so that the energy expended by the users is minimized. The decisions and allocations are made centrally, using the knowledge of the (single-input single-output) channels from all users. For indivisible and infinitesimally divisible computational tasks we will derive closed-form and quasi-closed-form solutions, respectively, to the energy minimization problem when the full capabilities of the multiple access channel are exploited. We will also provide corresponding solutions for some simplified multiple access schemes, namely, time-division multiple access (TDMA), and, in the case of indivisible tasks, sequential decoding without time sharing (SDwts), and independent decoding (ID).

2.2.1 Principles of Proposed Approach

The broad principles that underlie our approach to the resource allocation problem arise from the observation that the communication rates employed by each user must lie in the intersection of two regions. First, the rates must be large enough to meet

the latency constraints imposed by the computational tasks. Second, the rates must lie within the achievable rate region for the chosen multiple access scheme. For the simple two-user case that we will consider in this chapter (see Secs 2.2.2 and 2.3), these regions are illustrated in Fig. 2.1, where the shaded regions are the achievable rate region for (a) independent decoding, (b) TDMA, and (c) sequential decoding without time sharing, and (d) the capacity region of the multiple access channel,[1] for a given channel environment and a given set of operating power constraints. The L-shaped dashed lines in each figure denote the boundary of the set of rate pairs that will enable the latency constraints of the given applications to be satisfied. Rate pairs above and to the right of the boundary will enable the latency constraints to be satisfied, and we will call the set of those rate pairs the latency region. In the scenario marked L, the latency constraints are quite long and in all four cases there is an intersection between the achievable rate region for the channel and the latency region. Therefore, for each multiple access scheme there are rate pairs that are feasible for the computational offloading problem. In the scenario marked S, the latency constraints are quite short and there is no intersection in the cases of independent decoding or TDMA. However, if one takes advantage of the full capabilities of the multiple access channel there is an intersection (see Fig. 2.1(d)), and there are pairs of rates that will enable the latency constraints to be satisfied.

Another useful interpretation of the components in Fig. 2.1 arises from the fact that the operating power levels of the transmitters control the size of the achievable rate regions: as the operating power is reduced, those regions shrink. If we consider the latency region for long latency constraints, marked by L, it appears that there

[1]Rate pairs on the "dominant face" of the capacity region can be achieved by joint decoding or by employing "time sharing" between the "corner points" of the region, each of which can be achieved by sequential decoding (Cover and Thomas, 2012).

is not much "room" to shrink the independent decoding and TDMA regions while retaining an intersection with the latency region. However, there appears to be considerable room to shrink the capacity region. This suggests that by taking advantage of the full capabilities of the multiple access channel we can reduce the energy required to offload the latency-constrained tasks.

2.2.2 Specific Contributions of this Chapter

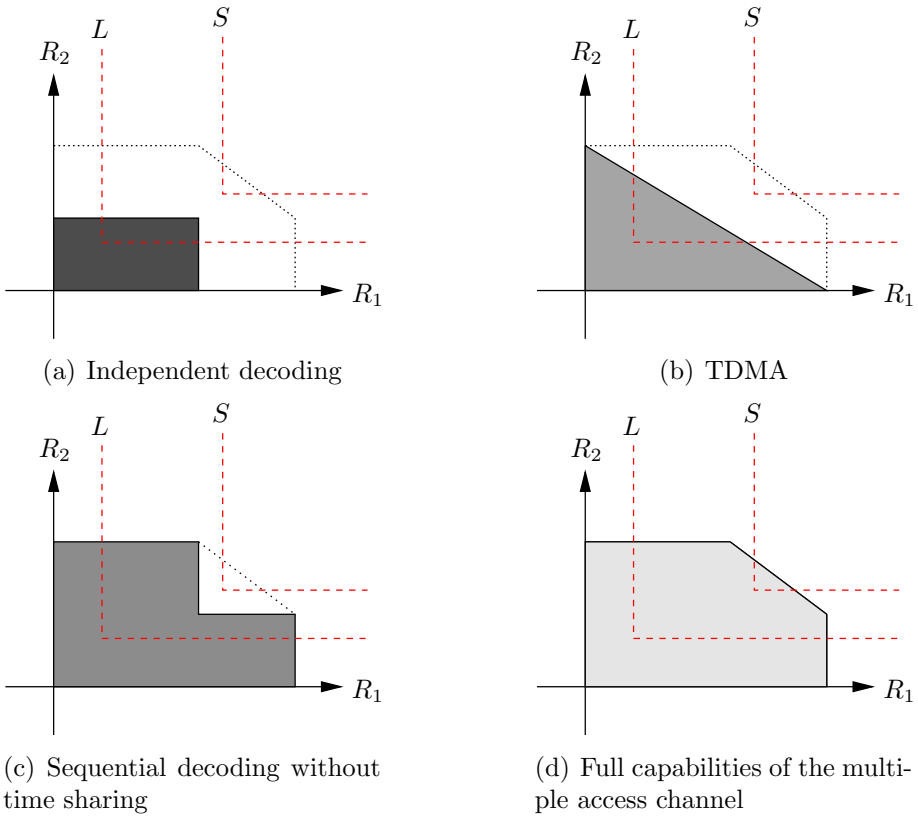
The specific problems considered in this chapter concern a two-user system in which each user has a latency-constrained task that they wish to complete with the possible assistance of a single-antenna access point with plentiful computationally resources. We consider both indivisible computational tasks, for which a binary offloading decision must be made, and infinitesimally divisible tasks, for which the fraction of the task to be offloaded is to be determined. In terms of communication resources, each user has a single antenna and a maximum allowable transmission power. The communication channels are assumed to be known and constant over the latency interval. A distinguishing feature of our system is the observation that when users are seeking to offload (a fraction of) their tasks, the communication system can operate in a combination of three modes: one in which both users are transmitting and the others in which only one user is transmitting. In addition to making the offloading decisions, or determining the fractions to be offloaded, our goal includes determining the durations of the time slots in which the system operates in each mode and the rates and powers allocated to each slot, so that the sum of the computational and communication energies expended by the mobile devices is minimized.

Indivisible Tasks—Binary Computational Offloading

One of our key results in the case of indivisible tasks is a closed-form solution to the mobile energy minimization problem in the case in which the full capabilities of the multiple access channel are exploited when both users are offloading. That solution shows that only two of the three available time slots are required. We then obtain a closed-form expression for the optimal solution in the TDMA case. Quite naturally, that solution can also be achieved in two time slots. In the cases of independent decoding and sequential decoding without time sharing, the optimal solution may have three active time slots, and we provide quasi-closed-form solutions for both those cases. These solutions each depend on the solution of different three-variable optimization problems. Based on the structure of each of those problems we propose a coordinate descent method in which each subproblem is convex. That approach is guaranteed to converge to a stationary point (Hong *et al.*, 2016, Theorem 1), and in all our numerical experiments it converged to the globally optimal solution.

Our closed-form solutions enable us to show that when TDMA is able to satisfy the latency constraints, the optimized solution in the case of independent decoding takes the form of TDMA. (In TDMA systems the decoders operate independently.) We will also show that when the channel gains are the same, if the power budgets of the users are above an explicit threshold, the optimized TDMA solution is globally optimal. A consequence of that result is that independent decoding and sequential decoding (wts) are also optimal when the channel gains are the same.

In a complementary way, our numerical results will illustrate that when the channel gains are quite different and the latency constraints are reasonably tight, taking



(a) Independent decoding

(b) TDMA

(c) Sequential decoding without time sharing

(d) Full capabilities of the multiple access channel

Figure 2.1: The set of feasible rates for the problems that we will consider is the intersection of the achievable rate region (shaded) and the region of the rates that will satisfy the latency constraints, which is above and to the right of the L-shaped dashed lines.

advantage of the full capabilities of the multiple access channel offers significant reductions in the energy required to offload the applications and substantially broadens the range of channel gains over which offloading can be achieved while respecting the power constraints of the devices. Our results will also show that a large fraction of these gains can be achieved by sequential decoding without time sharing.

Infinitesimally Divisible Tasks—Partial Computational Offloading

In Section 2.8 we extend our approach to the limiting case of infinitesimally divisible computational tasks that may be offloaded using the full multiple access scheme or TDMA. Our focus in that section is on data-partitionable applications (Wang *et al.*, 2016), in which a simple-to-describe operation is applied, independently, to different blocks of data. For both the full multiple access and TDMA schemes, we obtain quasi-closed-form solutions that depend on the solution of (different) three-variable optimization problems. For the full multiple access scheme a coordinate descent approach that is guaranteed to yield a stationary point is employed (see Hong *et al.*, 2016, Theorem 1), and in the TDMA case the problem is convex and hence, an optimal solution can be easily obtained. The structure of these subproblems enables us to show that, as in the case of indivisible tasks, when the channel gains are the same, if the transmission power budgets are above a threshold, then the optimized TDMA partial offloading solution is globally optimal. However, when the channel gains are quite different, performing complete offloading using full multiple access scheme can result in significantly lower mobile energy consumption than partial offloading using TDMA.

2.3 System Model

The goal of this chapter is to develop insight into the impact of the choice of the multiple access scheme on the energy consumed by a computational offloading system. To do so, we will consider a two-user system in which each user seeks to obtain the results of a latency-constrained computational task with the possible assistance

of a single access point with plentiful computational resources. The nature of the computational tasks that the users are to execute has a significant impact on the way this problem is formulated (e.g., Muñoz *et al.*, 2015; Wang *et al.*, 2016). If the components of the task are tightly coupled, the problem must be executed either by the user or by the access point alone. That is, the task must be completely offloaded or not offloaded at all (e.g., Kumar and Lu, 2010; Wu *et al.*, 2013; Sardellitti *et al.*, 2015). In contrast, tasks with independent or loosely coupled components can benefit from the parallelism between the mobile device and the access point, with a portion of the task being offloaded and the remainder being computed locally (e.g., Zhang *et al.*, 2013; Wang *et al.*, 2016). In this chapter, we will first consider the case in which both users have an indivisible task; see Sections 2.4–2.7. Then, in Section 2.8, we consider the limiting case in which both users have an infinitesimally divisible task that can be partially offloaded, with the remainder of the task computed locally. For ease of exposition, in this section we will establish the system model for the case of indivisible tasks. The extension of this model to infinitesimally divisible tasks is provided in Section 2.8.

In the case of indivisible tasks, the access point will decide whether or not each user will offload its task. It is assumed that the access point knows the energy that each user would expend in order to complete its task locally before the latency deadline. (If the task cannot be completed locally in time that energy is notionally set to $+\infty$.) That local computational energy is then compared to the energy that would be expended to offload the task to the access point in such a way that the result can be returned to the user before the deadline. That transmission energy is dependent on the allocation of the available communication resources.

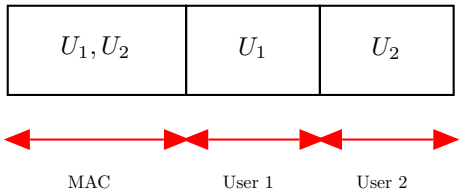


Figure 2.2: Different time slots in computational offloading for two users. In the first time slot both users are offloading simultaneously, while in the second only user one is offloading and in the third only user two is offloading.

In the general case, the users' tasks will have different description lengths and different latency constraints, and the users will offload their descriptions at different rates. As a result, when both users are offloading, one user may complete its transmission before the other. In order to take advantage of that fact in minimizing the users' energy consumption, we will adopt the time slotted structure in Fig. 2.2. In the first time slot, both users are using the channel, and in the second and third time slots, user 1, the user with smaller latency, and user 2 complete the offloading of their applications, respectively. Since the length of any time slot can be set to zero, this time slotted structure naturally incorporates scenarios in which only one user is offloading.

The problem that we will consider is to minimize the total energy consumption of the users. For each user the energy consumption is either the local computation energy or the offloading energy. The variables that we can manipulate are the offloading decisions, the duration of each time slot (or equivalently, the fractions of the number of bits offloaded in each time slot), and each user's transmission power and rate in each slot. The constraints on those variables arise from the latency of each offloaded task, the bounds on the transmission power of each user, and the set of achievable rates of the chosen multiple access scheme. That is, we seek solutions to problems of

the form

$$\min_{\substack{\text{rates, powers,}\\ \text{time slot durations,}\\ \text{offloading decisions}}} \quad \text{Transmission energy + Local computational energy} \tag{2.1a}$$

$$\text{s.t.} \quad \text{Latency constraint of each task,} \tag{2.1b}$$

$$\text{Power constraint on each user,} \tag{2.1c}$$

$$\text{Achievable rate region.} \tag{2.1d}$$

In order to formulate the problem described in (2.1), let B_k denote the total number of bits needed to describe the problem of user k and let γ_{ki} denote the fraction of bits that user k transmits at time slot i. The number of bits transmitted by user k in time slot i can then be written as $\gamma_{ki}B_k$. (We will assume that B_k is large enough that γ_{ki} can be modeled as a continuous variable in $[0, 1]$.) We will let T_s denote the symbol interval of the system, and we let τ_i denote the length (in channel uses) of the i^{th} time slot. Hence, the i^{th} time slot has a duration $\tau_i T_s$. If P_{ki} and R_{ki} denote the transmission power and data rate (in units per channel use) for an offloading user, k, in time slot i, then the offloading energy consumption of user k is $E_{\text{off}_k} = T_s \sum_i P_{ki}\tau_i$. We will let $\bar{P}_k$ denote the maximum operating power level of user k and we impose the power constraint $P_{ki} \leq \bar{P}_k$. The energy consumption of local execution, E_{loc_k}, and the time that would be needed to complete the task locally, t_{loc_k}, are determined by the number of CPU cycles required to execute the computation task and the structure of user k's CPU. In the binary computation offloading case, these parameters are constant and are assumed to be known by the access point.

We will adopt a computational model in which the full description of an offloaded task must be received by the access point before computation can begin, and the

results only become available once the entire task has been completed. Under that model, the latency of an offloaded task is the sum of the time taken to transmit the description of the problem to the access point, t_{UL_k}, the time taken to compute the result at the access point, t_{exe_k}, and the time taken to return the result to the user, t_{DL_k}. Accordingly, the latency constraint for user k can be written as

$$t_{\mathrm{UL}_k} + t_{\mathrm{exe}_k} + t_{\mathrm{DL}_k} \leq L_k. \tag{2.2}$$

As outlined in the Introduction, we will focus on scenarios in which there are sufficient computational resources available at the access point so that the execution time of the offloaded task of the k^{th} user can be considered independent of the computational loads imposed by the other users. Furthermore, we will assume that sufficient communication resources are available on the downlink for the time taken to return the result to user k to be considered as a constant. In other words, we assume that it is the allocation of communication resources on the uplink that is the bottleneck of the offloading problem.

The uplink communication environment that we will consider is a narrowband multiple access channel with two single-antenna users offloading to a single-antenna access point. If the signal transmitted by the k^{th} user at a given time instant is denoted by s_k and the corresponding channel is denoted by h_k, then the signal received at the access point is

$$y = h_1 s_1 + h_2 s_2 + v, \tag{2.3}$$

where v is a sample from a zero mean circular white Gaussian noise process with variance σ^2. We will let $\alpha_k = \frac{|h_k|^2}{\sigma^2}$ denote the effective channel gain of the channel

Table 2.1: Parameters and variables

Symbol	Quantity
α_k	effective channle gain of user k, $\lvert h_k\rvert^2/\sigma^2$
γ_{ki}	the fraction of bits transmitted by user k in slot i
L_k	latency of user k
T_k	execution time plus downlink time of user k
$\tilde{L}_k$	$(L_k - T_k)/T_s$
P_{ki}	power transmitted by user k in slot i
$\bar{P}_k$	power constraint on user k
R_{ki}	rate transmitted by user k in slot i
τ_1	duration of first time slot
T_s	symbol interval
E_{loc_k}	local energy consumption of user k

of user k and we will assume that α_1 and α_2 are known by the access point. To establish the constraints on the rates we will make the assumption that each time slot τ_i is long enough that the achievable rate region for a finite block length can be approximated by the limiting region for long block lengths. (Recent work suggests that conventional rate limits provide insightful guidelines for communication over finite block length even when the blocks are quite short (Polyanskiy *et al.*, 2010; MolavianJazi and Laneman, 2012).)

For ease of later reference, we have summarized the definitions of the key parameters and variables in Table 2.1.

2.4 Binary Computation Offloading

In this section, we consider the case in which both users have indivisible computational tasks. As outlined in the previous section, in this setting, the access point will evaluate

the total mobile energy consumption in each of the four possible cases of local or offloaded computation. The energy of local computation is presumed to be known by the access point (if local computation can meet the latency deadline). In the following section we will find the closed-form optimal solution of the problem in (2.1) when only one user is offloading. The formulation of the "complete" computation offloading problem, in which it is decided that both users should offload their tasks, will be presented in Section 2.4.2 and solutions will be obtained in Sections 2.5 and 2.6.

2.4.1 Single Offloading User

When only one user is offloading, the duration of the first time slot is zero, (i.e., $\tau_1 = 0$), and the solution to the minimal transmission energy problem has a simple closed-form expression (Salmani and Davidson, 2016). If we consider the case in which user 1 is offloading, then, if we simplify our notation so that P_1 and R_1 denote the power and rate employed by user 1, respectively, the problem in (2.1) can be written as

$$\begin{aligned} \min_{P_1,R_1} \quad & T_s\big(\tfrac{B_1}{R_1}\big)P_1 + E_{\mathrm{loc}_2} \\ \text{s.t.} \quad & \tfrac{B_1}{R_1} \leq \tilde{L}_1, && \text{(2.4a)} \\ & 0 \leq P_1 \leq \bar{P}_1, && \text{(2.4b)} \\ & 0 \leq R_1 \leq \log_2(1+\alpha_1 P_1), && \text{(2.4c)} \end{aligned}$$

where $\tilde{L}_k = \frac{L_k - T_k}{T_s}$, and $T_k = t_{\mathrm{exe}_k} + t_{\mathrm{DL}_k}$. In this case, since user 1 is (actively) transmitting to the access point, the transmission rate, R_1, is greater than zero with the

constraint in (2.4a) providing the lower bound on R_1. In this setting, the local execution energy consumption of user 2 is constant, and hence the optimization problem is minimizing the energy consumption of user 1. Considering the constraints in (2.4), it can be shown that when $\frac{B_1}{\bar{L}_1} \geq \log_2(1+\alpha_1 \bar{P}_1)$ the problem is infeasible, which means that user 1 cannot meet its latency constraint by offloading and it should execute its task locally. Otherwise, the optimal communication resource allocation to user 1 is $R_1 = \frac{B_1}{\bar{L}_1}$, and $P_1 = \frac{2^{R_1}-1}{\alpha_1}$.

2.4.2 Both Users Offloading: Complete Computation Offloading

In the case in which both users are offloading, we observe that the durations of the second and third time slots can be written as $\tau_2 = \frac{B_1-\gamma_{11}B_1}{R_{12}}$ and $\tau_3 = \frac{B_2-\gamma_{21}B_2}{R_{23}}$, respectively. Using our ordering of the users (so that $L_1 \leq L_2$), the problem of minimizing the sum of the users' transmission energies required to meet the latency constraints, subject to the power constraints and the achievable rate region, $\mathcal{R}$, of the

chosen multiple access scheme can be formulated as

$$\min_{\substack{P_{11},P_{12},P_{21},P_{23},\\ R_{11},R_{12},R_{21},R_{23},\\ \gamma_{11},\gamma_{21},\tau_1}} \quad T_s\tau_1(P_{11}+P_{21}) + T_s(\tfrac{B_1-\gamma_{11}B_1}{R_{12}})P_{12} + T_s(\tfrac{B_2-\gamma_{21}B_2}{R_{23}})P_{23} \tag{2.5a}$$

$$\text{s.t.} \quad 0 \le \gamma_{11}, \gamma_{21} \le 1, \tag{2.5b}$$

$$\tau_1 + \tfrac{B_1-\gamma_{11}B_1}{R_{12}} \le \tilde{L}_1, \tag{2.5c}$$

$$\tau_1 + \tfrac{B_1-\gamma_{11}B_1}{R_{12}} + \tfrac{B_2-\gamma_{21}B_2}{R_{23}} \le \tilde{L}_2, \tag{2.5d}$$

$$0 \le P_{k1}, P_{k2} \le \bar{P}_k, \quad k = 1, 2, \tag{2.5e}$$

$$0 \le R_{12} \le \log_2(1+\alpha_1 P_{12}), \tag{2.5f}$$

$$0 \le R_{23} \le \log_2(1+\alpha_2 P_{23}), \tag{2.5g}$$

$$\{R_{11}, R_{21}\} \in \mathcal{R}. \tag{2.5h}$$

Analogous to the single-user formulation in (2.4), if user 1 is (actively) transmitting in time slot 2 and user 2 is (actively) transmitting in the slot 3, the corresponding rates will be positive and will be lower bounded by the expressions in (2.5c) and (2.5d). If any of those single-user slots is not used, the corresponding rate, power, slot duration, and number of bits transmitted in that slot is zero.

One of the key results in this chapter is a closed-form expression for the optimal solution to the problem in (2.5) when the full multiple access scheme is employed; i.e., when $\mathcal{R}$ in (2.5h) is the capacity region of the multiple access channel; see Section 2.5. We also provide a closed-form solution in the case of TDMA (Section 2.6.1) and quasi-closed-form solutions for the cases of sequential decoding without time sharing (Section 2.6.2) and independent decoding (Section 2.6.3). In the case of full multiple access scheme we will show (see Appendix 2.B) that only two time slots are required,

and, quite naturally, this is also the case for TDMA. However, for the two other multiple access schemes there are scenarios in which all three time slots are employed.

2.5 Complete Computation Offloading: Full Multiple Access Scheme

The general case of the problem in (2.5) consists of three time slots; cf. Fig. 2.2. In the first time slot, both users are offloading using the full capabilities of the multiple access channel, while in the second time slot user 1 is offloading and in the third slot user 2 is offloading. (Recall that the lengths of the time slots satisfy $\tau_i R_{ki} = \gamma_{ki} B_k$ and that we have ordered the users so that $L_1 \leq L_2$.) The problem can be written explicitly as

$$\min_{\substack{P_{11},P_{12},P_{21},P_{23},\\ R_{11},R_{12},R_{21},R_{23},\tau_1}} \quad T_s\tau_1(P_{11}+P_{21}) + T_s(\tfrac{B_1-\tau_1 R_{11}}{R_{12}})P_{12} + T_s(\tfrac{B_2-\tau_1 R_{21}}{R_{23}})P_{23} \tag{2.6a}$$

$$\text{s.t.} \quad (2.5\text{b}) - (2.5\text{g}), \tag{2.6b}$$

$$0 \leq R_{k1} \leq \log_2(1+\alpha_k P_{k1}), \quad k=1,2, \tag{2.6c}$$

$$R_{11}+R_{21} \leq \log_2(1+\alpha_1 P_{11}+\alpha_2 P_{21}). \tag{2.6d}$$

Although the problem in (2.6) is cast in the generic three time slot setting, we show in Appendix 2.B that it is sufficient to consider a two slot system consisting of a multiple access time slot of length $\tau_1 = \tilde{L}_1$ and a slot in which user 2 transmits alone. In other words, there is an optimal solution to (2.6) in which $\tau_1 = \tilde{L}_1$, $P_{12} = 0$, $R_{12} = 0$, and $\tau_1 R_{11} = B_1$. This result not only simplifies the derivation of an optimal solution to (2.6), it also simplifies the implementation of the system. With this

simplification, the problem in (2.6) reduces to

$$\min_{\substack{P_{11},P_{21},P_{23},\\ R_{21},R_{23}}} \quad T_s\tilde{L}_1(P_{11}+P_{21}) + T_s(\tfrac{B_2-\tilde{L}_1R_{21}}{R_{23}})P_{23} \tag{2.7a}$$

$$\tfrac{B_1}{L_1} \leq \log_2(1+\alpha_1 P_{11}), \tag{2.7b}$$

$$0 \leq R_{21} \leq \log_2(1+\alpha_2 P_{21}), \tag{2.7c}$$

$$\tfrac{B_1}{L_1} + R_{21} \leq \log_2(1+\alpha_1 P_{11}+\alpha_2 P_{21}), \tag{2.7d}$$

$$0 \leq R_{23} \leq \log_2(1+\alpha_2 P_{23}), \tag{2.7e}$$

$$\tilde{L}_1 + \tfrac{B_2-\tilde{L}_1R_{21}}{R_{23}} \leq \tilde{L}_2, \tag{2.7f}$$

$$0 \leq P_{11} \leq \bar{P}_1, \quad 0 \leq P_{21}, P_{23} \leq \bar{P}_2. \tag{2.7g}$$

Our approach to solving the problem in (2.7), and indeed the other problems that we will consider in this chapter, will be to (i) (precisely) decompose the problem into inner and outer problems, (ii) determine a closed-form or quasi-closed-form expression for the optimal solution of the inner problem in terms of the variables of the outer problem, and (iii) solve the outer problem and subsequently obtain optimal values for the inner problem (e.g., Salmani and Davidson, 2016). In the first step, we decompose the problem in (2.7) in such a way that the transmission rate and transmission power of the time slot in which user 2 transmits alone can be obtained in terms of the rates and powers of the first time slot, namely,

$$\begin{array}{ll} \min\limits_{P_{11},P_{21},R_{21}} & \min\limits_{P_{23},R_{23}} \quad (2.7a) \\ \text{s.t.} \quad (2.7b)-(2.7d),(2.7g) & \text{s.t.} \quad (2.7e)-(2.7g). \end{array} \tag{2.8}$$

Since the first part of the objective function in (2.7) is independent of P_{23} and R_{23},

the inner optimization in (2.8) is

$$\min_{P_{23},R_{23}} \quad (\tfrac{B_2-\tilde{L}_1R_{21}}{R_{23}})P_{23} \tag{2.9a}$$

$$0 \leq R_{23} \leq \log_2(1+\alpha_2 P_{23}), \tag{2.9b}$$

$$0 \leq P_{23} \leq \bar{P}_2, \tag{2.9c}$$

$$\tilde{L}_1 + \tfrac{B_2-\tilde{L}_1R_{21}}{R_{23}} \leq \tilde{L}_2. \tag{2.9d}$$

For a given value of R_{23} the objective in (2.9) is increasing in terms of P_{23}. Since the lower bounds on P_{23} are separable, its optimal value is the minimum feasible value; i.e.,

$$P_{23} = \tfrac{2^{R_{23}}-1}{\alpha_2}. \tag{2.10}$$

That enables us to reduce the inner optimization problem to

$$\min_{R_{23}} \quad (\tfrac{B_2-\tilde{L}_1R_{21}}{\alpha_2})\tfrac{2^{R_{23}}-1}{R_{23}} \tag{2.11a}$$

$$0 \leq R_{23} \leq \log_2(1+\alpha_2 \bar{P}_2), \tag{2.11b}$$

$$\tilde{L}_1 + \tfrac{B_2-\tilde{L}_1R_{21}}{R_{23}} \leq \tilde{L}_2. \tag{2.11c}$$

The constraint in (2.11b) is obtained from the constraint in (2.9c) and it guarantees the feasibility of the problem in (2.9) in terms of P_{23}. As shown in Appendix 2.A, the objective function in (2.11) is increasing in terms of R_{23}. Since (2.11c) imposes a lower bound on R_{23} and the right hand side of (2.11b) imposes an upper bound, the optimal value of R_{23} is obtained when equality holds in (2.11c), so long as that value satisfies (2.11b). If it does not, the problem in (2.11) is infeasible and so is that in

(2.6). Therefore, when it is feasible the optimal value of R_{23} is

$$R_{23} = \tfrac{B_2 - \tilde{L}_1 R_{21}}{\tilde{L}_2 - \tilde{L}_1}. \tag{2.12}$$

Having found closed-form solutions for R_{23} and P_{23}, we can begin to solve the outer problem in (2.8), namely,

$$\min_{P_{11},P_{21},R_{21}} \quad T_s\tilde{L}_1(P_{11}+P_{21}) + T_s(\tfrac{\tilde{L}_2-\tilde{L}_1}{\alpha_2})\big(2^{\frac{B_2-\tilde{L}_1R_{21}}{\tilde{L}_2-\tilde{L}_1}} - 1\big) \tag{2.13a}$$

$$\text{s.t.} \quad (2.7\text{b}) - (2.7\text{d}), (2.7\text{g}), \tag{2.13b}$$

$$\tfrac{B_2-\tilde{L}_1R_{21}}{\tilde{L}_2-\tilde{L}_1} \leq \log_2(1+\alpha_2\bar{P}_2). \tag{2.13c}$$

To do so, we decompose (2.13) as

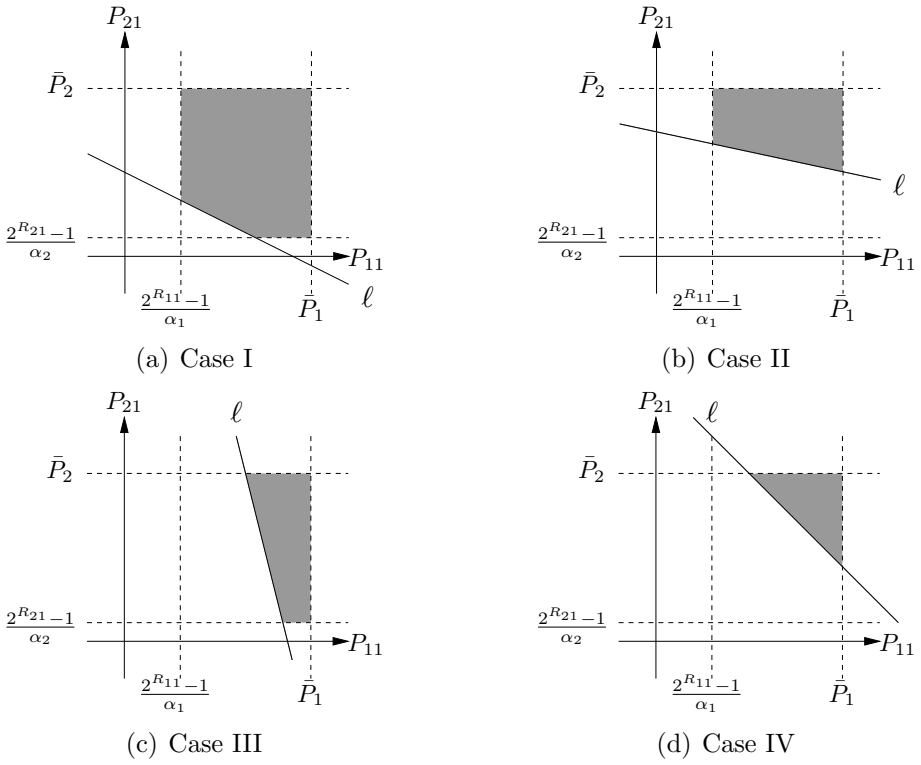
$$\begin{array}{ll} \min\limits_{R_{21}} & \min\limits_{P_{11},P_{21}} \ (2.13\text{a}) \\ \text{s.t.} \ \ (2.13\text{c}), & \text{s.t.} \ \ (2.7\text{b}) - (2.7\text{d}), (2.7\text{g}). \end{array} \tag{2.14}$$

For a given rate R_{21}, the constraints in (2.7b), (2.7c), and (2.7g) construct a rectangular feasibility region of power pair (P_{11}, P_{21}). Since the objective function in (2.13a) is an increasing linear function of P_{11} and P_{21}, it can be shown that at optimality the constraint in (2.7d) holds with equality. Let us define

$$\ell = \{(P_{11},P_{21}) | \alpha_1 P_{11} + \alpha_2 P_{21} + 1 = 2^{B_1/\tilde{L}_1 + R_{21}}\}, \tag{2.15a}$$

$$\ell_k = \{(P_{11},P_{21}) | P_{ki} = (2^{R_{ki}} - 1)/\alpha_k\}. \tag{2.15b}$$

The line ℓ describes the power pairs for which the constraint in (2.7d) is active, and the



(a) Case I (b) Case II

(c) Case III (d) Case IV

Figure 2.3: (P_{11}, P_{21}) feasibility region in different cases.

lines ℓ_k describe the power pairs for which the constraint in (2.7b) and the right hand side of the constraint in (2.7c) are active. Based on the intersection of the line ℓ with the rectangular region constructed by the constraints in (2.7b), (2.7c), and (2.7g), the feasibility regions for the pair (P_{11}, P_{21}) will have different shapes; see Fig. 2.3. If ℓ does not have any intersection with the rectangular region, the problem is not feasible. (Note that the line ℓ cannot lie below the point $(P_{11}, P_{21}) = (\frac{2^{R_{11}}-1}{\alpha_1}, \frac{2^{R_{21}}-1}{\alpha_2})$; see (2.15).)

According to the feasibility regions in Fig. 2.3 and the fact that the optimal values of P_{11} and P_{21} lie on the line ℓ, the inner problem in (2.14) can be written, in terms

of P_{11}, as

$$\min_{P_{11}} \quad T_s\tilde{L}_1\big((1-\tfrac{\alpha_1}{\alpha_2})P_{11}+\tfrac{2^{B_1/\tilde{L}_1+R_{21}-1}}{\alpha_2}\big) \tag{2.16a}$$

$$\text{s.t.} \quad (2.7\text{b})-(2.7\text{c}),(2.7\text{g}). \tag{2.16b}$$

Since the objective in (2.16) is linear in P_{11}, it is sufficient to examine the values of P_{11} on the boundary of its feasible region, which can be written as

$$\max\{\tfrac{2^{B_1/\tilde{L}_1}-1}{\alpha_1},\tfrac{2^{B_1/\tilde{L}_1+R_{21}-1-\alpha_2\bar{P}_2}}{\alpha_1}\}\leq P_{11}\leq \min\{\bar{P}_1,\tfrac{2^{R_{21}}(2^{B_1/\tilde{L}_1}-1)}{\alpha_1}\}, \tag{2.17}$$

and the parameter $\frac{\alpha_1}{\alpha_2}$ defines whether the upper bound or the lower bound on P_{11} is the optimal solution; i.e.,

$$\begin{cases} P_{11}=\max\{\tfrac{2^{B_1/\tilde{L}_1}-1}{\alpha_1},\tfrac{2^{B_1/\tilde{L}_1+R_{21}-1-\alpha_2\bar{P}_2}}{\alpha_1}\}, & \text{if } \tfrac{\alpha_1}{\alpha_2}\leq 1, \quad (2.18\text{a})\\ P_{11}=\min\{\bar{P}_1,\tfrac{2^{R_{21}}(2^{B_1/\tilde{L}_1}-1)}{\alpha_1}\}, & \text{if } \tfrac{\alpha_1}{\alpha_2}>1. \quad (2.18\text{b}) \end{cases}$$

For any value of $\frac{\alpha_1}{\alpha_2}$, we have two candidates for the optimal solution for P_{11} (and, consequently, for P_{21}). By substituting the closed-form expressions for the transmission powers and solving the problem in (2.14) the candidate for the minimum objective function in (2.14) can be obtained.

2.5.1 Case A: $\frac{\alpha_1}{\alpha_2} \leq 1$

Using (2.18a) and (2.15a), the candidate solutions of the winner problem when the channel gain of the second user is larger than that of the first user are

$$\begin{cases} P_{11} = \frac{2^{B_1/\tilde{L}_1}-1}{\alpha_1},\ P_{21} = \frac{2^{B_1/\tilde{L}_1}(2^{R_{21}}-1)}{\alpha_2}, & \text{(2.19a)} \\ P_{11} = \frac{2^{B_1/\tilde{L}_1+R_{21}}-\alpha_2\bar{P}_2-1}{\alpha_1},\ P_{21} = \bar{P}_2. & \text{(2.19b)} \end{cases}$$

Case A-I

By substituting the values in (2.19a) the outer optimization problem in (2.14) can be written as

$$\min_{R_{21}} \quad T_s\tilde{L}_1\big(\tfrac{2^{B_1/\tilde{L}_1}-1}{\alpha_1} + \tfrac{2^{B_1/\tilde{L}_1}(2^{R_{21}}-1)}{\alpha_2}\big) + T_s\big(\tfrac{\tilde{L}_2-\tilde{L}_1}{\alpha_2}\big)\big(2^{\frac{B_2-\tilde{L}_1R_{21}}{\tilde{L}_2-\tilde{L}_1}} - 1\big) \quad \text{(2.20a)}$$

$$\text{s.t.} \quad \tfrac{B_2-(\tilde{L}_2-\tilde{L}_1)\log_2(1+\alpha_2\bar{P}_2)}{\tilde{L}_1} \leq R_{21} \leq \log_2(1+\alpha_2\bar{P}_2). \quad \text{(2.20b)}$$

In Appendix 2.C we show that the objective function in (2.20) is a convex function of R_{21}. Since the constraints are linear in terms of R_{21}, the optimization problem in (2.20) is convex and there are three possible values for the optimal value of R_{21}. It it is feasible, $\frac{B_2-(\tilde{L}_2-\tilde{L}_1)\log_2(1+\alpha_2\bar{P}_2)}{\tilde{L}_1} \leq \log_2(1+\alpha_2\bar{P}_2)$, the value for which the first derivative of the objective function is equal to zero, $R_{21_{\text{d}}} = \frac{B_2}{\tilde{L}_2} - \frac{(\tilde{L}_2-\tilde{L}_1)B_1}{\tilde{L}_1\tilde{L}_2}$ is the optimal solution. Otherwise, either the upper bound or the lower bound on R_{21} is the optimal solution. (If $\frac{B_2-(\tilde{L}_2-\tilde{L}_1)\log_2(1+\alpha_2\bar{P}_2)}{\tilde{L}_1} > \log_2(1+\alpha_2\bar{P}_2)$, the problem is infeasible.)

Case A-II

Using the transmission powers in (2.19b), the rate optimization problem can be written as

$$\min_{R_{21}} \quad T_s\tilde{L}_1\big(\tfrac{2^{B_1/\tilde{L}_1+R_{21}}-\alpha_2\bar{P}_2-1}{\alpha_1}+\bar{P}_2\big)+T_s\big(\tfrac{\tilde{L}_2-\tilde{L}_1}{\alpha_2}\big)\big(2^{\frac{B_2-\tilde{L}_1R_{21}}{\tilde{L}_2-\tilde{L}_1}}-1\big) \tag{2.21a}$$

$$\text{s.t.} \quad \text{(2.20b)}. \tag{2.21b}$$

The objective function in (2.21) is convex in terms of R_{21} and the constraints are linear. Hence, the optimal value for R_{21} can be at either end point of its feasibility interval or at the point at which the derivative of the objective is equal to zero.

2.5.2 Case B: $\frac{\alpha_1}{\alpha_2} > 1$

Using (2.18b) and (2.15a), the candidate solutions of the inner problem when the channel gain of the first user is greater than that of the second user are

$$\begin{cases} P_{11}=\frac{2^{R_{21}}(2^{B_1/\tilde{L}_1}-1)}{\alpha_1},\ P_{21}=\frac{2^{R_{21}}-1}{\alpha_2}, & \text{(2.22a)} \\ P_{11}=\bar{P}_1,\ P_{21}=\frac{2^{B_1/\tilde{L}_1+R_{21}}-\alpha_1\bar{P}_1-1}{\alpha_2}. & \text{(2.22b)} \end{cases}$$

Case B-I

By substituting the values in (2.22a), the outer problem in (2.14) becomes

$$\min_{R_{21}} \quad T_s\tilde{L}_1\big(\tfrac{2^{B_1/\tilde{L}_1+R_{21}}-2^{R_{21}}}{\alpha_1}+\tfrac{2^{R_{21}}-1}{\alpha_2}\big)+T_s\big(\tfrac{\tilde{L}_2-\tilde{L}_1}{\alpha_2}\big)\big(2^{\frac{B_2-\tilde{L}_1R_{21}}{\tilde{L}_2-\tilde{L}_1}}-1\big) \tag{2.23a}$$

$$\text{s.t.} \quad \text{(2.20b)}. \tag{2.23b}$$

Applying the techniques used in Appendix 2.C, it can be shown that the objective function in (2.23) is a convex function of R_{21}. Since the constraints in (2.23) are the same as those in (2.20), there are three possible values for the optimal value of R_{21}. If it is feasible, the value for which the derivative of the objective is zero, $R_{21_{\mathrm{d}}} = \frac{B_2}{\tilde{L}_2} - \frac{(\tilde{L}_2-\tilde{L}_1)\phi_1}{\tilde{L}_2}$, where $\phi_1 = \log_2\big(\frac{\alpha_2}{\alpha_1}(2^{B_1/\tilde{L}_1}-1)+1\big)$, is the optimal solution. Otherwise, either the upper bound or the lower bound on R_{21} is the optimal solution, or the problem is infeasible.

Case B-II

By substituting the values in (2.22b), the rate optimization problem for this candidate solution is

$$\min_{R_{21}} \quad T_s\tilde{L}_1\big(\tfrac{2^{B_1/\tilde{L}_1+R_{21}-\alpha_1\bar{P}_1-1}}{\alpha_2}+\bar{P}_1\big)+T_s\big(\tfrac{\tilde{L}_2-\tilde{L}_1}{\alpha_2}\big)\big(2^{\frac{B_2-\tilde{L}_1R_{21}}{\tilde{L}_2-\tilde{L}_1}}-1\big) \tag{2.24a}$$

$$\text{s.t.} \quad \text{(2.20b)}. \tag{2.24b}$$

The objective function in (2.24) has the same structure as the objective function in (2.21). Accordingly, the optimal solution for the transmission rate R_{21} is either at one of the end points of the feasible interval, or at the point in which the derivative of the objective function is equal to zero if that point lies within the feasibility region.

2.5.3 Algorithm for Solving (2.6)

By considering the derivations in the subsections above, we obtain a closed-form expression for an optimal solution to (2.6). The steps that need to be taken to obtain this solution are summarized in Algorithm 1. For a complete solution to the offloading

problem for indivisible tasks, the energy obtained from Algorithm 1 would need to be compared to the total mobile energy consumption for the scenarios in which only one user offloads (assuming local computation is feasible for the other user), and for the case where both users compute locally. The optimal offloading decision corresponds to the scenario with the lowest mobile energy consumption.

Algorithm 1 : An optimal solution to (2.6)

Input data: values of $\{B_k\}$, $\{\bar{P}_k\}$, $\{L_k\}$, $\{T_k\}$, $\{\alpha_k\}$, and T_s.
if $\frac{B_2-(\tilde{L}_2-\tilde{L}_1)\log_2(1+\alpha_2\bar{P}_2)}{\tilde{L}_1} > \log_2(1+\alpha_2\bar{P}_2)$ **then**
The problem is infeasible; c.f. (2.20).
else if $\frac{\alpha_1}{\alpha_2} \leq 1$ **then**
Generate a partial candidate solution according to Case A-I (Section 2.5.1). That is, let $R_{21}^{\star}$ denote the optimal solution to (2.20), and calculate $P_{11}^{\star}$ and $P_{21}^{\star}$ using (2.19a).
Generate a partial candidate solution according to Case A-II (Section 2.5.1). That is, let $R_{21}^{\star}$ denote the optimal solution to (2.21), and calculate $P_{11}^{\star}$ and $P_{21}^{\star}$ using (2.19b).
else
Generate a partial candidate solution according to Case B-I (Section 2.5.2). That is, let $R_{21}^{\star}$ denote the optimal solution to (2.23), and calculate $P_{11}^{\star}$ and $P_{21}^{\star}$ using (2.22a).
Generate a partial candidate solution according to Case B-II (Section 2.5.2). That is, let $R_{21}^{\star}$ denote the optimal solution to (2.24), and calculate $P_{11}^{\star}$ and $P_{21}^{\star}$ using (2.22b).
end if
Complete each partial candidate solution by choosing $R_{23}^{\star}$ according to (2.12), $P_{23}^{\star}$ according to (2.10), and setting $\tau_1 = \tilde{L}_1$, $P_{12} = 0$, $R_{12} = 0$ and $R_{11} = \frac{B_1}{\tau_1}$. Calculate the objective value for each candidate solution, and choose the solution that corresponds to the minimum value.

2.6 Complete Computation Offloading: Suboptimal Multiple Access Methods

Having obtained a closed-form expression for an optimal solution to the minimum energy offloading problem when the full capabilities of the multiple access channel are employed, we now address that problem when suboptimal multiple access methods are employed. That is, we will solve the variant of the problem in (2.5) in which the achievable rate region constraint of the first time slot, cf. (2.5h), is the rate region of the chosen suboptimal scheme rather than the capacity region. We will consider the cases of TDMA, sequential decoding without time sharing (SDwts), and independent decoding (ID).

2.6.1 Time Division Multiple Access

In the TDMA scheme only one user is transmitting at a time and the communication resource is fully assigned to that user. Quite naturally, that means that the optimal energy consumption can be achieved using only two time slots. By simplifying our notation in an intuitive way and observing that the durations of the time slots are

$\frac{B_1}{R_1}$ and $\frac{B_2}{R_2}$, respectively, the optimization problem in the TDMA case becomes

$$\min_{\substack{P_1,P_2,\\ R_1,R_2}} \quad T_s\big(\tfrac{B_1P_1}{R_1}+\tfrac{B_2P_2}{R_2}\big) \tag{2.25a}$$

$$\text{s.t.} \quad 0 \le R_k \le \log_2(1+\alpha_k P_k), \quad k=1,2, \tag{2.25b}$$

$$0 \le P_k \le \bar{P}_k, \quad k=1,2, \tag{2.25c}$$

$$\tfrac{B_1}{R_1} \le \tilde{L}_1, \tag{2.25d}$$

$$\tfrac{B_1}{R_1}+\tfrac{B_2}{R_2} \le \tilde{L}_2. \tag{2.25e}$$

The optimal powers in (2.25) are those that achieve equality in the upper bounds in (2.25b), and hence the problem can be simplified to

$$\min_{R_1,R_2} \quad T_s\Big(\tfrac{B_1}{R_1}\big(\tfrac{2^{R_1}-1}{\alpha_1}\big)+\tfrac{B_2}{R_2}\big(\tfrac{2^{R_2}-1}{\alpha_2}\big)\Big) \tag{2.26a}$$

$$\text{s.t.} \quad 0 \le R_1 \le \log_2(1+\alpha_1\bar{P}_1), \tag{2.26b}$$

$$0 \le R_2 \le \log_2(1+\alpha_2\bar{P}_2), \tag{2.26c}$$

$$\text{(2.25d) and (2.25e).} \tag{2.26d}$$

The problem in (2.26) can be decomposed as

$$\begin{array}{ll} \min\limits_{R_1} & \min\limits_{R_2} \quad \text{(2.26a)} \\ \text{s.t.} \quad \text{(2.25d)}, \text{(2.26b)}, & \text{s.t.} \quad \text{(2.25e)}, \text{(2.26c)}. \end{array} \tag{2.27}$$

The objective function of the problem in (2.27) is increasing in terms of R_2 and the optimal transmission rate for the second user is achieved when (2.25e) holds with

equality, i.e., $R_2 = \frac{B_2 R_1}{\tilde{L}_2 R_1 - B_1}$. The outer optimization problem is then

$$\min_{R_1} \quad T_s\Big(\frac{B_1}{R_1}\big(\frac{2^{R_1}-1}{\alpha_1}\big) + \frac{\tilde{L}_2 R_1 - B_1}{\alpha_2 R_1}\big(2^{\frac{B_2 R_1}{\tilde{L}_2 R_1 - B_1}} - 1\big)\Big) \tag{2.28a}$$

$$\text{s.t.} \quad \text{(2.25d) and (2.26b)}, \tag{2.28b}$$

$$\frac{B_2 R_1}{\tilde{L}_2 R_1 - B_1} \leq \log_2(1 + \alpha_2 \bar{P}_2). \tag{2.28c}$$

It is shown by Salmani and Davidson (2017a) that the objective in (2.28) is convex. Since the constraints are linear, the optimal solution is either where the derivative of the objective is zero or at one of the end points of the feasibility interval. For each of those values for R_1 the corresponding values for R_2, P_1 and P_2 can be obtained and the quadruple that provides the smallest objective value in (2.25) is the optimal solution.

2.6.2 Sequential Decoding without time sharing

In the sequential decoding without time sharing scheme, the received signal from one user is decoded considering the interference from the other user as noise. Presuming that this message is correctly decoded, the interference of the decoded user is then reconstructed and subtracted from the received signal and the other user is decoded without interference. We will look at the case in which the system can choose the order in which the users are decoded, but that order remains fixed for the duration of the multiple access interval. Hence, we use the qualifier "without time sharing" (wts) in our description. The achievable rate region of sequential decoding (wts) is shown in Fig. 2.1(c). For this scheme there are scenarios in which the optimal energy consumption requires that all three time slots of the system be employed. Hence, the

problem of finding the minimum energy consumption for this scheme is

$$\min_{\substack{P_{11},P_{12},P_{21},P_{23},\\ R_{11},R_{12},R_{21},R_{23},\tau_1}} \quad T_s\tau_1(P_{11}+P_{21})+T_s(\tfrac{B_1-\tau_1R_{11}}{R_{12}})P_{12}+T_s(\tfrac{B_2-\tau_1R_{21}}{R_{23}})P_{23} \tag{2.29a}$$

$$\text{s.t.} \quad (2.5\text{b})-(2.5\text{g}), \tag{2.29b}$$

$$\{R_{11},R_{21}\}\in\tilde{\mathcal{R}}_1\cup\tilde{\mathcal{R}}_2, \tag{2.29c}$$

where

$$\tilde{\mathcal{R}}_1=\big\{\{R_{11},R_{21}\}|0\le R_{11}\le\log_2(1+\alpha_1P_{11}),\quad 0\le R_{21}\le\log_2(1+\tfrac{\alpha_2P_{21}}{1+\alpha_1P_{11}})\big\},$$
$$\tilde{\mathcal{R}}_2=\big\{\{R_{11},R_{21}\}|0\le R_{11}\le\log_2(1+\tfrac{\alpha_1P_{11}}{1+\alpha_2P_{21}}),\quad 0\le R_{21}\le\log_2(1+\alpha_2P_{21})\big\}.$$

The problem in (2.29) can be decomposed and a closed-form solution for the transmission rates and the transmission powers of the second and third time slots can be obtained in terms of the transmission rates and transmission powers of the first time slot (Salmani and Davidson, 2017a). The remaining optimization problem is

$$\min_{\substack{P_{11},P_{21},\\ R_{11},R_{21},\tau_1}} \quad T_s\tau_1(P_{11}+P_{21})+T_s(\tfrac{\tilde{L}_1-\tau_1}{\alpha_1})(2^{\frac{B_1-\tau_1R_{11}}{\tilde{L}_1-\tau_1}}-1)+T_s(\tfrac{\tilde{L}_2-\tilde{L}_1}{\alpha_2})(2^{\frac{B_2-\tau_1R_{21}}{\tilde{L}_2-\tilde{L}_1}}-1)$$

$$\text{s.t.} \quad (2.5\text{b}),(2.5\text{e}) \tag{2.31a}$$

$$\tfrac{B_1-\tau_1R_{11}}{\tilde{L}_1-\tau_1}\le\log_2(1+\alpha_1\bar{P}_1), \tag{2.31b}$$

$$\tfrac{B_2-\tau_1R_{21}}{\tilde{L}_2-\tilde{L}_1}\le\log_2(1+\alpha_2\bar{P}_2), \tag{2.31c}$$

$$\{R_{11},R_{21}\}\in\tilde{\mathcal{R}}_1\cup\tilde{\mathcal{R}}_2. \tag{2.31d}$$

Depending on the corner point at which sequential decoding scheme is operating, a closed-form solution for the powers can be obtained in terms of the rates (Salmani

and Davidson, 2017a), namely,

$$\begin{cases} P_{11} = \frac{2^{R_{11}}-1}{\alpha_1}, \quad P_{21} = \frac{2^{R_{11}}(2^{R_{21}}-1)}{\alpha_2} & \text{for } \tilde{\mathcal{R}}_1 \\ P_{11} = \frac{2^{R_{21}}(2^{R_{11}}-1)}{\alpha_1}, \quad P_{21} = \frac{2^{R_{21}}-1}{\alpha_2} & \text{for } \tilde{\mathcal{R}}_2 \end{cases}$$

Since the rate regions for each decoding order, $\tilde{\mathcal{R}}_1$ and $\tilde{\mathcal{R}}_2$, are rectangular (see Fig. 2.1(c)), the optimal rate pair lies at the "dominant" corner (i.e., the "North-East" corner) of one of the rectangles. To determine which rectangle, and hence the optimal triple (R_{11}, R_{21}, τ_1) in (2.31), we will first determine the optimal triple for each decoding order and then select the triple that generates the lower energy solution. The optimization problem for (R_{11}, R_{21}, τ_1) has a similar structure in each case, and in the case of $\tilde{\mathcal{R}}_1$ it is

$$\begin{aligned} \min_{R_{11},R_{21},\tau_1} \quad & T_s\tau_1\big(\tfrac{2^{R_{11}}-1}{\alpha_1} + \tfrac{2^{R_{11}}(2^{R_{21}}-1)}{\alpha_2}\big) + T_s(\tfrac{\tilde{L}_1-\tau_1}{\alpha_1})(2^{\frac{B_1-\tau_1 R_{11}}{\tilde{L}_1-\tau_1}} - 1) \\ & \quad + T_s(\tfrac{\tilde{L}_2-\tilde{L}_1}{\alpha_2})(2^{\frac{B_2-\tau_1 R_{21}}{\tilde{L}_2-\tilde{L}_1}} - 1) && \text{(2.32a)} \\ \text{s.t.} \quad & \text{(2.5b)}, && \text{(2.32b)} \\ & \tfrac{2^{R_{11}}-1}{\alpha_1} \le \bar{P}_1, && \text{(2.32c)} \\ & \tfrac{2^{R_{11}}(2^{R_{21}}-1)}{\alpha_2} \le \bar{P}_2, && \text{(2.32d)} \\ & \text{(2.31b) and (2.31c)}. && \text{(2.32e)} \end{aligned}$$

This problem is not known to be convex, but a variety of solution strategies can be developed, including algorithms based on augmented Lagrangian techniques (Nocedal and Wright, 2006, Chapter 17), sequential quadratic programming (Nocedal and Wright, 2006, Chapter 18), (Lawrence and Tits, 2001), and successive convex

approximation (Razaviyayn *et al.*, 2013; Scutari *et al.*, 2017). In this chapter we will employ a simpler strategy that is based on the observation that when any two of the variables in (2.32) are fixed, the objective function is convex in the remaining variable and the constraints can be written so that they are linear in that variable. That observation suggests the adoption of a coordinate descent method for solving (2.32). The convexity of (2.32) in each coordinate (alone), and other properties of the objective and the constraints, enable us to show that the coordinate descent method is guaranteed to converge to a stationary solution of (2.32) (Hong *et al.*, 2016, Theorem 1). In all our numerical experiments, only some of which are shown in Section 2.9, our coordinate descent method converged to the globally optimal solution.

2.6.3 Independent Decoding

In the independent decoding scheme, the received signals of both users are decoded independently, with the interference of the other user being considered as noise. The achievable rate region in this scheme is depicted in Fig. 2.1(a). As in the case of sequential decoding (wts), the optimal solution may activate all three time slots. Therefore, the minimum energy consumption optimization problem in the independent decoding case can be written as

$$\min_{\substack{P_{11},P_{12},P_{21},P_{23},\\ R_{11},R_{12},R_{21},R_{23},\tau_1}} \quad T_s\tau_1(P_{11}+P_{21}) + T_s(\tfrac{B_1-\tau_1 R_{11}}{R_{12}})P_{12} + T_s(\tfrac{B_2-\tau_1 R_{21}}{R_{23}})P_{23} \tag{2.33a}$$

$$\text{s.t.} \quad (2.5\text{b}) - (2.5\text{g}), \tag{2.33b}$$

$$0 \le R_{11} \le \log_2(1+\tfrac{\alpha_1 P_{11}}{1+\alpha_2 P_{21}}), \tag{2.33c}$$

$$0 \le R_{21} \le \log_2(1+\tfrac{\alpha_2 P_{21}}{1+\alpha_1 P_{11}}), \tag{2.33d}$$

where (2.33c) and (2.33d) describe the achievable rate region in the first time slot. In terms of the variables of the second and third time slots, the problem in (2.33) is similar to that in (2.29) for the sequential decoding (wts) case. Accordingly, the optimal transmission rates and transmission powers of these slots can be obtained in terms of the rates and powers of the first time slot (cf. Section 2.6.2), and the problem can be simplified to

$$\min_{\substack{P_{11},P_{21},\\ R_{11},R_{21},\tau_1}} \quad T_s\tau_1(P_{11}+P_{21}) + T_s(\tfrac{\tilde{L}_1-\tau_1}{\alpha_1})(2^{\frac{B_1-\tau_1 R_{11}}{\tilde{L}_1-\tau_1}}-1) + T_s(\tfrac{\tilde{L}_2-\tilde{L}_1}{\alpha_2})(2^{\frac{B_2-\tau_1 R_{21}}{\tilde{L}_2-\tilde{L}_1}}-1) \tag{2.34a}$$

$$\text{s.t.} \quad (2.5b), (2.5e), (2.31b), (2.31c), (2.33c), (2.33d). \tag{2.34b}$$

By decomposing the problem in (2.34), the powers of the first time slot can be obtained in terms of the rates of that slot (Salmani and Davidson, 2017a),

$$P_{11} = \tfrac{2^{R_{21}}(2^{R_{11}}-1)}{\alpha_1(2^{R_{11}}+2^{R_{21}}-2^{R_{11}+R_{21}})}, \tag{2.35}$$

$$P_{21} = \tfrac{2^{R_{11}}(2^{R_{21}}-1)}{\alpha_2(2^{R_{11}}+2^{R_{21}}-2^{R_{11}+R_{21}})}. \tag{2.36}$$

Accordingly, the remaining optimization problem reduces to

$$\min_{R_{11},R_{21},\tau_1} \quad f(R_{11},R_{21},\tau_1) \tag{2.37a}$$

$$\text{s.t.} \quad (2.5b), (2.31b), (2.31c), \tag{2.37b}$$

$$\tfrac{2^{R_{21}}(2^{R_{11}}-1)}{\alpha_1(2^{R_{11}}+2^{R_{21}}-2^{R_{11}+R_{21}})} \leq \bar{P}_1, \tag{2.37c}$$

$$\tfrac{2^{R_{11}}(2^{R_{21}}-1)}{\alpha_2(2^{R_{11}}+2^{R_{21}}-2^{R_{11}+R_{21}})} \leq \bar{P}_2, \tag{2.37d}$$

$$f(R_{11},R_{21},\tau_1) = T_s\tau_1\Big(\frac{2^{R_{21}}(2^{R_{11}}-1)}{\alpha_1(2^{R_{11}}+2^{R_{21}}-2^{R_{11}+R_{21}})}+\frac{2^{R_{11}}(2^{R_{21}}-1)}{\alpha_2(2^{R_{11}}+2^{R_{21}}-2^{R_{11}+R_{21}})}\Big) + T_s(\frac{\tilde{L}_1-\tau_1}{\alpha_1})(2^{\frac{B_1-\tau_1R_{11}}{\tilde{L}_1-\tau_1}}-1)+T_s(\frac{\tilde{L}_2-\tilde{L}_1}{\alpha_2})(2^{\frac{B_2-\tau_1R_{21}}{\tilde{L}_2-\tilde{L}_1}}-1). \quad (2.38)$$

where $f(R_{11},R_{21},\tau_1)$ is shown in (2.38) at the top of this page.

Similar to the problem in (2.32), we can show that the objective function in (2.37) is convex in each of the variables when the other two are given, and the constraints can be written so that they are linear in the corresponding variable. Hence, we will adopt a coordinate descent approach to solving (2.37). It can be shown that that approach is guaranteed to converge to a stationary solution, and in all our numerical experiments it converged to the globally optimal solution.

2.7 On the Choice of the Multiple Access Scheme for Complete Computation Offloading

The closed-form expressions that we have obtained for the optimal communication resource allocation in the case of complete computation offloading with the full multiple access scheme and TDMA, and the quasi-closed-form expressions that we have obtained for the cases of independent decoding and sequential decoding (without time sharing), enable us to gain insight into the impact of the choice of the multiple access scheme.

The first result is that whenever TDMA produces a solution that is feasible for the offloading problem, that solution is also an optimal solution for the independent decoding case. (This is consistent with the fact that in the TDMA case the decoders

work independently.) As one might expect, there are scenarios in which a three-time-slot independent decoding scheme provides a feasible solution to the offloading problem, but TDMA does not, and we will provide some examples of such scenarios in Section 2.9.1. However, whenever TDMA is feasible it is optimal for the independent decoding case, and in certain circumstances it will have a more straightforward implementation. We formally state this property in the following proposition.

Proposition 1. *If the problem in* (2.28) *is feasible, let* $P_k^\star$ *and* $R_k^\star$ *denote the solution to the problem in* (2.25) *that is derived in Section 2.6.1. In that case, an optimal solution to the problem in* (2.33) *is* $P_{11} = P_{21} = 0, P_{12} = P_1^\star, P_{23} = P_2^\star, R_{11} = R_{21} = 0, R_{12} = R_1^\star, R_{23} = R_2^\star$ *and* $\tau_1 = 0$.

Proof. See Appendix 2.D. □

Our second result shows that when the channel gains of both users are equal and the power budgets are above an explicit threshold then the optimal resource allocation for the TDMA scheme reduces the energy consumption to the same level as the optimal resource allocation for the full multiple access scheme. In other words, when the channel gains are equal, simplifying the implementation by constraining the multiple access scheme to be TDMA does not result in loss of optimality (for sufficiently large power budgets). Having said that, as we will show in Section 2.9, when the channel gains are significantly different, exploiting the full capabilities of the multiple access channel enables substantial reduction in the energy required to offload the tasks. The optimality of the TDMA scheme is formalized in the following proposition.

Proposition 2. *Let* $P_{ki}^\star$ *denote an optimal solution for* P_{ki} *in the problem in* (2.6)

when $|h_1|^2 = |h_2|^2$. *If* $P_{11}^\star + P_{21}^\star \leq \min\{\bar{P}_1, \bar{P}_2\}$, *then TDMA can obtain the optimal energy consumption of the full multiple access scheme.*

Proof. See Appendix 2.E. □

Since in a time-slotted system the optimal TDMA scheme is an optimal independent decoding scheme whenever it is feasible (Proposition 1), a consequence of Proposition 2 is that, for power budgets above the threshold, independent decoding is also optimal when the channel gains are the same. Since the achievable rate region of the sequential decoding (wts) scheme is no smaller than that of independent decoding, Proposition 2 also implies that sequential decoding (wts) is optimal when the channel gains are equal.

2.8 Partial Computation Offloading

Up until this point, we have considered indivisible computational tasks that are either completely offloaded or executed locally. For divisible computational tasks, we have the opportunity to take advantage of the implicit parallelism of the mobile station and the access point by offloading a portion of the computational task to the access point, with the remainder being executed locally.

Our first observation in the development of resource allocation algorithms for the partial offloading case is that the transmission energy and the communication latency associated with offloading a portion of the task depend on its description length, whereas the computational energy and latency associated with executing the remaining portion locally depend on the number of operations required. In this section we will focus on the class of "data-partitionable" tasks (Wang *et al.*, 2016).

Such tasks involve a relatively simple-to-describe action being applied, independently, to multiple blocks of data. As such, the number of operations required to complete a fraction of the task can be modeled as being a function of the description length (Wang *et al.*, 2016; Muñoz *et al.*, 2015; Zhang *et al.*, 2013). For simplicity we will consider the limiting case in which the tasks can be partitioned finely enough that the partition can be modeled by a continuous variable.

In the generic scenario of partial offloading, both users will be offloading a portion of their tasks and the time slotted communication model in Fig. 2.2 applies. (Note that without loss of generality we have ordered the users so that $L_1 \leq L_2$.) Based on the outcomes of the indivisible case, we will focus on the full multiple access and TDMA schemes, and hence, we need only consider two of the time slots ($\tau_2 = 0$). As in our earlier model, γ_{ki} denotes the fraction of its task description that user k offloads in time slot i, but in the partial offloading case ($\gamma_{11}+\gamma_{12}$) and ($\gamma_{21}+\gamma_{23}$) lie in the interval $[0, 1]$ rather than at one of the end points. The total energy consumption of a user in the partial offloading case is the summation of the energy consumed in transmitting the offloaded portion to the access point, E_{off_k}, and the energy consumed in the local execution of the remaining fraction, E_{loc_k}. Moreover, the latency constraint of each user must be applied to both the execution time of the local component, t_{loc_k}, and the total time that it takes to transmit the offloaded component, execute it at the access point, and send the results back to the user.

For a given choice of offloading fractions, E_{off_k} takes the same form as in the indivisible case, and the time taken to upload, compute, and return the results of the offloaded portion has the same three components as in (2.2). The uploading time takes a familiar form (cf. (2.5a)), $t_{\text{UL}_k} = T_s \sum_i \frac{\gamma_{ki} B_k}{R_{ki}}$, and we will assume that the

time taken to return the results to the user, t_{DL_k}, is independent of the fraction of the task that is offloaded. For the class of problems that we are considering, the number of operations to be performed depends on the description length, and hence the execution time at the access point (which has plentiful computational resources) can be modeled as $t_{\mathrm{exe}_k} = \delta_c \sum_i \gamma_{ki} B_k$, where δ_c denotes the constant processing time of one bit.

The energy consumed in computing a portion of the task locally, and the time incurred in doing so, are dependent on the computational architecture at the user. Hence, in our initial formulation we will represent them generically using a function $\mathcal{F}_k(\cdot)$ of the number of bits in the retained description, and t_{loc_k}, respectively. In this thesis we will solve the optimal offloading problems for the dynamic voltage scaling architecture (Zhang *et al.*, 2013), which provides energy-optimal local computation; see Section 2.8.1. A solution to a related problem for a local computational model that resembles the one we have used for the access point was provided by Salmani and Davidson (2017c).

With the above computation and communication models in place, the problem of minimizing the total energy consumption of a system with partial offloading can be

formulated as

$$\min_{\substack{P_{11},P_{21},P_{23},\\ R_{11},R_{21},R_{23},\\ \gamma_{11},\gamma_{21},\gamma_{23}}} \quad T_s\Big((\tfrac{\gamma_{11}B_1}{R_{11}})P_{11}+(\tfrac{\gamma_{21}B_2}{R_{21}})P_{21}+(\tfrac{\gamma_{23}B_2}{R_{23}})P_{23}\Big)+\mathcal{F}_1\big((1-\gamma_{11})B_1\big)$$

$$+\mathcal{F}_2\big((1-\gamma_{21}-\gamma_{23})B_2\big) \tag{2.39a}$$

$$\text{s.t.} \quad 0\le\gamma_{11}\le 1, \tag{2.39b}$$

$$0\le\gamma_{21}\le 1, \quad 0\le\gamma_{23}\le 1, \tag{2.39c}$$

$$0\le\gamma_{21}+\gamma_{23}\le 1, \tag{2.39d}$$

$$T_s(\tfrac{\gamma_{11}B_1}{R_{11}})+\delta_c\gamma_{11}B_1\le\bar{L}_1, \tag{2.39e}$$

$$T_s(\tfrac{\gamma_{21}}{R_{21}}+\tfrac{\gamma_{23}}{R_{23}})B_2+\delta_c(\gamma_{21}+\gamma_{23})B_2\le\bar{L}_2, \tag{2.39f}$$

$$t_{\mathrm{loc}_k}\le\bar{L}_k, \quad k=1,2, \tag{2.39g}$$

$$0\le P_{k1},P_{k2}\le\bar{P}_k, \quad k=1,2, \tag{2.39h}$$

$$0\le R_{23}\le\log_2(1+\alpha_2P_{23}), \tag{2.39i}$$

$$\{R_{11},R_{21}\}\in\mathcal{R}, \tag{2.39j}$$

where $\bar{L}_k=L_k-t_{\mathrm{DL}_k}$.

2.8.1 Energy-Optimal Local Execution

In this thesis, we will consider the dynamic voltage scaling approach to local computation (Zhang *et al.*, 2013; Wang *et al.*, 2016). This approach involves adjusting the CPU cycle frequency of the mobile devices so as to minimize the energy required to complete a task within a given deadline. Indeed, for the class of problems that we are considering, the minimum energy required for local processing of μ_kB_k bits with

a latency constraint of L_k is (Zhang *et al.*, 2013)

$$E_{\text{loc}_k} = \frac{M_k(\mu_k B_k)^3}{L_k^2}, \tag{2.40}$$

where M_k is a constant that depends on the chip architecture. This expression not only gives us the form of $\mathcal{F}_k(\cdot)$ in (2.39), it also ensures that the local component of the task is completed before the deadline. Therefore, we can remove the local computational latency constraints in (2.39g).

2.8.2 Full Multiple Access Scheme

Using the same insights as those used in the complete computation offloading case, we can show that the optimal solution for the problem in (2.39) is obtained when the constraints in (2.39e) and (2.39f) hold with equality. Therefore, we can find closed-form expressions for the optimal solutions for γ_{ki} in terms of the other parameters of the problem,

$$\gamma_{11} = \frac{\bar{L}_1 R_{11}}{B_1(T_s+\delta_c R_{11})}, \tag{2.41a}$$

$$\gamma_{21} = \frac{\bar{L}_1 R_{21}}{B_2(T_s+\delta_c R_{11})}, \tag{2.41b}$$

$$\gamma_{23} = \frac{R_{23}}{B_2(T_s+\delta_c R_{23})}\big(\bar{L}_2 - \frac{T_s+\delta_c R_{21}}{T_s+\delta_c R_{11}}\bar{L}_1\big), \tag{2.41c}$$

where (2.41b) results from the fact that $\tau_1 = \frac{\gamma_{11}B_1}{R_{11}} = \frac{\gamma_{21}B_2}{R_{21}}$.

By substituting these closed-form expressions for γ_{ki} into (2.39), the remaining

optimization problem can be written as

$$\min_{\substack{P_{11},P_{21},P_{23},\\ R_{11},R_{21},R_{23}}} \quad \tfrac{T_s\bar{L}_1}{T_s+\delta_c R_{11}}(P_{11}+P_{21}) + \tfrac{T_s}{T_s+\delta_c R_{23}}\big(\bar{L}_2 - \tfrac{T_s+\delta_c R_{21}}{T_s+\delta_c R_{11}}\bar{L}_1\big)P_{23} + \tfrac{M_1}{L_1^2}\big(B_1 - \tfrac{\bar{L}_1 R_{11}}{T_s+\delta_c R_{11}}\big)^3$$
$$+ \tfrac{M_2}{L_2^2}\Big(B_2 - \tfrac{\bar{L}_2 R_{23}(T_s+\delta_c R_{11})+\bar{L}_1 T_s(R_{21}-R_{23})}{(T_s+\delta_c R_{11})(T_s+\delta_c R_{23})}\Big)^3 \tag{2.42a}$$

$$\text{s.t.} \quad (2.39\text{b}) - (2.39\text{d}), (2.39\text{h}) - (2.39\text{i}), \tag{2.42b}$$

$$0 \le R_{k1} \le \log_2(1+\alpha_k P_{k1}), \quad k=1,2, \tag{2.42c}$$

$$R_{11}+R_{21} \le \log_2(1+\alpha_1 P_{11}+\alpha_2 P_{21}). \tag{2.42d}$$

In the next step toward solving the problem we obtain closed-form expressions for the transmission powers by decomposing the problem in (2.42) as

$$\min_{R_{11},R_{21},R_{23}} \quad (2.42\text{a}) \qquad \min_{P_{11},P_{21},P_{23}} \quad (2.42\text{a}) \tag{2.43}$$
$$\text{s.t.} \quad (2.39\text{b}) - (2.39\text{d}), \qquad \text{s.t.} \quad (2.39\text{h}), (2.39\text{i}), (2.42\text{c}) - (2.42\text{d}).$$

Given a set of transmission rates (R_{11}, R_{21}, R_{23}), the optimal solution for P_{23} is the minimum feasible value, i.e., $P_{23} = \frac{2^{R_{23}}-1}{\alpha_2}$ and closed-form expressions for the transmission powers of the users in the first time slot can be obtained by employing the technique that was explained in Section 2.5. In this section, we solve the problem for the first subcase of the scenario $\frac{\alpha_1}{\alpha_2} \le 1$, which forms an analogy with the first subcase of the complete computation offloading scenario (see Section 2.5.1). The problem in the other cases can be solved by following similar steps.

Given the closed-form solutions for all the fractions γ_{ki} and all the transmission powers P_{ki} in terms of the transmission rates, the problem of minimizing the total energy consumption of the users can be reduced to the following three-variable

optimization problem

$$\min_{R_{11},R_{21},R_{23}} \quad \frac{T_s\bar{L}_1}{T_s+\delta_c R_{11}}\left(\frac{2^{R_{11}}-1}{\alpha_1}+\frac{2^{R_{11}}(2^{R_{21}}-1)}{\alpha_2}\right)+\frac{T_s}{T_s+\delta_c R_{23}}\left(\bar{L}_2-\frac{T_s+\delta_c R_{21}}{T_s+\delta_c R_{11}}\bar{L}_1\right)\frac{2^{R_{23}}-1}{\alpha_2}$$
$$+\frac{M_1}{L_1^2}\left(B_1-\frac{\bar{L}_1 R_{11}}{T_s+\delta_c R_{11}}\right)^3$$
$$+\frac{M_2}{L_2^2}\left(B_2-\frac{\bar{L}_2 R_{23}(T_s+\delta_c R_{11})+\bar{L}_1 T_s(R_{21}-R_{23})}{(T_s+\delta_c R_{11})(T_s+\delta_c R_{23})}\right)^3 \tag{2.44a}$$

$$\text{s.t.} \quad (2.39\text{b})-(2.39\text{d}), \tag{2.44b}$$

$$0\le R_{ki}\le \log_2(1+\alpha_k\bar{P}_k), \quad k=1,2. \tag{2.44c}$$

It is shown in Appendix 2.F that the objective function of the problem in (2.44) is a quasi-convex function of each of the variables when the other two variables are given. Therefore, the coordinate descent algorithm can be applied to find a stationary solution for the transmission rates (Hong *et al.*, 2016, Theorem 1). In all our numerical experiments that approach converged to the globally optimal solution.

2.8.3 Time Division Multiple Access Scheme

For a two-user offloading system that employs the TDMA scheme, each user operates in its own time slot. By simplifying the notation in a natural way, the total energy

minimization problem can be written as

$$\min_{\substack{P_1,P_2,R_1,R_2,\\ \gamma_1,\gamma_2}} \quad T_s(\tfrac{\gamma_1 B_1}{R_1})P_1 + T_s(\tfrac{\gamma_2 B_2}{R_2})P_2 + \tfrac{M_1}{L_1^2}\big((1-\gamma_1)B_1\big)^3 + \tfrac{M_2}{L_2^2}\big((1-\gamma_2)B_2\big)^3 \quad (2.45a)$$

$$\text{s.t.} \quad 0 \le \gamma_k \le 1, \quad k = 1,2, \quad (2.45b)$$

$$T_s(\tfrac{\gamma_1 B_1}{R_1}) + \delta_c \gamma_1 B_1 \le \bar{L}_1, \quad (2.45c)$$

$$T_s(\tfrac{\gamma_1 B_1}{R_1}) + T_s(\tfrac{\gamma_2 B_2}{R_2}) + \delta_c \gamma_2 B_2 \le \bar{L}_2, \quad (2.45d)$$

$$0 \le P_{k1}, P_{k2} \le \bar{P}_k, \quad k = 1,2, \quad (2.45e)$$

$$0 \le R_k \le \log_2(1+\alpha_k P_k), \quad k = 1,2. \quad (2.45f)$$

Since only one user is transmitting during each time slot, it can be shown that the optimal transmission power of each user is the minimum feasible value, i.e., $P_k = \frac{2^{R_k}-1}{\alpha_k}$. Moreover, it can be shown that for any optimal solution of the problem in (2.45), the constraint in (2.45d) holds with equality, i.e.,

$$\gamma_2 = \tfrac{\bar{L}_2 - T_s(\gamma_1 B1/R_1)}{B_2(T_s/R_2+\delta_c)}. \quad (2.46)$$

By substituting the obtained closed-form expressions we can rewrite the problem in (2.45) as

$$\min_{R_1,R_2,\gamma_1} \quad T_s(\tfrac{\gamma_1 B_1}{R_1})(\tfrac{2^{R_1}-1}{\alpha_1}) + T_s(\tfrac{\bar{L}_2 - T_s(\gamma_1 B_1/R_1)}{T_s+R_2\delta_c})(\tfrac{2^{R_2}-1}{\alpha_2}) + \tfrac{M_1}{L_1^2}\big((1-\gamma_1)B_1\big)^3 + \tfrac{M_2}{L_2^2}\big(B_2 - \tfrac{\bar{L}_2 - T_s(\gamma_1 B1/R_1)}{T_s/R_2+\delta_c}\big)^3 \tag{2.47a}$$

$$\text{s.t.} \quad 0 \le \gamma_1 \le 1, \tag{2.47b}$$

$$0 \le \tfrac{\bar{L}_2 - T_s(\gamma_1 B1/R_1)}{B_2(T_s/R_2+\delta_c)} \le 1, \tag{2.47c}$$

$$0 \le R_k \le \log_2(1+\alpha_k \bar{P}_k), \quad k=1,2. \tag{2.47d}$$

The three-variable optimization problem in (2.47) is convex in terms of each of the variables when the other two variables are given. Hence, by applying coordinate descent optimization methods a stationary solution of the problem can be obtained. In all of our numerical experiments the coordinate descent algorithm converged to the globally optimal solution.

In the case of complete computation offloading, we were able to show that when the channel gains of both users are equal and the power budgets are above a threshold, the optimized TDMA scheme obtains the optimal energy consumption of the full multiple access scheme; see Proposition 2. As we will formalize in the following proposition, we can extend that result to the case of partial offloading.

Proposition 3. *Let $\gamma_{ki}^\star$ and $P_{ki}^\star$ denote an optimal solution for γ_{ki} and P_{ki} in* (2.44) *when $|h_1|^2 = |h_2|^2$. If $P_{11}^\star + P_{21}^\star \le \min\{\bar{P}_1, \bar{P}_2\}$, then the TDMA scheme can obtain the optimal energy consumption of the full multiple access scheme with the offloaded portions of the first and second users equal to $\gamma_1 = \gamma_{11}^\star$ and $\gamma_2 = \gamma_{21}^\star + \gamma_{23}^\star$, respectively.*

Proof. Let $\hat{\gamma}_{ki}$ denote offloading fractions of an arbitrary instance of the full multiple

access scheme. If we select $\gamma_1 = \hat{\gamma}_{11}$ and $\gamma_2 = \hat{\gamma}_{21}+\hat{\gamma}_{23}$, then we can apply Proposition 2 to show that the optimized TDMA scheme achieves the same energy consumption as the full multiple access scheme with offloading fractions $\hat{\gamma}_{ki}$. The proposition follows by looking at the case where the offloading fractions of the full multiple access scheme are optimal; i.e., $\hat{\gamma}_{ki} = \gamma^{\star}_{ki}$. □

2.9 Numerical Results

In this section we will illustrate the performance of the multiple access computation offloading schemes that we have considered in some simple proof-of-concept experiments that highlight the insights that have been developed. We consider a two-user communication system in which the users have the opportunity to offload their latency-constrained computational tasks to a computationally-rich access point. In Section 2.9.1 we will illustrate the impact of the choice of the multiple access scheme in the case of complete computation offloading. Then, in Section 2.9.2 we will compare the performance of binary and partial computation offloading under the full multiple access and TDMA schemes.

2.9.1 Complete Computation Offloading

In our first experiment, we examine the total energy usage of two offloading users as the power gain of user 1's channel, $|h_1|^2$, changes. In particular, we set the power budgets, the latencies, the channel gain of the second user and the receiver noise variance to be constant values, namely, $\bar{P}_1 = 0.3$, $\bar{P}_2 = 0.5$, $L_1 = 2.5$s, $L_2 = 3.3$s, $|h_2|^2 = 0.1$, and $\sigma^2 = 0.1$, respectively. The number of bits that are needed to

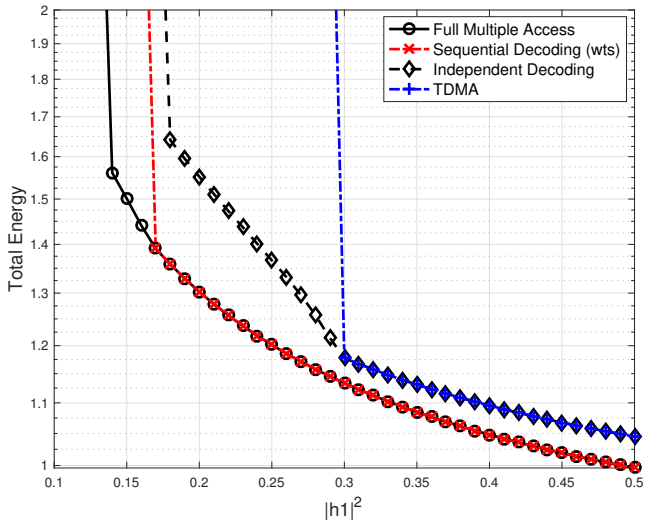


Figure 2.4: Energy required to offload the tasks for the optimal, sequential decoding without time sharing, TDMA, and independent decoding schemes as a function of $|h_1|^2$.

describe the tasks to be offloaded are $B_1 = B_2 = 10^6$ and we set the sum of the time of execution of the application in the cloud and the time it takes to download the result to the mobile users to be $T_1 = T_2 = 0.5$s. The symbol period of the channel is set to be $T_s = 10^{-6}$s.

It can be seen from Fig. 2.4 that for small values of the power gain of user 1 only the full multiple access scheme is able to offload both tasks while meeting the constraints. This implies that the optimized transmission rates of the first time slot are on the dominant face of the capacity region of the multiple access channel, cf. Fig. 2.1(d). For larger values of $|h_1|^2$, sequential decoding (wts) is also feasible and it can be seen that it has the same total energy consumption as the full multiple access scheme. As the value of $|h_1|^2$ increases, the independent decoding scheme becomes feasible, but the total energy consumption of independent decoding is significantly

greater than the energy consumed by the full multiple access and sequential decoding (wts) schemes. As $|h_1|^2$ is increased further, the TDMA scheme eventually becomes feasible. As argued in Section 2.7, once it becomes feasible, it achieves the same energy consumption as the independent decoding scheme. (In TDMA, the decoders work independently.) However, for this range of values of $|h_1|^2$ the full multiple access and sequential decoding (wts) schemes are able to offload both tasks using less energy.

Figs 2.5 and 2.6 exhibit the duration of each time slot in the TDMA and independent decoding schemes for the same system parameters as those in Fig. 2.4, respectively. (It has been shown that in the full multiple access scheme the duration of time slots are independent from the channel gains and they only depend on the latency constraints of the users; i.e., $\tau_1 = \tilde{L}_1 = 2.5$s, $\tau_2 = 0$, and $\tau_3 = \tilde{L}_2 - \tilde{L}_1 = 0.8$s.) It can be seen in Fig. 2.6 that when $|h_1|^2 < 0.3$ all three time slots are needed for the independent decoding scheme to achieve the optimal energy consumption. In those scenarios, the TDMA scheme is not feasible; see Figs 2.4 and 2.5. Figs 2.5 and 2.6 also illustrate the observation made in Section 2.7 that when the TDMA scheme is feasible it is optimal for the independent decoding scheme. That is, one can construct an optimal signalling scheme for independent decoding that has $\tau_1 = 0$.

In our next experiment we will illustrate the impact of the latency of the second user's task, L_2. To do so, we tighten the first user's latency constraint to $L_1 = 1.8$s and we provide user 2 with a larger channel gain, $|h_2|^2 = 0.24$. The number of bits needed to describe the tasks are changed to $B_1 = 3 \times 10^6$ and $B_2 = 5 \times 10^6$, and the receiver noise variance is set to be $\sigma^2 = 2 \times 10^{-3}$. The other system parameters remain the same.

In Fig. 2.7 we plot the energy required to offload both tasks as a function of

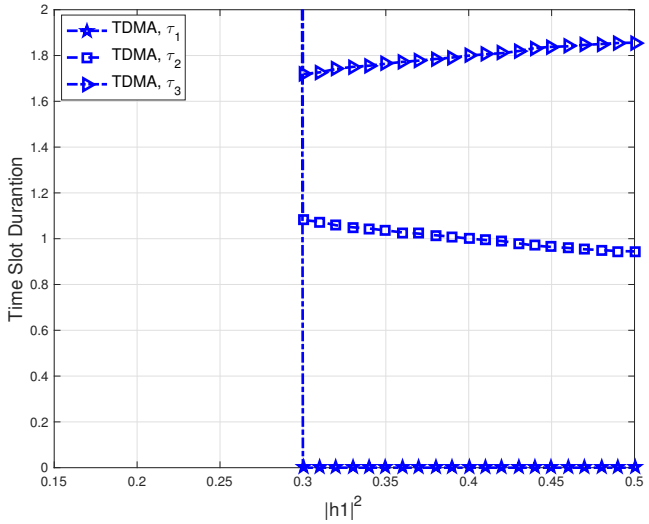


Figure 2.5: Duration of each time slot in the TDMA scheme.

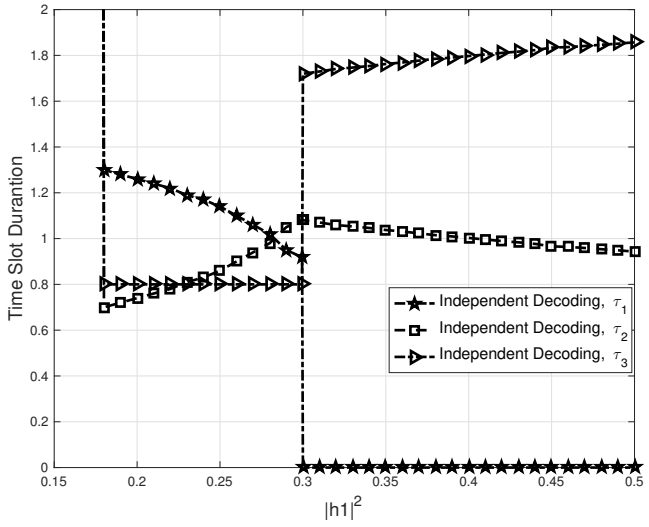


Figure 2.6: Duration of each time slot in the independent decoding scheme.

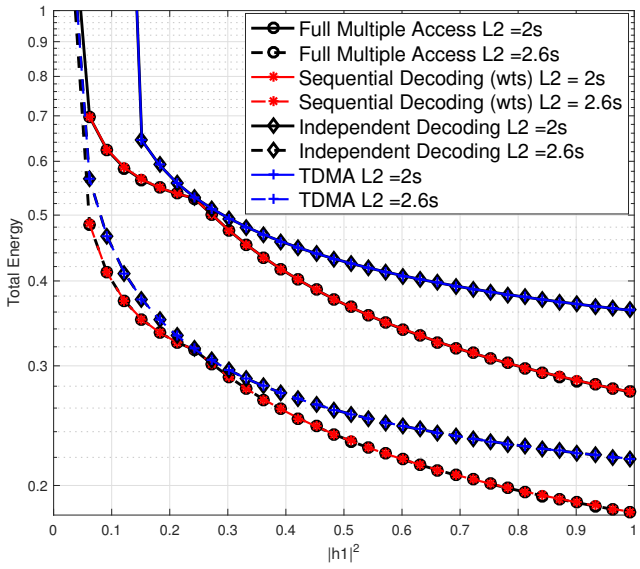


Figure 2.7: Energy required to offload the tasks as a function of $|h_1|^2$ for two different values for the second user's latency, $L_2 = 2$s and $L_2 = 2.6$s.

the first user's channel gain, $|h_1|^2$, for two values of the latency for the second user, namely $L_2 = 2$s and $L_2 = 2.6$s.

The sets of curves for the two latencies exhibit similar characteristics, but these characteristics, and the reduced energy consumption of the full multiple access scheme, are more pronounced in the case where the latency is tighter. (As expected more energy is required to offload the tasks in that case.) Both sets of curves in Fig. 2.7 demonstrate the ability of the full multiple access and sequential decoding (wts) schemes to take advantage of skewed channel conditions. In particular, when $|h_1|^2$ is small these schemes are able to offload both tasks, whereas TDMA and independent decoding are unable to do so. (The extended range of the full multiple access and sequential decoding (wts) schemes is quite significant in the lower latency case.) When $|h_1|^2$ is large, the full multiple access and sequential decoding (wts) schemes are able

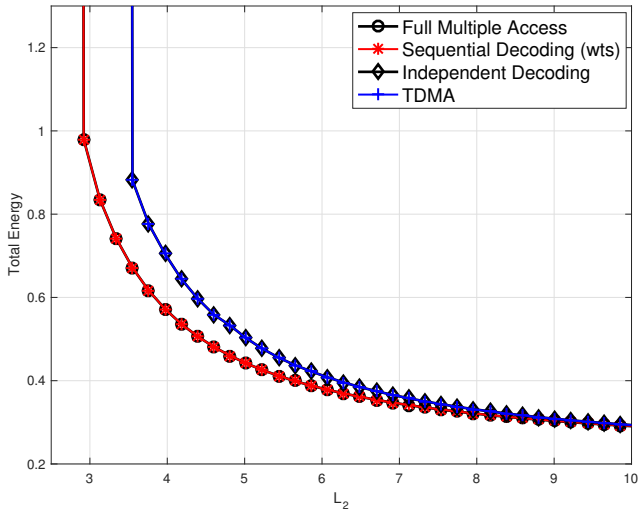


Figure 2.8: Energy required to offload the tasks as a function of L_2, the latency of the second user's application.

to provide a substantial reduction in the energy required to perform the offloading. The cusp that can be seen in each of the curves for the full multiple access scheme, which occurs at $\alpha_1 = \alpha_2$, represents switching between the cases in Algorithm 1 (from Case A to Case B). Fig. 2.7 also illustrates the impact of Proposition 2; namely that when the channel gains are equal and the power budgets are above an explicit threshold, TDMA, independent decoding and sequential decoding (wts) are all able to achieve the same minimum energy consumption as the full multiple access scheme.

In Fig. 2.8 we plot the energy required to offload both tasks as a function of the second user's latency constraint, L_2, for the case in which $|h_1|^2 = 0.6$ and $|h_2|^2 = 0.06$. In order to guarantee the feasibility of all four multiple access schemes we set the latency of user 1 equal to $L_1 = 2$s and the transmitted number of bits for the first and the second users are $B_1 = 4 \times 10^6$ bits and $B_2 = 8 \times 10^6$ bits, respectively. The other parameters are the same as those that were used for Fig. 2.7.

Fig. 2.8 reinforces the observation from Fig. 2.7 that as the latency constraints are tightened, the ability of the full multiple access scheme to exploit all the capabilities of the multiple access channel offers increasing performance gains. Fig. 2.8 also illustrates the fact that the full multiple access scheme can satisfy tighter latency constraints than the TDMA and independent decoding schemes. (For all the values of L_2 that we considered in Fig. 2.8, sequential decoding (wts) is optimal and TDMA is an optimal scheme for independent decoding.)

2.9.2 Binary and Partial Computation Offloading

In the second phase of our numerical analysis we study the case in which the computational tasks of the users are infinitesimally divisible and hence partial offloading can be employed. In this phase, we consider the full multiple access and TDMA schemes, and we examine the total energy consumption of partial computation offloading under different parameter settings. We also compare the energy consumption of partial offloading to that of the binary offloading scheme that would be used if the tasks were considered to be indivisible.

We first illustrate the performance of the full multiple access and TDMA schemes as a function of the channel gain of user 1, $|h_1|^2$. We set the power budgets, the latencies, the channel gain of the second user and the receiver noise variance to be constant values, namely, $\bar{P}_1 = \bar{P}_2 = 0.5$, $L_1 = 1.5$s, $L_2 = 2$s, $|h_2|^2 = 0.5$, and $\sigma^2 = 10^{-3}$, respectively. The number of bits to describe the problems are $B_1 = 2\times10^6$ and $B_2 = 6 \times 10^6$ and we set the time it takes to download the result to the mobile users to be $t_{\mathrm{DL}_1} = t_{\mathrm{DL}_2} = 0.2$s. The symbol period of the channel is set to be $T_s = 10^{-6}$s. As explained in Section 2.8, we consider data-partitionable computational

tasks for which the (optimal) local energy consumption can be modeled as a function of number of bits; see Section 2.8.1. In order to be consistent with the measurements in (Miettinen and Nurminen, 2010), we set the constants M_k in the local energy consumption expression in (3.6) to $M_1 = M_2 = 10^{-18}$ (Zhang *et al.*, 2013; Wang *et al.*, 2016).

Fig. 2.9 illustrates the total energy consumption of the users in the partial and binary computation offloading scenarios for the full multiple access and TDMA schemes, and Fig. 2.10 illustrates the corresponding fraction of the bits that each user offloads to the access point in the partial offloading scenario. It can be seen from Fig. 2.9 that in both the partial and binary offloading scenarios taking advantage of the full capabilities of the channel enables the users to execute their tasks with substantially less energy consumption compared to the TDMA scheme and the gap between the energy usage of these schemes becomes larger as the channel gain of the first user increases. Fig. 2.9 also illustrates the fact that, since the power budgets are above the threshold in Proposition 3, when the channel gains are equal TDMA can achieve the minimum energy consumption. Another observation in Fig. 2.9 is that for large values of $|h_1|^2$, binary offloading with the full multiple access scheme achieves lower energy consumption than partial offloading using TDMA. This is due to the fact that the full multiple access scheme's ability to utilize all the capabilities of the channel overcomes the limitations of binary offloading when the channel gains are sufficiently different.

It can be seen in Fig. 2.10 that using the full capabilities of the channel enables the users to offload larger fraction of bits to the access point than TDMA. This results in a lower total energy consumption; see Fig. 2.9. Moreover, when the channel gains

are equal, the portions that the users offload in the TDMA scheme are the optimal portions offloaded by the corresponding users in the full multiple access scheme, which verifies Proposition 3.

Fig. 2.10 exhibits that in the TDMA case, as one would expect, an increase in the channel gain of the first user leads to an increased fraction of bits that each user offloads. For the full multiple access scheme, an increase in the channel gain of the first user leads to an increase in the fraction of bits offloaded by that user. This observation can be verified from the expression in (2.41a) and the fact that by increasing the channel gain of the first user, a higher transmission rate will be employed by that user. However, the offloaded fraction of the second user does not change in a monotonic manner. When $|h_1|^2 \leq |h_2|^2$, an increase in the channel gain of user 1 results in a decrease in the portion of bits offloaded by user 2, while for $|h_1|^2 \geq |h_2|^2$ the offloaded portion of user 2 increases. That is because the minimum energy consumption of the system depends on the ratio between the channel gains (see Section 2.5), which also affects the portion of bits offloaded by the second user; see (2.41b) and (2.41c).

Our final numerical experiments examine the average performance of the full multiple access and TDMA schemes under a simple fading channel model for both binary and partial computation offloading. The channel model has a (large-scale) path-loss exponent of 3 and the small-scale fading has a Rayleigh distribution. The latency constraints of the tasks are $L_1 = 1.7$s, $L_2 = 2$s and the descriptions of the tasks require $B_1 = 2 \times 10^6$ and $B_2 = 5 \times 10^6$ bits, respectively. The second user is placed at a distance of 500m from the access point, and in Fig. 2.11 we plot the average energy required to offload the tasks as user 1 moves from a position 100m from the

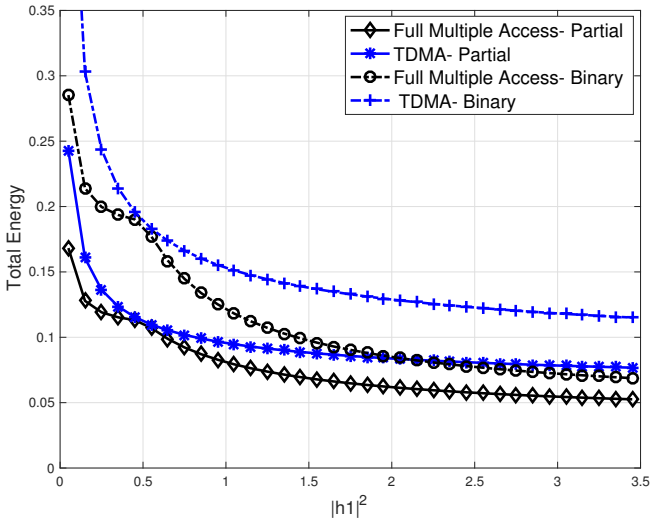


Figure 2.9: The total energy consumption of the two-user system with the full multiple access and TDMA schemes in the cases of binary and partial computation offloading as a function of $|h_1|^2$.

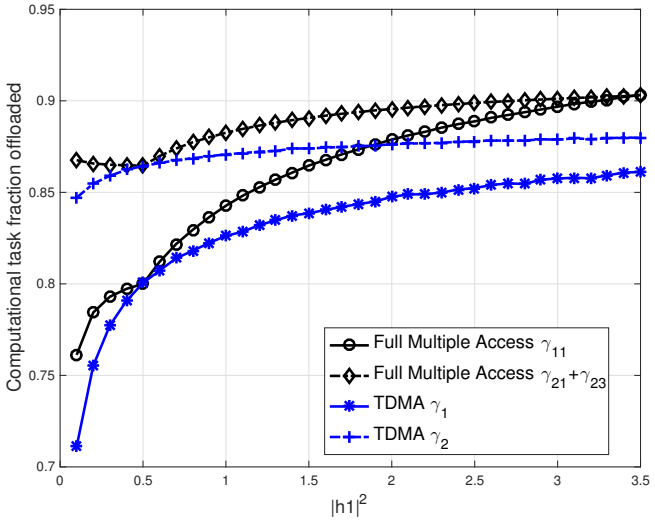


Figure 2.10: The fraction of the total number of bits offloaded by each user with the full multiple access and TDMA schemes in the partial computation offloading case as a function of $|h_1|^2$.

access point to a position 900m away. The average is taken over 10^5 realizations of the channel pairs for which the two schemes provided a feasible solution in the complete and partial computation offloading cases. The other parameters of the problem are set to be the same as those in the experiment that produced Figs 2.9 and 2.10.

Figs 2.11 and 2.12 demonstrate that the insights that were developed analytically in Section 2.7 for individual channel realizations also reflect the performance on average. In particular, when the users are at similar distances from the access point, then the channel gains are likely to be similar and hence we would expect the performance of TDMA to be close to that of the full multiple access scheme. This is indeed the case in Figs 2.11 and 2.12. When the users are at significantly different distances from the access point, their channel gains are likely to be quite different, and hence we would expect the full multiple access scheme to have significantly better performance than TDMA. Once again, Fig. 2.11 confirms that insight, and Fig. 2.12 shows how the full multiple access scheme enables a greater fraction of each task to be offloaded. Indeed, when user 1 is far from the access point, binary offloading with the full multiple access scheme consumes less energy than partial offloading with TDMA.

2.10 Conclusion

In this work, we have obtained closed-form and quasi-closed-form solutions to problems of optimizing the communication resource allocation so as to minimize the energy that the users expend in a computational offloading system with two users and plentiful computational resources. We have provided solutions for both indivisible and infinitesimally divisible computational tasks, and we consider the full multiple access

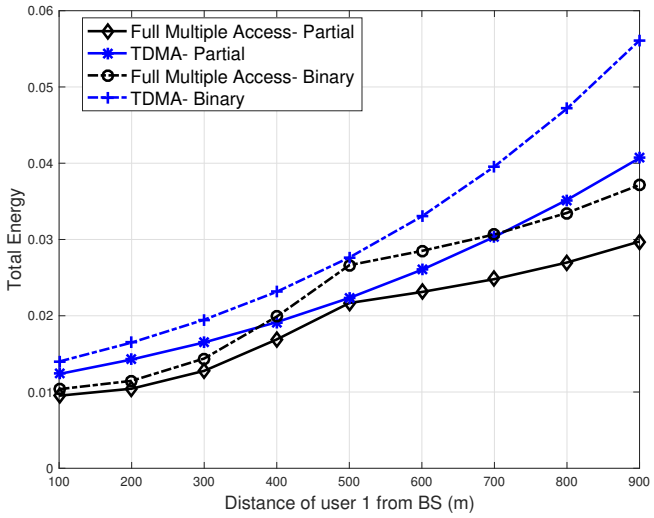


Figure 2.11: Average energy required to offload the computational tasks for the full multiple access and TDMA schemes against the distance of user 1 from the access point in the binary and partial computation offloading scenarios. User 2 is 500m from the access point.

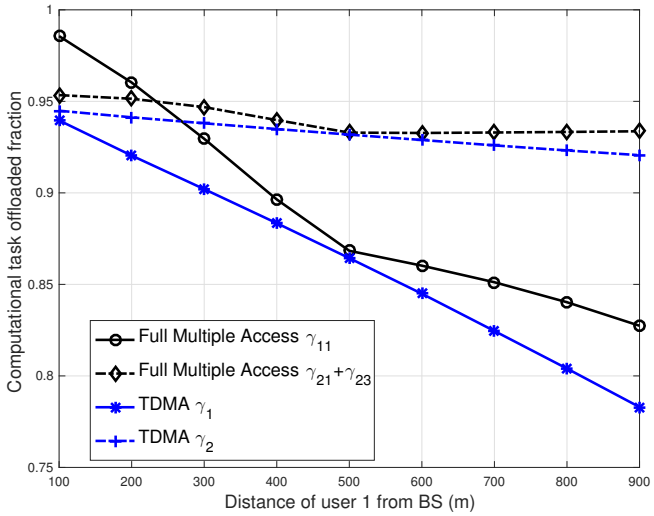


Figure 2.12: Average computation fraction offloaded by the users for the full multiple access and TDMA schemes against the distance of user 1 from the access point in the partial computation offloading scenario. User 2 is 500m from the access point.

scheme along with some of the simplified schemes, namely, TDMA, sequential decoding (without time sharing), and independent decoding. In broad terms, the structure of our solutions suggests that if the channel gains of the users are similar (or if the latency constraints are loose), then the implementation simplicity of TDMA may outweigh the energy reduction offered by the full multiple access scheme. However, when the latency constraints are tight and the channel gains are quite different from each other, the full multiple access scheme provides a substantial reduction in the energy required to complete the tasks.

2.A Objective Function in (2.11) is Increasing

After removing the positive coefficient $(\frac{B_2-\tilde{L}_1 R_{21}}{\alpha_2})$, the first derivative of the objective function in (2.11) with respect to R_{23} is

$$\frac{d}{dR_{23}}\Big(\frac{2^{R_{23}}-1}{R_{23}}\Big) = \Big[\frac{R_{23}2^{R_{23}}\ln 2-(2^{R_{23}}-1)}{R_{23}^2}\Big]. \tag{2.48}$$

To show that the numerator on the right hand side of (2.48) is positive for $R_{23} > 0$, we observe that

$$\frac{d}{dR_{23}}\Big(R_{23}2^{R_{23}}\ln 2 - (2^{R_{23}}-1)\Big) = R_{23}2^{R_{23}}(\ln 2)^2, \tag{2.49}$$

which is positive for $R_{23} > 0$. Since the numerator of (2.48) is zero for $R_{23} = 0$, the positivity of (2.49) implies that the expression in (2.48) is positive for $R_{23} > 0$ and hence that the objective in (2.11) is an increasing function of R_{23} over the feasible set.

(a) Three-time-slot system (b) Two-time-slot system

Figure 2.13: In the three-time-slot system there is one time slot (the second) in which user 1 is the only user transmitting. In the two-time-slot system user 1 completes its transmission in the first time slot and only user 2 has a slot in which it transmits alone.

2.B Optimality of Two Time Slot Scenario

Let us consider the three-time-slot system illustrated in Fig. 2.13(a), with the notation for the time slot durations, the transmission powers and transmission rates being as defined in Section 2.3. Let us also consider the two-time-slot system depicted in Fig. 2.13(b), in which the power and the rate of the k^{th} user in the first time slot are denoted by P'_k and R'_k, respectively. We will show that for the full multiple access scheme, if the users' tasks can be offloaded using the three-time-slot system with a given energy, then there exists a power and rate allocation for the two-time-slot system that can offload the tasks using the same energy. It will suffice to assume that the power and rate allocations in the second time slot of the two-time-slot system are the same as those in the third time slot of the three-time-slot system. Furthermore, it will suffice to add the constraint that the energy consumption of each user is the same in both systems.

For the energy and the number of transmitted bits of each user to be the same,

P'_k and R'_k must satisfy

$$\tau_1 P_{11} + (\gamma - \tau_1) P_{12} = \gamma P'_1 \tag{2.50a}$$

$$\tau_1 P_{21} + (\tilde{L}_2 - \gamma) P_{22} = \gamma P'_2 + (\tilde{L}_2 - \gamma) P_{22}, \tag{2.50b}$$

$$\tau_1 R_{11} + (\gamma - \tau_1) R_{12} = \gamma R'_1, \tag{2.50c}$$

$$\tau_1 R_{21} + (\tilde{L}_2 - \gamma) R_{22} = \gamma R'_2 + (\tilde{L}_2 - \gamma) R_{22}. \tag{2.50d}$$

The solution of that set of linear equations is

$$P'_1 = \tfrac{\tau_1}{\gamma} P_{11} + \tfrac{\gamma - \tau_1}{\gamma} P_{12} \quad \text{and} \quad P'_2 = \tfrac{\tau_1}{\gamma} P_{21}, \tag{2.51a}$$

$$R'_1 = \tfrac{\tau_1}{T} R_{11} + \tfrac{\gamma - \tau_1}{\gamma} R_{12} \quad \text{and} \quad R'_2 = \tfrac{\tau_1}{\gamma} R_{21}. \tag{2.51b}$$

What remains is to show that these power and rate allocations satisfy the rate region constraint for the first time slot of the two-time-slot system, namely,

$$R'_1 \leq \log_2(1 + \alpha_1 P'_1), \quad R'_2 \leq \log_2(1 + \alpha_2 P'_2), \tag{2.52a}$$

$$R'_1 + R'_2 \leq \log_2(1 + \alpha_1 P'_1 + \alpha_2 P'_2). \tag{2.52b}$$

The inequalities in (2.52) can be rewritten in terms of the rates and powers of the three-time-slot case as follows,

$$\tfrac{\tau_1}{T} R_{11} + \tfrac{\gamma - \tau_1}{\gamma} R_{12} \leq \log_2\big(1 + \alpha_1 \rho\big), \tag{2.53a}$$

$$\tfrac{\tau_1}{\gamma} R_{21} \leq \log_2\big(1 + \alpha_2 (\tfrac{\tau_1}{\gamma} P_{21})\big), \tag{2.53b}$$

$$\tfrac{\tau_1}{T} R_{11} + \tfrac{\gamma - \tau_1}{\gamma} R_{12} + \tfrac{\tau_1}{\gamma} R_{21} \leq \log_2\big(1 + \alpha_1 \rho + \alpha_2 (\tfrac{\tau_1}{\gamma} P_{21})\big), \tag{2.53c}$$

where $\rho = \frac{\tau_1}{\gamma} P_{11} + \frac{\gamma - \tau_1}{\gamma} P_{12}$. To establish the validity of the inequalities in (2.53), we use the rate constraints of the three-time-slot case and the concavity of the logarithm. For example, since the power and rate allocations for the three-time-slot case are assumed to be valid, we have that $R_{21} \leq \log_2(1 + \alpha_2 P_{21})$. Using that and the concavity of the logarithm we have that

$$\frac{\tau_1}{\gamma} R_{21} \leq \frac{\tau_1}{\gamma} \log_2(1 + \alpha_2 P_{21}) \leq \log_2\big(1 + \alpha_2(\tfrac{\tau_1}{\gamma} P_{21})\big), \tag{2.54}$$

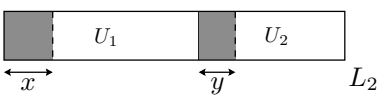
and hence that (2.53b) holds. The inequalities in (2.53a) and (2.53c) can be established in an analogous way.

2.C Convexity of Objective Function in (2.20a)

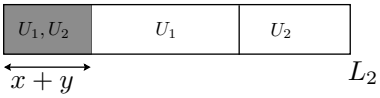
Since the exponential function is convex, the second term of (2.20a) is convex. To prove the convexity of the third term, and hence the convexity of the function as a whole, we let $f_3(R_{21})$ denote the third term of the objective function and evaluate its second derivative, $f_3''(R_{21}) = \ln(2)^2 (\frac{\tilde{L}_1^2}{\alpha_2(\tilde{L}_2 - \tilde{L}_1)})\ 2^{\frac{B_2 - \tilde{L}_1 R_{21}}{\tilde{L}_2 - \tilde{L}_1}}$. By our standing assumption that $\tilde{L}_2 \geq \tilde{L}_1$, $f_3''(R_{21}) \geq 0$ and hence $f_3(R_{21})$ is convex.

2.D Optimality of TDMA for Independent Decoding

Consider the structure of TDMA signalling illustrated in Fig. 2.14(a), where x and y are portions of the duration of the first and second users' transmissions, respectively.



(a) TDMA scheme with two time slots



(b) Independent decoding scheme with three time slots

Figure 2.14: The structure of a TDMA scheme (a), and the corresponding independent decoding scheme (b).

In the independent decoding signalling structure illustrated in Fig. 2.14(b), the intervals x and y are combined, with both users transmitting simultaneously and being decoded independently. Let P_k and R_k denote the transmission power and transmission rate of the k^{th} user in the TDMA scheme, respectively, and P'_k and R'_k denote those quantities in the first time slot of the independent decoding scheme. Since the number of transmitted bits in both cases must be the same, we have that

$$(x+y)R'_1 = xR_1 = x\log_2(1+\alpha_1 P_1) \tag{2.55a}$$

$$(x+y)R'_2 = yR_2 = y\log_2(1+\alpha_2 P_2). \tag{2.55b}$$

We will show that we cannot find a set of transmission rates and transmission powers for which the energy consumption in independent decoding scheme is less than the energy consumption of the TDMA scheme; i.e., there are no allocations that satisfy $(x+y)(P'_1+P'_2) \leq xP_1 + yP_2$.

As shown in Section 2.6.3, the optimal transmission rates and transmission powers in the independent decoding scheme can be written as

$$R'_1 = \log_2\big(1+\tfrac{\alpha_1 P'_1}{1+\alpha_2 P'_2}\big) \qquad R'_2 = \log_2\big(1+\tfrac{\alpha_2 P'_2}{1+\alpha_1 P'_1}\big). \tag{2.56}$$

Considering (2.55) and (2.56) the problem of finding the rates for the independent

decoding scheme such that the total energy consumption is less than that of TDMA can be written as

$$\begin{aligned} \text{find} \quad & P_1', P_2' && \text{(2.57a)} \\ \text{s.t.} \quad & \log_2\big(1+\tfrac{\alpha_1 P_1'}{1+\alpha_2 P_2'}\big) = \lambda \log_2(1+\alpha_1 P_1), && \text{(2.57b)} \\ & \log_2\big(1+\tfrac{\alpha_2 P_2'}{1+\alpha_1 P_1'}\big) = (1-\lambda) \log_2(1+\alpha_2 P_2), && \text{(2.57c)} \\ & P_1' + P_2' \leq \lambda P_1 + (1-\lambda) P_2, && \text{(2.57d)} \\ & 0 \leq \lambda \leq 1, && \text{(2.57e)} \end{aligned}$$

where $\lambda = \frac{x}{x+y}$. Using the constraints in (2.57b) and (2.57c) we can rewrite P_1' and P_2' in terms of P_1 and P_2, namely,

$$\alpha_1 P_1' = \frac{\phi_2^{(1-\lambda)}(\phi_1^{\lambda}-1)}{\phi_1^{\lambda}+\phi_2^{(1-\lambda)}-\phi_1^{\lambda}\phi_2^{(1-\lambda)}}, \quad \alpha_2 P_2' = \frac{\phi_1^{\lambda}(\phi_2^{(1-\lambda)}-1)}{\phi_1^{\lambda}+\phi_2^{(1-\lambda)}-\phi_1^{\lambda}\phi_2^{(1-\lambda)}} \tag{2.58}$$

in which $\phi_1 = (1+\alpha_1 P_1)$ and $\phi_2 = (1+\alpha_2 P_2)$. Accordingly, the problem in (2.57) can be written as

$$\begin{aligned} \text{find} \quad & \lambda && \text{(2.59a)} \\ \text{s.t.} \quad & \frac{\big(\phi_2^{(1-\lambda)}(\phi_1^{\lambda}-1)\big)/\alpha_1 + \big(\phi_1^{\lambda}(\phi_2^{(1-\lambda)}-1)\big)/\alpha_2}{\phi_1^{\lambda}+\phi_2^{(1-\lambda)}-\phi_1^{\lambda}\phi_2^{(1-\lambda)}} \leq \lambda P_1 + (1-\lambda) P_2, && \text{(2.59b)} \\ & 0 \leq \lambda \leq 1. && \text{(2.59c)} \end{aligned}$$

It can be seen that the right hand side of the constraint in (2.59b) is a linear function of λ, and that for the values $\lambda = 0$ and $\lambda = 1$ the constraint holds with equality. To show that for values of λ that lie within the interval $(0, 1)$ the constraint in (2.59b)

cannot be satisfied, we write the first derivative of the left hand side function of (2.59b) as

$$\frac{d}{d\lambda} = \frac{\phi_1^{\lambda}\phi_2^{(1-\lambda)}}{(\phi_1^{\lambda}+\phi_2^{(1-\lambda)}-\phi_1^{\lambda}\phi_2^{(1-\lambda)})^2} \times \Big((\ln\phi_1+\ln\phi_2)(\tfrac{1}{\alpha_1}-\tfrac{1}{\alpha_2}) - \tfrac{\ln\phi_2}{\alpha_1}\phi_1^{\lambda} + \tfrac{\ln\phi_1}{\alpha_2}\phi_2^{(1-\lambda)}\Big), \quad (2.60)$$

which has a positive value for $\lambda = 0$ and a negative value for $\lambda = 1$. In addition, there is only one value of λ for which the derivative is equal to zero. This is because the function $f(\lambda) = \frac{\ln\phi_2}{\alpha_1}\phi_1^{\lambda} - \frac{\ln\phi_1}{\alpha_2}\phi_2^{(1-\lambda)}$ is increasing in terms of λ and accordingly it has only one intersection with the constant value $(\ln\phi_1+\ln\phi_2)(\frac{1}{\alpha_1}-\frac{1}{\alpha_2})$. Based on these observations we can conclude that for any value of $\lambda \in (0,1)$ the function in the left hand side of (2.59b) is greater than the function in the right hand side. Hence, the objective function of TDMA has a value that is no larger than that of independent decoding if TDMA is feasible.

2.E Optimality of Suboptimal Methods for Equal Channel Gains

Consider a two-time-slot system that employs the full multiple access scheme in the first time slot which is of the length $\tau_1 = \tilde{L}_1$, and a three-time-slot system that employs independent decoding in the first time slot. For the latter system the first user has the opportunity to use the first and the second time slots to transmit its task to the access point, and hence $\tau_1 + \tau_2 = \tilde{L}_1$. It will suffice to assume that the total energy consumption and the number of transmitted bits in the last slot of both systems are equal, and to show that we can find a set of rate and power allocations

and the slot durations for the independent decoding scheme so that in its first two time slots it transmits the required bits with the same energy as the first slot of the optimal scheme. Our resource allocation for the independent decoding scheme will adopt a TDMA structure, and hence the result applies to TDMA, too.

Let R_k and P_k denote the parameters of the k^{th} user in the first slot of the full multiple access scheme and let R'_{ki} and P'_{ki} denote the corresponding parameters in the i^{th} time slot of the independent decoding scheme. In order to guarantee that the energy consumption and the transmitted bits in both schemes are equal, the independent decoding scheme must satisfy

$$\tilde{L}_1 R_1 = \tau_1 R'_{11} + (\tilde{L}_1 - \tau_1) R'_{12}, \tag{2.61a}$$

$$\tilde{L}_1 R_2 = \tau_1 R'_{21}, \tag{2.61b}$$

$$\tilde{L}_1 (P_1 + P_2) = \tau_1 (P'_{11} + P'_{21}) + (\tilde{L}_1 - \tau_1) P'_{12}. \tag{2.61c}$$

Using the closed-form solutions obtained in Sections 2.5 and 2.6.3, we can rewrite (2.61c) for equal channel gains as

$$\tfrac{\tilde{L}_1}{\alpha} 2^{R_1+R_2} = \tfrac{\tau_1}{\alpha} \big(\tfrac{2^{R'_{11}+R'_{21}}}{2^{R'_{11}}+2^{R'_{21}}-2^{R'_{11}+R'_{21}}} \big) + \tfrac{\tilde{L}_1-\tau_1}{\alpha} 2^{R'_{12}}, \tag{2.62}$$

where $\alpha = \alpha_1 = \alpha_2$. It will suffice to set $R'_{11} = 0$, so the independent decoding scheme adopts a TDMA structure. In that case, (2.62) can be written as

$$\tfrac{\tilde{L}_1}{\alpha} 2^{R_1+R_2} = \tfrac{\tau_1}{\alpha} 2^{R'_{21}} + \tfrac{\tilde{L}_1-\tau_1}{\alpha} 2^{R'_{12}}. \tag{2.63}$$

From (2.61a) and (2.61b), we have that $R_1 + R_2 = \frac{\tau_1}{\tilde{L}_1} R'_{21} + \frac{\tilde{L}_1-\tau_1}{\tilde{L}_1} R'_{12}$. Accordingly,

(2.63) takes the form

$$2^{(\lambda R'_{21}+(1-\lambda)R'_{12})} = \lambda 2^{R'_{21}} + (1-\lambda)2^{R'_{12}}, \tag{2.64}$$

where $\lambda = \tau_1/\tilde{L}_1$. Since the exponential function is strictly convex, if $\tau_1 \in (0, \tilde{L}_1)$ the equality in (2.64) holds if and only if $R'_{12} = R'_{21}$. In that case, using (2.63), $R'_{12} = R'_{21} = R_1 + R_2$, and from (2.61b) $\tau_1 = \tilde{L}_1 R_2/(R_1 + R_2)$. If the power constraints satisfy $\bar{P}_1, \bar{P}_2 \geq (2^{R_1+R_2} - 1)/\alpha$, the rates R'_{12} and R'_{21} are achievable and TDMA (and hence independent decoding) can achieve the minimum energy consumption.

2.F Quasi-Convexity of the Objective Function in (2.44)

A function f with a scalar argument is quasi-convex if at least one of the following conditions holds (Boyd and Vandenberghe, 2004),

(a) f is non-increasing,

(b) f is non-decreasing,

(c) there is a (turning) point, c, such that for any $x \leq c$ the function $f(x)$ is non-increasing and for any $x \geq c$ the function $f(x)$ is non-decreasing.

Here, we will show that when R_{21} and R_{22} are held constant and the objective function in (2.44) is viewed as a function of R_{11}, either the condition (b) or the condition (c) holds. An analogous approach can be employed to prove the quasi-convexity with respect to R_{21} and R_{22}.

For a given pair (R_{21}, R_{22}), the derivative of the objective with respect to R_{11} can be written as

$$\frac{T_s \bar{L}_1}{(T_s+\delta_c R_{11})^2} \times F_r, \tag{2.65}$$

where

$$\begin{aligned} F_r = \; & -\delta_c\big(\tfrac{2^{R_{11}}-1}{\alpha_1} + \tfrac{2^{R_{11}}(2^{R_{21}}-1)}{\alpha_2}\big) + (T_s+\delta_c R_{11})\big(\tfrac{2^{R_{11}}\ln 2}{\alpha_1} + \tfrac{2^{R_{11}}\ln 2(2^{R_{21}}-1)}{\alpha_2}\big) \\ & + \big(\tfrac{\delta_c(T_s+\delta_c R_{21})}{T_s+\delta_c R_{23}}\big)\tfrac{2^{R_{23}}-1}{\alpha_2} - \mathcal{F}_1'\big(B_1 - \tfrac{\bar{L}_1 R_{11}}{T_s+\delta_c R_{11}}\big) \\ & + \mathcal{F}_2'\big(B_2 - \tfrac{\bar{L}_2 R_{23}}{T_s+\delta_c R_{23}} - \tfrac{\bar{L}_1 T_s(R_{21}-R_{23})}{(T_s+\delta_c R_{11})(T_s+\delta_c R_{23})}\big) \times \big(\tfrac{\delta_c(R_{21}-R_{23})}{T_s+\delta_c R_{23}}\big). \end{aligned}$$

As $\frac{T_s \bar{L}_1}{(T_s+\delta_c R_{11})^2}$ is always positive, to show that either condition (b) or condition (c) holds, it is sufficient to show that F_r is non-decreasing. In order to show that, we will show that the derivative of F_r with respect to R_{11} is always non-negative. The derivative is

$$\begin{aligned} \tfrac{dF_r}{dR_{11}} = \; & (T_s+\delta_c R_{11})\big(\tfrac{2^{R_{11}}(\ln 2)^2}{\alpha_1} + \tfrac{2^{R_{11}}(\ln 2)^2(2^{R_{21}}-1)}{\alpha_2}\big) + \mathcal{F}_1''\big(B_1 - \tfrac{\bar{L}_1 R_{11}}{T_s+\delta_c R_{11}}\big)\big(\tfrac{\bar{L}_1 T_s}{(T_s+\delta_c R_{11})^2}\big) \\ & + \mathcal{F}_2''\big(B_2 - \tfrac{\bar{L}_2 R_{23}}{T_s+\delta_c R_{23}} - \tfrac{\bar{L}_1 T_s(R_{21}-R_{23})}{(T_s+\delta_c R_{11})(T_s+\delta_c R_{23})}\big) \times \big(\tfrac{\delta_c(R_{21}-R_{23})}{T_s+\delta_c R_{23}}\big)^2\big(\tfrac{\bar{L}_1 T_s}{(T_s+\delta_c R_{11})^2}\big). \end{aligned}$$

Since $\mathcal{F}_k(\cdot)$ in (3.6) is a convex function for non-negative arguments, its second derivative is non-negative. Accordingly, $\frac{dF_r}{dR_{11}}$ is non-negative, and hence F_r is non-decreasing.

Chapter 3

Uplink Resource Allocation for Multiple Access Computational Offloading

3.1 Abstract

The mobile edge computing framework offers the opportunity to reduce the energy that devices must expend to complete computational tasks. The extent of that energy reduction depends on the nature of the tasks, and on the choice of the multiple access scheme. In this chapter, we first address the uplink communication resource allocation for offloading systems that exploit the full capabilities of the multiple access channel (FullMA). For indivisible tasks we provide a closed-form optimal solution of the energy minimization problem when a given set of users with different latency constraints are offloading, and a tailored greedy search algorithm for finding a good set of offloading users. For divisible tasks we develop a low-complexity algorithm to

find a stationary solution. To highlight the impact of the choice of multiple access scheme, we also consider the TDMA scheme, which, in general, cannot exploit the full capabilities of the channel, and we develop low-complexity optimal resource allocation algorithms for indivisible and divisible tasks under that scheme. The energy reduction facilitated by FullMA is illustrated in our numerical experiments. Further, those results show that the proposed algorithms outperform existing algorithms in terms of energy consumption and computational cost.

3.2 Introduction

The rapid development of mobile device technology and wireless communication networks is bringing the vision of ubiquitous computing to fruition, at least for tasks of modest complexity. However, as the demand for ubiquity in computationally-intensive and latency-sensitive tasks increases, the limited computation, memory and energy resources of mobile and other small scale devices present significant impediments to progress. The Mobile Edge Computing (MEC) framework seeks to address these impediments by offering the devices the opportunity to offload (a portion of) their computational tasks to a local shared computational resource. This offloading option enables the users to execute more computationally complex applications within a certain deadline, and can also prolong the battery lifetime of the devices (Mao *et al.*, 2017b; Hu *et al.*, 2015; Mach and Becvar, 2017).

In order to exploit the opportunities provided by the MEC framework, a computational offloading system must address a number of challenges, including the energy that each user would expend to offload (a portion of) its computational task to the access point (Barbera *et al.*, 2013), the latency requirements of the tasks (Lei *et al.*,

2013), contention for the limited communication resources (Mao *et al.*, 2017b), and, in some cases, contention for the limited shared computation resources at the access point (Liu *et al.*, 2013). In order to address those challenges, the available resources must be effectively allocated to the users. The resource allocation problem, which usually targets the energy consumption of the users while ensuring that the latency constraints of the tasks are met, can be formulated as a joint optimization problem over the available communication and computation resources (Sardellitti *et al.*, 2015; Muñoz *et al.*, 2015, 2014; Salmani and Davidson, 2016; Wang *et al.*, 2017a, 2018a; Chen *et al.*, 2018; Salmani and Davidson, 2018b, and Chapter 2). The predominant factor in determining the structure of that optimization problem is the nature of the users' computational tasks. Two classes of tasks that are widely considered in the literature are "indivisible tasks" and "divisible tasks" (Khan *et al.*, 2014; Mach and Becvar, 2017). A task is indivisible if its components are tightly coupled. Such a task must be either completely offloaded or executed locally (e.g., Kumar and Lu, 2010; Wu *et al.*, 2013; Sardellitti *et al.*, 2015). On the other hand, a divisible task has independent or loosely coupled components, and can be partitioned. Hence, the mobile device can benefit from the implicit parallelism between the access point and the device by offloading a portion of the task while the remainder is executed locally (e.g., Zhang *et al.*, 2013; Wang *et al.*, 2016). Accordingly, the resource allocation problem is structurally different in the cases of binary offloading (for indivisible tasks) and partial offloading (for divisible tasks). We will address the cases of indivisible and data-partitionable divisible (Wang *et al.*, 2016) tasks in this thesis.

In a multi-user offloading system, irrespective of whether it is binary offloading or

partial offloading, the choice of the multiple access scheme can have a significant impact on the energy consumption, especially when the latency constraints are tight. In most of the previous work on such systems, the multiple access schemes employed by the system have been restricted to schemes that are simple, but are unable to exploit the full capabilities of the channel. Those schemes include Time Division Multiple Access (TDMA) and (Orthogonal) Frequency Division Multiple Access ((O)FDMA), which avoid interference by allocating orthogonal channels to the users, in time and frequency, respectively, and independent decoding, in which the receiver treats interference as noise. For example, Sardellitti *et al.* (2015) and Wang *et al.* (2018a) consider the energy minimization problem in a TDMA-based multi-user offloading system is considered for binary and partial offloading cases, respectively. An FDMA-based partial offloading system is considered by Mao *et al.* (2017a). The energy minimization problems for TDMA and OFDMA-based multi-user partial offloading systems were addressed by You *et al.* (2017), and the corresponding problem for independent decoding was addressed by Sardellitti *et al.* (2015). All of those multiple access schemes limit the range of the rates at which the users can operate reliably, and hence the optimal energy consumption cannot be obtained.

To address those limitations, the main focus of this work is to find the optimal user energy consumption of a K-user (binary or partial) offloading system that employs a multiple access scheme that exploits the full capabilities of the multiple access channel. That is, a scheme that enables reliable operation at rates that approach the boundary of the capacity region. Examples of such schemes include Gaussian signalling with joint decoding, and Gaussian signalling with optimally-ordered sequential decoding and time sharing (Cover and Thomas, 2012; El-Gamal and Cover,

1980). For simplicity, we will refer to any such scheme as a "full" multiple access (FullMA) scheme. For any FullMA scheme, we will provide efficient algorithms for optimally allocating the available communication resources to the K users, each of which wishes to complete its computational task within its own specific deadline, either locally or by offloading (a portion of) the task to an access point that has substantial computation resources. We will consider this problem for both indivisible and divisible computational tasks. For the indivisible case, the combinatorial structure of the binary offloading problem of deciding which users will offload their tasks and which will complete them locally suggests a natural decomposition into an outer search strategy for the offloading decisions and the inner optimization of the communication resources for given offloading decisions. That inner subproblem will be referred to as the "complete offloading" problem. For the case of partial offloading of data-partitionable divisible tasks (Wang *et al.*, 2016), the fraction of each task to be offloaded will be optimized jointly with the communication resource allocation.

Our strategy for solving the resource allocation problem for systems with a FullMA scheme is based on the insights developed in the two-user case in Chapter 2 (see also Salmani and Davidson, 2016, 2018b), which suggests algebraic decompositions of the problem. We will exploit the polymatroid structure of the capacity region of the multiple access channel (see Tse and Hanly, 1998) in both the complete offloading and partial offloading cases. In the complete offloading case, we will obtain closed-form optimal solution for the energy minimization problem. That solution also forms the core of a tailored greedy search algorithm for good solutions to the binary offloading problem. In the partial offloading case, our decomposition strategy enables us to obtain closed-form solutions for some of the design variables and to obtain a stationary

solution of the energy minimization problem by employing a simple coordinate descent algorithm over the K remaining variables.

To highlight the impact of the choice of the multiple access scheme on the energy consumption of an offloading system, we will also address the energy minimization problem for the TDMA scheme. We will show that that problem can be written as a jointly convex K-dimensional optimization problem. In our simulation results, we will show that although there are scenarios in which TDMA provides good performance, there are others in which exploiting the full capabilities of the multiple access channel enables a substantial energy consumption reduction.

A special case of our total energy minimization problem for a K-user offloading system with a full multiple access scheme has appeared (Wang *et al.*, 2018b). Wang *et al.* (2018b) assumed that the latency constraints of all the users are the same, while we consider the more general case in which different users have different latency requirements and we exploit the maximum allowable latency constraint of each user to reduce the user energy consumption. Wang *et al.* (2018b) obtain solutions for the transmission rates, transmission powers, and the fraction of offloaded bits in the partial offloading scenario iteratively using a variant of the ellipsoidal algorithm. For a K-user system, that algorithm imposes a computational cost of $O(K^3)$ operations per iteration. In contrast, the closed-form optimal solutions for the transmission powers and the fraction of offloaded bits provided by the decomposition-based approach developed herein result in an algorithm whose computational cost is only $O(K \log K)$. Since the class of scenarios for which the proposed algorithm is developed includes the scenario of equal latencies for which Wang *et al.* (2018b) developed their algorithm, our algorithm has the same performance as that of Wang *et al.* (2018b) in the equal

latency case. However, our numerical results, and those of Wang *et al.* (2018b), show that the number of iterations required by the corresponding algorithm of Wang *et al.* (2018b) can be quite large. As a result, in the case of a single-antenna access point the proposed algorithm has a significant computational advantage.

An analogous equal-latency assumption has also been considered for the TDMA scheme by You *et al.* (2017). An additional difference between the problem considered by You *et al.* (2017) and that proposed herein is that we have considered the dynamic voltage scaling approach (Wang *et al.*, 2016) for computation energy management in the mobile devices. This approach guarantees the minimum local energy consumption in the users subject to the latency constraints. In our numerical results, we will show that the energy consumption of the problem formulation proposed for the TDMA case in this chapter is significantly lower than the energy consumption of the problem formulation of You *et al.* (2017).

3.3 System Model

We will consider a system consisting of K single-antenna users, each of which has a computational task that is to be executed within its own specific latency constraint, and an access point that is equipped with sufficiently large computational resources that the offloaded tasks can be processed without contention. The offloading users are served over a single time slot by the single-antenna (coherent) receiver at the access point, and the channels between the users and the access point are assumed to be frequency-flat and quasi-static. We will adopt the conventional discrete-time baseband equivalent model with symbol interval T_s. Therefore, if $s_k[n]$ denotes the transmitted signal by the k^{th} user at the n^{th} channel use, and if h_k denotes the channel

from the k^{th} user to the access point, then the signal received at the access point at the n^{th} channel use is

$$y[n] = \sum_{k=1}^{K} h_k s_k[n] + v[n], \tag{3.1}$$

where $v[n]$ is an additive circular zero mean white Gaussian noise of variance σ^2.

In order to explore the impact of the multiple access scheme on the energy consumption of the offloading devices, we will tackle the following generic energy minimization problem

$$\begin{aligned} \min_{\substack{\text{offloading fraction,} \\ \text{communication resources}}} \quad & \text{Total device energy consumption} && \text{(3.2a)} \\ \text{s.t.} \quad & \text{Offloading fraction constraints,} && \text{(3.2b)} \\ & \text{Latency constraints,} && \text{(3.2c)} \\ & \text{Achievable rate region constraints.} && \text{(3.2d)} \end{aligned}$$

The constraints on the fraction of the computational task that is offloaded by each user are determined by the nature of the tasks that the users seek to offload. If the tasks are indivisible, the offloading fraction for each user is either zero (local execution) or one (complete offloading of the task). Alternatively, if the tasks are data-partitionable divisible tasks, in which a simple-to-describe operation is applied, independently, to different blocks of data (Wang *et al.*, 2016), the offloading fraction can be modeled as taking any value in $[0, 1]$. Regarding the constraints on the latencies, we will consider the general case in which each user has its specific latency constraint, independent from the latencies of other users. Finally, the achievable rate region describes the set of rates at which reliable communication can be achieved for a given set of transmission

powers (e.g., Cover and Thomas, 2012). Different multiple access schemes manage the interference between users in different ways and hence have different achievable rate regions. The capacity region is the convex hull of all achievable rate regions and we will call any multiple access scheme that can operate reliably at all points in the capacity region a "full" multiple access scheme.

Now, in order to formulate the generic energy minimization problem, let R_k and P_k denote the data rate and power (in units per channel use) employed by user k when it is transmitting, respectively. In addition, let $\{R_k\}_{k=1}^K$ and $\{P_k\}_{k=1}^K$ denote the sets of transmission rates and transmission powers for all users, respectively. In some cases we will simplify that notation to $\{R_k\}$ and $\{P_k\}$. We will use the generic notation $\mathcal{R}\left(\{P_k\}_{k=1}^K\right)$ to denote the achievable rate region of a multiple access scheme, and hence the rate region constraint can be written as $\{R_k\}_{k=1}^K \in \mathcal{R}\left(\{P_k\}_{k=1}^K\right)$, (e.g., Cover and Thomas, 2012; El-Gamal and Cover, 1980; El-Gamal and Kim, 2011). In specifying that constraint for a particular multiple access scheme, we will assume that the data blocks are long enough for the asymptotic characterization to be valid. Under the asymptotic assumption, the achievable rate region of a FullMA scheme is the capacity region (see (3.10) below), and for the TDMA scheme, since each user transmits in a different interval, the rate R_k at which it can reliably communicate during that interval is upper-bounded by the classical single-user capacity expression, (e.g., Cover and Thomas, 2012).[1]

If B_k denotes the total number of bits describing the task of user k, then let $\gamma_k B_k$ define the number of bits offloaded by the k^{th} user, where $\gamma_k \in \{0, 1\}$ for the binary offloading case, and $\gamma_k \in [0, 1]$ for the partial offloading case. Accordingly, the time

[1]Extensions to rate regions for finite block lengths (e.g., Polyanskiy *et al.*, 2010; MolavianJazi and Laneman, 2012) will be guided by the insight developed herein.

it takes for user k to offload (the portion of) its task is $t_{\mathrm{UL}_k} = T_s \frac{\gamma_k B_k}{R_k}$. The energy that it expends in doing so is $\frac{\gamma_k B_k}{R_k} P_k$.

In order to satisfy the latency constraints in (3.2c), both the offloaded portion of each user's task and the locally retained portion must be completed within that user's specified latency. To formulate those constraints, we observe that the structure of data-partitionable tasks is such that the time that it takes for the access point to process the offloaded portion can be modeled as a simple multiple of its size (Zhang *et al.*, 2013),

$$t_{\mathrm{exe}_k} = \delta_c \gamma_k B_k, \tag{3.3}$$

where δ_c is the time it takes to process one bit at the access point. For indivisible tasks, $\gamma_k \in \{0, 1\}$, and we can use the expression in (3.3) if we scale δ_c so that $\delta_c B_k$ is equal to the time that it would take for the access point to complete the task.

The time that it takes for user k to communicate (a portion of) the problem to the access point is the sum of any time it has to wait until it can access the channel, t_{w_k}, and the actual offloading time t_{UL_k}. For FullMA schemes each user has immediate access to the channel and hence $t_{\mathrm{w}_k} = 0$, whereas for the TDMA scheme users have to wait until their turn; see Section 3.4.3. If the time it takes for the access point to send the results back to the k^{th} user is denoted by t_{DL_k}, then the latency constraint of that offloading user can be written as

$$t_{\mathrm{w}_k} + t_{\mathrm{UL}_k} + t_{\mathrm{exe}_k} + t_{\mathrm{DL}_k} \leq L_k, \tag{3.4}$$

in which L_k denotes the maximum allowable latency for user k. The time t_{DL_k} depends on a number of different factors, including the description length of the results of the

(partially) offloaded task, which is often considerably shorter than the description length of the task itself. It also depends on the downlink signalling scheme chosen by the access point, and the energy that the access point expends on the downlink. Since our emphasis is on the minimization of the energy expended by the devices (and not the access point) through the selection of a multiple access scheme for the uplink and the corresponding resource allocation, we will model t_{DL_k} as a (possibly different) constant for each user.

The local execution time takes a similar form to that in (3.3) when the users employ a conventional computational architecture. Hence, a local latency constraint for data-partitionable tasks takes the form $t_{\mathrm{loc}_k} = \delta_k(1-\gamma_k)B_k \leq L_k$, where δ_k is the time it takes for the k^{th} user to process one bit. A scaling analogous to that after (3.3) can be used for the binary offloading case.

To complete the generic formulation, we will let $E_{\mathrm{loc}_k}(\gamma_k)$ denote the energy that user k expends to complete its local computation within its latency constraint. That energy depends on the number of operations that the local processor must perform to complete (the retained portion of) the user's task, and on the energy required to perform each operation. As discussed after (3.5), the latter depends on the nature of the computational architecture of the device. For an indivisible task, the number of local operations is either zero (when the task is fully offloaded), or a constant (when the task is locally executed). That constant is determined by the complexity of the task. For data-partitionable divisible task, the number of local operations can be modeled as being proportional to the fraction of the description that user k retains (Wang *et al.*, 2016).

Having developed this notation, the generic problem of minimizing the user energy

consumption of a system with K offloading users, which was described in (3.2), can be formulated as

$$\min_{\{R_k\},\{P_k\},\{\gamma_k\}} \sum_k \frac{\gamma_k B_k}{R_k} P_k + E_{\mathrm{loc}_k}(\gamma_k) \tag{3.5a}$$

$$\text{s.t.} \quad \gamma_k \in \{0,1\} \text{ or } \gamma_k \in [0,1], \quad \forall k, \tag{3.5b}$$

$$t_{\mathrm{w}_k} + T_s\big(\tfrac{\gamma_k B_k}{R_k}\big) + \delta_c \gamma_k B_k + t_{\mathrm{DL}_k} \leq L_k, \quad \forall k, \tag{3.5c}$$

$$\delta_k (1-\gamma_k) B_k \leq L_k, \quad \forall k, \tag{3.5d}$$

$$0 \leq P_k, \quad \forall k, \tag{3.5e}$$

$$\{R_k\}_{k=1}^K \in \mathcal{R}\left(\{P_k\}_{k=1}^K\right), \tag{3.5f}$$

where the constraints in (3.5b) are the offloading fraction constraints for binary or partial offloading, respectively, (3.5c) and (3.5d) capture the latency constraints on the offloaded and locally-executed portions of the task, and (3.5f) is the rate region constraint for the chosen multiple access scheme.

Our primary algorithm development for the solution of (3.5) will be tailored to devices with the dynamic voltage scaling computational architecture (Wang *et al.*, 2016). That architecture enables the device to adjust its CPU frequency and hence to minimize the energy it requires to complete (the local portion of) its task within the specified latency constraint. Since in that architecture the local latency constraint in (3.5d) is implicitly satisfied, it can be removed from (3.5). For a data-partitionable task, the minimized local computational energy can be expressed in the form (Wang *et al.*, 2016)

$$E_{\mathrm{loc}_k}(\gamma_k) = \tfrac{M_k}{L_k^2}\big((1-\gamma_k)B_k\big)^3, \tag{3.6}$$

where the coefficient M_k depends on the characteristics of the chip of user k. For the case of binary offloading with dynamic voltage scaling architecture, we will denote the minimized local energy computation by $\underline{E}_{\text{loc}_k}$, i.e.,

$$E_{\text{loc}_k}(0) = \underline{E}_{\text{loc}_k} \text{and } E_{\text{loc}_k}(1) = 0. \tag{3.7}$$

In Sections 3.4 and 3.5 we will focus on the development of algorithms for users that employ dynamic voltage scaling in the binary and partial offloading scenarios, respectively. However, with simple modifications the proposed algorithms can be applied to users with conventional computation architectures. The required modifications in the binary case are discussed at the end of Section 3.4, and the modifications for the case of partial offloading were illustrated for a two-user system in Chapter 2 (see also Salmani and Davidson, 2017c). In our numerical results in Section 3.6, we will illustrate that dynamic voltage scaling approach provides significant energy savings.

As mentioned in the Introduction, the problem in (3.5) is different from those of You *et al.* (2017) and Wang *et al.* (2018b). We allow the latency constraints of the users, L_k, to be different, which enables the users with larger latencies to benefit from their own available time to transmit. You *et al.* (2017) and Wang *et al.* (2018b) assume that the latency constraints of the users are the same, which forces the system to work with the minimum latency constraint among the users. If the latencies are different, doing that will increase the total energy consumption. In addition, for the partial offloading case, the formulations of You *et al.* (2017) and Wang *et al.* (2018b) assume that δ_c is small enough that the dependence of the execution time at the access point, t_{exe_k}, on the fraction of the task that is offloaded, γ_k, can be neglected.

We do not make that assumption in our formulations; see (3.3). Finally, in contrast to You *et al.* (2017), in our formulation we assume that the users can employ dynamic voltage scaling (Zhang *et al.*, 2013; Wang *et al.*, 2016) to minimize the energy that they expend in local computation.

The rest of this chapter addresses the energy minimization problem in (3.5) for two classes of computational tasks, namely indivisible tasks and data-partitionable divisible tasks, under two different multiple access schemes, namely FullMA and TDMA. In particular, we will consider that problem in the binary offloading case (for indivisible tasks) under FullMA in Section 3.4.2 and under the TDMA scheme in Section 3.4.3. We will tackle the energy minimization problem for a partial offloading system (for data-partitionable divisible tasks) under FullMA in Section 3.5.1, and under the TDMA scheme in Section 3.5.2.

3.4 Binary Offloading

In this section we will consider minimizing the total energy consumption of the K-user system when the computational tasks of the users are indivisible, i.e., the task of each user must be either totally offloaded to the access point or executed by the user. Since the offloading decision is binary, the problem of finding the optimal selection of offloading users that minimizes the total energy consumption is combinatorial. As a result, the joint offloading-decision and resource-allocation problem is typically partitioned, with the optimal resource allocation being found for given offloading decisions and a combinatorial search strategy being used to make the offloading decisions. Accordingly, in this section we first seek the optimal solution of energy minimization problem for the complete offloading case in which a subset of users is scheduled to

offload their tasks; see Section 3.4.1. Then, in Section 3.4.4, we will develop a low-complexity pruned greedy search technique that is tailored to the characteristics of the problem to find a set of offloading users that typically results in close-to-optimal energy consumption.

3.4.1 Complete Computation Offloading

Let $\mathcal{S} = \{1, 2, \ldots, K\}$ denote the set of all K users in the system and let $\mathcal{S}' \subseteq \mathcal{S}$, where $|\mathcal{S}'| = K'$, denote the subset of users scheduled to fully offload their tasks, i.e., $\gamma_k = 1, \forall k \in \mathcal{S}'$ and $\gamma_k = 0, \forall k \notin \mathcal{S}'$. (As mentioned above, the selection of $\mathcal{S}'$ is discussed in Section 3.4.4.) In that case, the total device energy consumption consists of the sum of the transmission energies of the users in $\mathcal{S}'$ and the sum of the local computational energies of the remaining users. The latter term can be minimized (while satisfying the latency constraint) by employing optimized dynamic voltage scaling (Wang *et al.*, 2016), which leads to the following expression for the total device energy:

$$E_{\text{total}} = \sum_{k \in \mathcal{S}'} \frac{B_k}{R_k} P_k + \sum_{j \in \mathcal{S} \setminus \mathcal{S}'} \underline{E}_{\text{loc}_j}. \tag{3.8}$$

Thus, the problem that remains is to minimize the energy consumed by the offloading devices

$$\min_{\{R_k\},\{P_k\}} \sum_{k\in\mathcal{S}'} \frac{B_k}{R_k} P_k \tag{3.9a}$$

$$\text{s.t.} \quad t_{\mathrm{w}_k} + T_s\big(\tfrac{\gamma_k B_k}{R_k}\big) + \delta_c \gamma_k B_k + t_{\mathrm{DL}_k} \leq L_k, \forall k \in \mathcal{S}', \tag{3.9b}$$

$$0 \leq P_k, \quad \forall k \in \mathcal{S}', \tag{3.9c}$$

$$\{R_k\}_{k=1}^{K'} \in \mathcal{R}\big(\{P_k\}_{k=1}^{K'}\big). \tag{3.9d}$$

As discussed in Section 3.3, the achievable rate region, $\mathcal{R}$, and the waiting time, t_{w_k}, depend on the chosen multiple access scheme. In the following sections, we will provide solutions to (3.9) for a FullMA scheme and for the TDMA scheme.

3.4.2 Full Multiple Access Scheme

For a FullMA scheme, the achievable rate region is the capacity region of the multiple access channel. Since there are K' users in $\mathcal{S}'$, that region can be described by the K' constraints of the form $0 \leq R_k$ and the $(2^{K'} - 1)$ constraints of the form (Cover and Thomas, 2012; El-Gamal and Cover, 1980)

$$\sum_{i\in\mathcal{N}} R_i \leq \log\big(1 + \sum_{i\in\mathcal{N}} \alpha_i P_i\big), \tag{3.10}$$

in which $\alpha_i = \frac{|h_i|^2}{\sigma^2}$ and $\mathcal{N} \subseteq \mathcal{S}'$. Furthermore, in a FullMA scheme the available channel is simultaneously assigned to all users, and hence $t_{\mathrm{w}_k} = 0, \forall k$. Therefore,

for a FullMA scheme, the problem in (3.9) becomes

$$\min_{\{R_k\},\{P_k\}} \sum_{k\in\mathcal{S}'} \frac{B_k}{R_k} P_k \tag{3.11a}$$

$$\text{s.t.} \quad T_s\big(\tfrac{B_k}{R_k}\big) \leq \tilde{L}_k, \quad \forall k \in \mathcal{S}' \tag{3.11b}$$

$$0 \leq R_k, \quad \forall k \in \mathcal{S}' \tag{3.11c}$$

$$2^{\sum_{i\in\mathcal{N}} R_i} \leq 1 + \sum_{i\in\mathcal{N}} \alpha_i P_i, \quad \forall \mathcal{N} \subseteq \mathcal{S}', \tag{3.11d}$$

where $\tilde{L}_k = L_k - \delta_c \gamma_k B_k - t_{\mathrm{DL}_k}$.

As the first step toward solving the problem in (3.11), we decompose the problem into an inner optimization over the transmission powers and an outer optimization over the rates:

$$\begin{array}{ll} \min\limits_{\{R_k\}} & \min\limits_{\{P_k\}} \ \sum_{k\in\mathcal{S}'} \frac{B_k}{R_k} P_k \\ \text{s.t.} \quad \text{(3.11b)} - \text{(3.11d)}, & \text{s.t.} \quad \text{(3.11d)}. \end{array} \tag{3.12}$$

For a fixed set of rates $\{R_k\}$, the inner optimization problem is a linear programme in $\{P_k\}$ and the feasibility region for the transmission powers is a polyhedron. Hence, in the search for an optimal solution it is sufficient to restrict attention to the vertices of the feasibility region. Each vertex is described by the simultaneous satisfaction of K' of the linear inequality constraints in (3.11d) with equality. As we show in the next section, by exploiting the polymatroid structure of the constraints in (3.11d) (e.g., Tse and Hanly, 1998), we can significantly reduce the number of the candidate vertices. In fact, we will show that we can find a closed-form optimal solution for the powers.

Closed-form optimal solutions for the powers

To begin, let us group the rate region constraints in (3.11d) into K' classes, where a constraint is assigned to class-ℓ if it involves the powers and rates of ℓ users; i.e., the constraint is assigned to class ℓ if $|\mathcal{N}| = \ell$. In Appendix 3.A we show that the vertices of the rate region that are candidates for optimality arise from the simultaneous satisfaction of at most one constraint from each of the classes. Since such vertices involve the simultaneous satisfaction of K' constraints, that implies that at optimality one constraint from each class holds with equality (see also Salmani and Davidson, 2018a).

Since class-K' contains only one constraint, that implies that at optimality

$$2^{\sum_{i=1}^{K'} R_i} = 1 + \sum_{i=1}^{K'} \alpha_i P_i. \tag{3.13}$$

Accordingly, the power of any arbitrary user, say user n, can be written in terms of the powers of the other users as

$$\alpha_n P_n = 2^{\sum_i R_i} - \sum_{i \neq n} \alpha_i P_i - 1. \tag{3.14}$$

By substituting this expression into (3.11a) and (3.11d), the inner optimization problem in (3.12) remains a linear programming problem, but now with $(K'-1)$ variables, namely,

$$\min_{\{P_k\}} \quad \sum\nolimits_{k \in \mathcal{S}' \setminus \{n\}} \big(\rho_k - \rho_n\big) \alpha_k P_k \tag{3.15a}$$

$$\text{s.t.} \quad 2^{\sum_{i \in \mathcal{N}} R_i} - 1 \leq \sum\nolimits_{i \in \mathcal{N}} \alpha_i P_i \leq 2^{R_n} \big(2^{\sum_{i \in \mathcal{N}} R_i} - 1\big), \quad \forall \mathcal{N} \subseteq \mathcal{S}' \setminus \{n\}, \tag{3.15b}$$

in which

$$\rho_k = \frac{B_k}{\alpha_k R_k}. \tag{3.16}$$

It can be seen that the constraints of the problem in (3.15) have a polymatroid structure, and hence the optimal solution results from simultaneous satisfaction of $(K'-1)$ constraints with at most one constraint from each class. For positive coefficients of the powers in the objective function in (3.15a) it can be shown that, analogous to (3.13), at optimality the single lower bound constraint in class-$(K'-1)$ is satisfied with equality. Accordingly, we can obtain a closed-form solution for the power of another arbitrary user by using an expression analogous to (3.14).

Based on the above discussion, we can obtain a sequence of closed-form solutions for all the powers for a given set of transmission rates if we can guarantee that in each step all the coefficients $\rho_k - \rho_n$ are positive. We can do that if we determine the permutation π so that

$$\rho_{\pi(1)} \geq \rho_{\pi(2)} \geq \cdots \geq \rho_{\pi(K'-1)} \geq \rho_{\pi(K')}, \tag{3.17}$$

and if we let $\pi^{-1}(\cdot)$ denote the inverse permutation; that is for user n there are $\pi^{-1}(n)-1$ users with values of ρ_k that are larger than ρ_n. Once the ordering in (3.17) has been determined, the first step of the algorithm is to obtain the closed-form solution for $P_{\pi^{-1}(K')}$ by substituting the expression in (3.14) with $n = \pi^{-1}(K')$ into (3.13); that is,

$$\alpha_{\pi^{-1}(K')} P_{\pi^{-1}(K')} = 2^{\sum_{i=1}^{\pi^{-1}(K')} R_i} - \sum_{i=1}^{\pi^{-1}(K')-1} \alpha_i P_i - 1. \tag{3.18}$$

The same procedure can then be applied in a sequential manner to find closed-form

solutions for all the powers. In the last step, we obtain

$$P_{\pi^{-1}(1)} = (2^{R_{\pi^{-1}(1)}} - 1)/\alpha_{\pi^{-1}(1)}. \tag{3.19}$$

This expression is only a function of this user's rate and channel, and does not depend on the powers of the other users. By retracing our steps, we obtain a closed-form solution for the optimal power of each user in terms of the rates of other users rather than their powers; i.e.,

$$P_k = \big(\tfrac{2^{R_k}-1}{\alpha_k}\big)2^{\sum_{j=1}^{\pi^{-1}(k)-1} R_j}. \tag{3.20}$$

We observe that the ordering in (3.17) not only ensures that the terms $(\rho_k - \rho_n)$ in (3.15a), and the corresponding terms in the subsequent instances of (3.17), are positive, it also determines the (optimal) decoding order that enables the rates that will be chosen in (3.22) below to be achieved by successive decoding. (Since these rates correspond to vertices of the capacity region, no time sharing is required.) In particular, it can be seen from (3.19) that the message from user $\pi(1)$ is being decoded after the messages from all other offloading users have been decoded and the corresponding interference canceled. Similarly, the expression in (3.18) reveals that the message from user $\pi(K')$ is the first message to be decoded, with the interference from the messages from the other users being treated as noise.

Closed-form optimal solutions for the rates

Now that we have the closed-form solutions for the transmission powers in (3.20), the outer optimization problem in (3.12) becomes

$$\min_{\{R_k\}} \quad \sum_{k\in\mathcal{S}'} \frac{B_k}{R_k}\left(\frac{2^{R_k}-1}{\alpha_k}\right)2^{\sum_{j=1}^{\pi^{-1}(k)-1} R_j} \tag{3.21a}$$

$$\text{s.t.} \quad \left(\frac{T_s B_k}{\tilde{L}_k}\right) \leq R_k, \quad \forall k \in \mathcal{S}'. \tag{3.21b}$$

It can be shown that the objective function in (3.21) is an increasing function with respect to each transmission rate and that the constraints on the transmission rates are separable. Hence, the optimal rate for each user is the minimum feasible rate according to its latency constraint,

$$R_k = \frac{T_s B_k}{\tilde{L}_k}. \tag{3.22}$$

Since this expression depends only on the parameters of the problem, we can obtain the ρ_k's in (3.16). Once those ρ_k's have been sorted, the optimal solutions for the transmission powers can be found using (3.20). These steps are summarized in Algorithm 2. The computational efficiency of the algorithm is apparent from the fact that the number of operations required is dominated by the sorting procedure in Step 3, which requires $O(K' \log K')$ operations.

Algorithm 2 : The optimal solution to (3.11)

Input data: $\mathcal{S}'$, $\{B_k\}$, $\{\tilde{L}_k\}$, $\{\alpha_k\}$, and T_s.
Step 1: Calculate the optimal rates $\{R_k\}$ using (3.22).
Step 2: Calculate the values $\{\rho_k\}$ using (3.16).
Step 3: Order $\{\rho_k\}$ according to (3.17) to find the optimal permutation π.
Step 4: Calculate the optimal powers using (3.20).

3.4.3 Time Division Multiple Access

In this section we will tackle the total energy minimization problem of a system with K' (completely) offloading users when TDMA is employed as the multiple access scheme. In the TDMA scheme there is only one user offloading at a time. Hence, there is no interference and the rate that each user employs when it transmits is bounded by the single-user capacity (e.g., Cover and Thomas, 2012). However, since the devices are transmitting one at a time, the allowable latency of each user must include the time that the user spends waiting for the devices scheduled to transmit earlier to complete their transmission. Therefore, the natural transmission schedule is in the order of increasing values of the transmission latency $\tilde{L}_k$, which was defined after (3.11). Without loss of generality we can order the users so that $\tilde{L}_1 \leq \tilde{L}_2 \leq \cdots \leq \tilde{L}_{K'}$, and in that case the waiting time of user k can be written as $t_{\mathrm{w}_k} = \sum_{i=1}^{k-1} t_{\mathrm{UL}_i} = \sum_{i=1}^{k-1} T_s\big(\frac{B_i}{R_i}\big)$. The device energy minimization problem in the TDMA case can then be written as

$$\min_{\{R_k\},\{P_k\}} \sum_{k\in\mathcal{S}'} \frac{B_k}{R_k} P_k \tag{3.23a}$$

$$\text{s.t.} \quad \sum_{i=1}^{k} T_s\big(\tfrac{B_i}{R_i}\big) \leq \tilde{L}_k, \quad \forall k \in \mathcal{S}', \tag{3.23b}$$

$$0 \leq P_k, \quad \forall k \in \mathcal{S}', \tag{3.23c}$$

$$0 \leq R_k \leq \log_2(1+\alpha_k P_k), \quad \forall k \in \mathcal{S}'. \tag{3.23d}$$

For a fixed set of transmission rates $\{R_k\}$, the objective in (3.23a) is an increasing function of each transmission power P_k, and the constraints on the powers are separable (because TDMA avoids interference between the users). Hence, the optimal solution for the transmission power for user k is, simply, the minimum power required

to achieve its target transmission rate, namely,

$$P_k = \frac{2^{R_k}-1}{\alpha_k}. \tag{3.24}$$

The remaining problem can be written in terms of the transmission rates as follows

$$\min_{\{R_k\}} \quad \sum_{k\in\mathcal{S}'} \frac{B_k}{\alpha_k}\left(\frac{2^{R_k}-1}{R_k}\right) \tag{3.25a}$$

$$\text{s.t.} \quad \sum_{i=1}^{k} T_s\left(\frac{B_i}{R_i}\right) \leq \tilde{L}_k, \quad \forall k \in \mathcal{S}', \tag{3.25b}$$

$$0 \leq R_k, \quad \forall k \in \mathcal{S}'. \tag{3.25c}$$

It can be shown that the objective function in (3.25) is jointly convex in the transmission rates and hence, the optimal solution to (3.25) can be efficiently obtained. The optimal solution to (3.23) is then the concatenation of these rates and the corresponding powers in (3.24).

3.4.4 Binary Computational Offloading

Now that we have obtained a closed-form optimal resource allocation for a given set of offloading users in the case of the full multiple access scheme, and a quasi-closed-form solution based on a convex optimization problem with K' variables in the case of TDMA, we can tackle the "outer" problem of finding an optimal set of offloading users. This is a combinatorial problem, with a search space of 2^K possibilities, but it admits a tree structure. Therefore, in addition to the branch-and-bound algorithm for finding an optimal set of offloading users, the problem is amenable to a wide variety of lower-complexity tree-search algorithms that typically provide offloading sets with low energy consumption. As an example, we will develop a customized greedy search

technique in which the search tree is (deterministically) pruned at each iteration.

Greedy search algorithm

To describe the proposed algorithm, we let $\mathcal{S}'$ denote the set of users that have already been chosen for offloading, and let $\mathcal{U}$ denote the set of users for which a decision as to whether or not to offload has yet to be made. We initialize the algorithm with all the users in $\mathcal{U}$ and none in $\mathcal{S}'$. The key steps in each iteration of the algorithm are an exploratory step, a deterministic pruning step, and a greedy user selection step that selects the "best" user to add to the offloading set (if any remain after the pruning step). These steps are summarized in steps 3, 4, and 6 in Algorithm 3. In the exploration step, for each user in $\mathcal{U}$ we obtain the energy consumption of the system if that user were to be added to the set of offloading users. In the case of FullMA scheme that can be computed using the closed-form expression in Algorithm 2 and in the case of TDMA it can be found by solving the convex optimization problem in (3.25) and using the expression in (3.24). In the pruning step we remove from $\mathcal{U}$ all those users for whom the exploration step revealed that (at this iteration) offloading would incur more energy consumption than local computation. These users can be "safely" removed, because in subsequent iterations there will be more users offloading and hence the energy required by any individual user to offload their task does not decrease as the iterations progress. In the greedy user selection for offloading step we select the user for which offloading offers the greatest reduction in the energy consumption of the system.

To analyze the computational effort required by the algorithm, let $Q^{(i)}$ denote the cardinality of the set $\mathcal{U}$ at the beginning of the i^{th} iteration; i.e., at Step 3.

At each iteration of the algorithm, the exploration step involves the solution of $Q^{(i)}$ complete offloading problems (Algorithm 2 for full multiple access scheme or (3.25) then (3.24) for TDMA scheme). The combination of the pruning and greedy selection steps requires $Q^{(i)}$ comparisons. At iteration i, there are i users in $\mathcal{S}'$ and hence, in the full multiple access case the cost of each complete offloading problem in Step 3 is $O(i \log i)$. Hence, the computational cost of Algorithm 3 in the full multiple access case is dominated by a term that is $O\big(\sum_i Q^{(i)} i \log i\big)$. In the worst case, no users are pruned in Step 4, so $Q^{(i)} \leq K - i + 1$ and hence the computation cost is at most $O\big(\sum_{i=1}^{K} (K - i + 1) i \log i\big)$. A loose upper bound for the argument of that expression is $K^3 \log K$. In our numerical results in Section 3.6 we will show that the proposed search strategy produces solutions that typically provide near optimal energy consumption and that it does so at a low computational cost.

We remark that an alternative greedy-based algorithm to solve the energy minimization problem of a binary offloading system was developed by Wang *et al.* (2018b) for systems in which all the users have the same latency constraint. The greedy choice at each iteration in that algorithm is similar to that in Algorithm 3; i.e., at each iteration a user which results in the maximum reduction of the total energy consumption is added to the set of offloading users. However, as we will illustrate in Section 3.6 the computational cost of the greedy algorithm of Wang *et al.* (2018b) is significantly higher than that of Algorithm 3. This is mainly due to the fact that each component of the equivalent to Step 2 of Algorithm 3 involves solving a problem using the ellipsoid algorithm. Using analysis similar to that in the previous paragraph, that results in a computational cost that is $O(K^5)$. In contrast, Algorithm 3 solves the corresponding problems using the closed-form expressions in Algorithm 2. The analogous

Algorithm 3 : Binary Offloading Solution

Input data: values of $\{B_k\}$, $\{\tilde{L}_k\}$, $\{\alpha_k\}$, $\{\underline{E}_{\text{loc}_k}\}$, T_s.
Step 1: Set $\mathcal{U} = \{1, 2, \ldots, K\}$, $\mathcal{S}' = \emptyset$, $E_{\text{off}}^{(0)} = 0$, $i = 0$.
Step 2: Set $\mathcal{V} = \emptyset$ and $i \leftarrow i + 1$.
for each $k \in \mathcal{U}$ **do**
Obtain the energy consumption of the system when user k is added to the set of offloading users, $E_{\text{total}_k}^{(i)}$; i.e., perform Alg. 2, or solve (3.25) then (3.24), for $\mathcal{S}' \cup \{k\}$.
if $E_{\text{off}}^{(i-1)} + \underline{E}_{\text{loc}_k} \leq E_{\text{total}_k}^{(i)}$ **then**
Add user k to the set of users to be pruned; i.e., $\mathcal{V} \leftarrow \mathcal{V} \cup \{k\}$
end if
end for
Step 3: Prune the selected users from the tree; i.e., $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{V}$.
Step 4: If $\mathcal{U} = \emptyset$, terminate the algorithm.
Step 5: Select the "best" user by choosing $k^\star = \arg\max_{k \in \mathcal{U}} \big(E_{\text{off}}^{(i-1)} + \underline{E}_{\text{loc}_k} - E_{\text{total}_k}^{(i)}\big)$.
Step 6: Update the offloading set and the undecided set; i.e., $\mathcal{S}' \leftarrow \mathcal{S}' \cup \{k^\star\}$ and $\mathcal{U} \leftarrow \mathcal{U} \setminus \{k^\star\}$.
Step 7: Update the offloading energy of the system; i.e., $E_{\text{off}}^{(i)} = E_{\text{total}_{k^\star}}^{(i)}$.
Step 8: If $\mathcal{U} = \emptyset$, stop. If not go to Step 3.

analysis for Algorithm 3 yields a computational cost that is $O(K^3 \log K)$.

Algorithm 3 can be modified for the case of conventional local communication architecture by replacing each $\underline{E}_{\text{loc}_k}$ by the energy required to complete the task locally using the conventional architecture, and by adjusting the initialization of the offloading set $\mathcal{S}'$ and the undecided set $\mathcal{U}$. The set $\mathcal{S}'$ is initialized with those users for which the task cannot be completed locally by the deadline, and $\mathcal{U}$ is initialized as $\{1, 2, \ldots, K\} \setminus \mathcal{S}'$. The initial offloading energy $E_{\text{off}}^{(0)}$ is set to be the optimal total energy consumption of the users in the initial set $\mathcal{S}'$.

Rounding-based algorithm

Wang *et al.* (2018b) also proposed a binary offloading algorithm for systems with equal latencies that is based on choosing the set of offloading users by rounding

the solution to the corresponding partial offloading problem, and then solving the complete offloading problem for that set. That rounding approach extends naturally to the formulation that we have considered. Furthermore, since the rounding process selects offloading users by rounding the optimal value of the offloading fraction in the partial offloading problem, $\gamma_k^\star \in [0, 1]$, to a value in $\{0, 1\}$, there is a natural extension to randomized rounding (cf. Raghavan and Tompson, 1987). In that case, multiple candidate sets of offloading users are selected according to independent Bernoulli distributions with the probability of offloading for user k being $\gamma_k^\star$. The hybrid scheme of deterministic and randomized rounding also arises naturally. Our numerical results in Section 3.6 will show that although the incorporation of randomized rounding offers better performance than deterministic rounding of the partial offloading solution, the proposed greedy search over the tree of complete offloading problems offers significant reductions in the energy consumption, at a computational cost that is similar to that of the deterministic rounding approach.

3.5 Partial offloading

Up until this point, we have considered computational tasks with tightly coupled components, which must be either totally offloaded or executed locally. If the computational tasks are divisible, the device energy consumption can be reduced by taking advantage of the parallelism between the access point and the devices. In that case, each user offloads a portion of its task and executes the remaining portion locally. As mentioned earlier, we will focus on "data-partitionable" tasks (Wang *et al.*, 2016). Such tasks involve a relatively simple-to-describe action being applied, independently, to multiple blocks of data. As such, the number of operations required to complete

a fraction of the task is modeled as being a function of the description length (Wang *et al.*, 2016; Muñoz *et al.*, 2015; Zhang *et al.*, 2013), cf. (3.3) and (3.6).

3.5.1 Full Multiple Access

In this section we will consider a K-user partial offloading system that employs a FullMA scheme. The users are assumed to adopt dynamic voltage scaling so that they can minimize their local computation energy consumption. (In this setting the local latency constraint is satisfied implicitly and E_{loc_k} takes the form in (3.6).) Considering that in a FullMA scheme $t_{\mathrm{w}_k} = 0, \forall k$, the energy minimization problem is

$$\min_{\{R_k\},\{P_k\},\{\gamma_k\}} \sum_k \tfrac{\gamma_k B_k}{R_k} P_k + \tfrac{M_k}{L_k^2}\big((1-\gamma_k)B_k\big)^3 \tag{3.26a}$$

$$\text{s.t.} \quad T_s\big(\tfrac{\gamma_k B_k}{R_k}\big) + \delta_c \gamma_k B_k \leq \bar{L}_k, \quad \forall k, \tag{3.26b}$$

$$0 \leq \gamma_k \leq 1, \quad \forall k, \tag{3.26c}$$

$$0 \leq P_k, \quad \forall k, \tag{3.26d}$$

$$\{R_k\}_{k=1}^K \in \mathcal{R}\big(\{P_k\}_{k=1}^K\big), \tag{3.26e}$$

where $\bar{L}_k = L_k - t_{\mathrm{DL}_k}$. The achievable rate region for FullMA was described in Section 3.4.2.

Using the insights generated from a two-user case in Chapter 2 (see also Salmani and Davidson, 2018b), it can be shown that optimal solution of the problem in (3.26) is obtained when each user utilizes its maximum allowable latency, i.e., the constraints in (3.26b) hold with equality. Accordingly, the closed-form solution for the optimal

fraction of bits offloaded by the k^{th} user is

$$\gamma_k = \frac{\bar{L}_k R_k}{B_k(T_s+\delta_c R_k)}, \tag{3.27}$$

and the problem in (3.26) can be reduced to

$$\min_{\{R_k\},\{P_k\}} \sum_k \frac{\bar{L}_k}{T_s+\delta_c R_k} P_k + \frac{M_k}{L_k^2}(B_k - \frac{\bar{L}_k R_k}{T_s+\delta_c R_k})^3 \tag{3.28a}$$

$$\text{s.t.} \quad 0 \leq \frac{\bar{L}_k R_k}{B_k(T_s+\delta_c R_k)} \leq 1, \quad \forall k, \tag{3.28b}$$

$$(3.26\text{d}), (3.26\text{e}), \tag{3.28c}$$

where (3.28b) results from the constraints in (3.26c). The problem in (3.28) can be decomposed as

$$\begin{array}{ll} \min\limits_{\{R_k\}} & \min\limits_{\{P_k\}} \ \sum_k \frac{\bar{L}_k}{T_s+\delta_c R_k} P_k \\ \text{s.t.} \ (3.26\text{e}), (3.28\text{b}), & \text{s.t.} \quad (3.26\text{d}), (3.26\text{e}). \end{array} \tag{3.29}$$

For a given set of transmission rates, the objective function in (3.29) has a structure that is analogous to that of the objective in (3.12). Hence, if the permutation π is defined such that

$$\rho'_{\pi(1)} \geq \rho'_{\pi(2)} \geq \cdots \geq \rho'_{\pi(K-1)} \geq \rho'_{\pi(K)}, \tag{3.30}$$

where

$$\rho'_k = \frac{\bar{L}_k}{\alpha_k(T_s+\delta_c R_k)}, \tag{3.31}$$

the following closed-form optimal solution for the transmission powers can be obtained

$$P_k = (\frac{2^{R_k}-1}{\alpha_k})2^{\sum_{j=1}^{\pi^{-1}(k)-1} R_j}. \tag{3.32}$$

As in the complete offloading case, the ordering in (3.30) also specifies the decoding order that enables the rates that will be found in (3.33) to be achieved using successive decoding.

Now the outer optimization problem in (3.29) becomes

$$\min_{\{R_k\}} \quad \sum_k \frac{\bar{L}_k}{\alpha_k}(\frac{2^{R_k}-1}{T_s+\delta_c R_k})2^{\sum_{j=1}^{\pi^{-1}(k)-1} R_j} + \sum_k \frac{M_k}{L_k^2}(B_k - \frac{\bar{L}_k R_k}{T_s+\delta_c R_k})^3 \tag{3.33a}$$

$$\text{s.t.} \quad 0 \leq \frac{\bar{L}_k R_k}{B_k(T_s+\delta_c R_k)} \leq 1, \quad \forall k. \tag{3.33b}$$

We have shown in Appendix 3.B that the objective function in (3.33) is quasi-convex in terms of each R_k when the other transmission rates are fixed. In addition, the constraints on the transmission rates are separable. Therefore, the coordinate descent algorithm can be employed to find a stationary solution for the transmission rates in (3.33) (e.g., Hong *et al.*, 2016, Theorem 1).

Using the obtained solutions for the transmission rates, we can update the values of the ρ_k's in (3.31) and consequently the optimal values of the transmission powers in (3.32). By substituting the updated transmission powers into the problem in (3.33), updated solutions for the transmission rates can be achieved. The resulting iterative algorithm is summarized in Algorithm 4. The computation cost of each iteration of Algorithm 4 is dominated by the ordering in Step 3 of the algorithm, the complexity of which is $O(K \log K)$. While the development of a formal convergence analysis of Algorithm 4 remains a work in progress, in our numerical experience, some of which

Algorithm 4 : Iterative algorithm for (3.26)

Input data: $\{B_k\}$, $\{\bar{L}_k\}$, $\{M_k\}$, $\{\alpha_k\}$, T_s, and δ_c.
Step 1: Initialize $\{R_k\}$ so that (3.26d) and (3.33b) are satisfied.
Step 2: Calculate the optimal $\{\gamma_k\}$ using (3.27).
Step 3: Calculate $\{\rho'_k\}$ using (3.31).
Step 4: Order $\{\rho'_k\}$ according to (3.30).
Step 5: Calculate the optimal powers using (3.32).
Step 6: Find a stationary point of the problem in (3.33).
Step 7: If the convergence criterion has been satisfied terminate the algorithm. Otherwise return to Step 2.

is reported in Section 3.6, the algorithm always converged quite fast; typically in 2–5 iterations and in no more than 10 iterations.

3.5.2 Time Division Multiple Access

For a TDMA-based partial offloading system, if we order the users such that $\bar{L}_1 \leq \bar{L}_2 \leq \cdots \leq \bar{L}_K$, where $\bar{L}_k$ was defined after (3.26), then, analogous to the binary offloading case, the waiting time for user k is $t_{\mathrm{w}_k} = \sum_{i=1}^{k-1} T_s\big(\frac{\gamma_i B_i}{R_i}\big)$, and the device energy minimization problem for the optimized dynamic voltage scaling architecture can be written as

$$\min_{\{R_k\},\{P_k\},\{\gamma_k\}} \sum_k \tfrac{\gamma_k B_k}{R_k} P_k + \tfrac{M_k}{L_k^2}\big((1-\gamma_k)B_k\big)^3 \tag{3.34a}$$

$$\text{s.t.} \quad \textstyle\sum_{i=1}^{k} T_s\big(\frac{\gamma_i B_i}{R_i}\big) + \delta_c \gamma_k B_k \leq \bar{L}_k, \quad \forall k, \tag{3.34b}$$

$$0 \leq \gamma_k \leq 1, \quad \forall k, \tag{3.34c}$$

$$0 \leq P_k, \quad \forall k, \tag{3.34d}$$

$$0 \leq R_k \leq \log_2(1+\alpha_k P_k), \quad \forall k. \tag{3.34e}$$

For a given set of $(\{R_k\},\{\gamma_k\})$ the objective is increasing in each P_k, and the

constraints on the powers are separable. Hence, the optimal powers are the minimum feasible values; i.e., $P_k = \frac{2^{R_k}-1}{\alpha_k}$. If we let $B'_k = \gamma_k B_k$ and $t_k = \frac{B'_k}{R_k}$ denote the offloaded portion of the computational task for the k^{th} user, and the time it takes to offload that portion to the access point, respectively, the problem in (3.34) can then be written as

$$\min_{\{B'_k\},\{t_k\}} \sum_k t_k \frac{2^{B'_k/t_k}-1}{\alpha_k} + \frac{M_k}{L_k^2}(B_k - B'_k)^3 \tag{3.35a}$$

$$\text{s.t.} \quad \sum_{i=1}^{k} T_s t_i + \delta_c B'_k \leq \bar{L}_k, \quad \forall k, \tag{3.35b}$$

$$0 \leq B'_k \leq B_k, \quad \forall k, \tag{3.35c}$$

$$0 \leq t_k, \quad \forall k. \tag{3.35d}$$

It is shown in Appendix 3.C that this problem is jointly convex in $\{B'_k\}$ and $\{t_k\}$ and hence, the optimal solution of the problem can be efficiently obtained.

3.6 Numerical Results

In this section we will evaluate the performance of the proposed energy minimization algorithms in both binary offloading and partial offloading scenarios, using either a full multiple access scheme (FullMA) or TDMA. We will compare the performance and computational cost of the proposed algorithms to those of Wang *et al.* (2018b) and You *et al.* (2017). The approach of Wang *et al.* (2018b) is a "full multiple access" approach, but is constrained to the case in which the latencies of the users are the same. Furthermore, the algorithm of Wang *et al.* (2018b) does not exploit as much of the algebraic structure of the problem as our algorithm and hence its computational

cost grows more quickly than that of the proposed algorithm; see the discussion in Section 3.4.4. The approach of You *et al.* (2017) tackles the energy minimization problem for partial offloading in the TDMA case. Like the approach of Wang *et al.* (2018b), it is also constrained to the case in which the latencies of all users are the same. The approach of You *et al.* (2017) is developed for conventional local computing architectures whereas the proposed approaches and those of Wang *et al.* (2018b) are developed for the dynamic voltage scaling architecture.

We will consider a cell of radius 1,000m over which the users are uniformly distributed. The symbol interval is $T_s = 10^{-6}$s and we consider a slowly fading channel model with a path-loss exponent of 3.7 and independent Rayleigh distributed small-scale fading. The receiver noise variance is set to $\sigma^2 = 10^{-13}$. The energy consumption in each experiment is averaged over 100 channel realizations. We assume that the time it takes to download the results to the mobile users is equal for all the users, $t_{\mathrm{DL}_k} = 0.2$s.

3.6.1 Binary Computation Offloading

In the first phase of our numerical experiments we will consider the case where the users seek to complete indivisible computational tasks, and hence they should either offload their task or complete it locally. We will begin by considering a four-user system in which the users latencies are different, $[L_1, L_2, L_3, L_4] = [1.2, 1.5, 1.8, 2.5]$s, and we will examine the energy consumption of FullMA and TDMA-based binary offloading systems as the (different) description lengths of the tasks grow (in proportion); $[B_1, B_2, B_3, B_4] = \zeta \times [2, 1, 3, 4] \times 10^6$ bits. In order to model the optimized

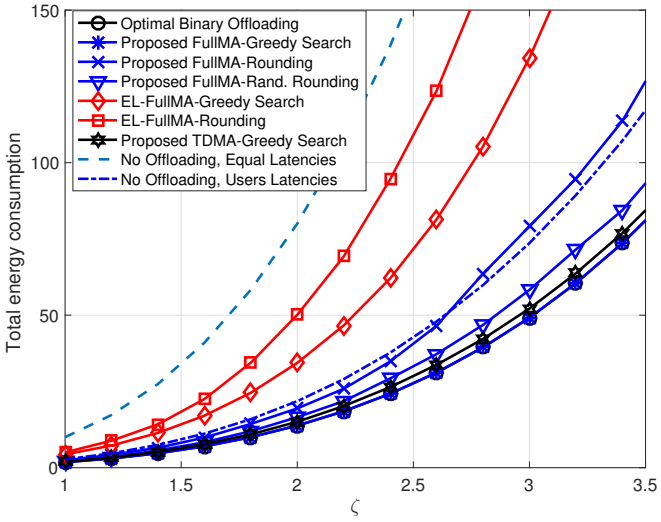


Figure 3.1: Average energy consumption of a binary offloading system with four users with different latency constraints versus the parameter that defines the required number of bits to describe the users' tasks, where EL-FullMA is the equal-latency FullMA-based approach proposed by Wang *et al.* (2018b).

energy consumption of local execution in each user, while meeting its latency constraint, we set $\underline{E}_{\mathrm{loc}_k} = \frac{M_k}{L_k^2} B_k^3$ (see (3.7)), and to be consistent with the measurements of Miettinen and Nurminen (2010), we set $M_k = 10^{-19}$ (Zhang *et al.*, 2013; Wang *et al.*, 2016). We apply the proposed greedy algorithm (Algorithm 3) to both a FullMA scheme and the TDMA scheme to find a good set of offloading users and the corresponding power and rate allocation. We will compare the energy consumption of these schemes to that of schemes in which the offloading set is chosen by deterministic rounding of the solution of the corresponding partial offloading problem, and to a scheme that selects the best solution from the deterministically rounded case and $(K-1)$ randomized roundings; see Section 3.4.4. In the case of FullMA we compare the performance and the computational cost of the proposed algorithm with those of the binary offloading algorithm proposed by Wang *et al.* (2018b).

Fig. 3.1 plots the average energy consumption of the four-user system as the problem sizes grow. Our first observation is that the proposed greedy search algorithm to find a set of offloading users provides close-to-optimal performance for both FullMA and TDMA, and significantly better performance than the deterministic rounding approach. In this setting, the optimized TDMA scheme performs quite well, but in other scenarios that we will consider (Figs 3.2, 3.4, and 3.5) an optimized FullMA scheme enables a significantly larger reduction in the energy consumption.

It can be seen from Fig. 3.1 that utilizing the maximum available latencies of the users enables the proposed algorithm to substantially reduce the energy consumption compared to the algorithm of Wang *et al.* (2018b), in which the users are assumed to have the same latency constraints. The performance gap increases quite quickly as the sizes of the problems increase. In the "No Offloading" approach in Fig. 3.1, the users complete their tasks locally employing the dynamic voltage scaling approach, by which they can minimize the local energy consumption subject to their latency constraints. Interestingly, the energy consumption when all users complete their tasks locally using the maximum available latency is substantially less than that of the latency-equal algorithm proposed by Wang *et al.* (2018b) and the case in which the offloading set is chosen by deterministically rounding the solution to the partial offloading problem with different latencies.

In order to compare the computational costs of the proposed FullMA algorithm with that of Wang *et al.* (2018b), Table 3.1 provides the average CPU times. These times are essentially independent of the description length of the tasks. All the algorithms were coded in MATLAB, with similar diligence paid to the efficiency of the programs. The convex optimization subproblems in the method of Wang *et al.*

Table 3.1: Average CPU times required for the proposed algorithms and the algorithms proposed by Wang *et al.* (2018b) for a four-user binary offloading system that employs a full multiple access scheme.

Algorithm	**Average CPU time (sec)**
Proposed FullMA-Greedy Search	4.6×10^{-5}
Proposed FullMA-Rounding	4.0×10^{-5}
EL-FullMA-Greedy Search (Wang *et al.*, 2018b)	1.7×10^{3}
EL-FullMA-Rounding (Wang *et al.*, 2018b)	0.2×10^{3}

(2018b) were solved using SDPT3 (Toh *et al.*, 1999) through the CVX interface (Grant *et al.*, 2008). The CPU times were evaluated on a MacBook Pro with a Core i5 processor running at 3.1GHz, and 8GB of RAM. It can be seen that the closed-form optimal solution that we have obtained for any given set of offloading users significantly reduces the computational cost of our proposed algorithm in comparison to the algorithm of Wang *et al.* (2018b). As discussed in Section 3.4.4, the main reason for such a significant computational cost reduction is that at each iteration of the proposed algorithm the optimal closed-form solution for a given set of offloading users is obtained with the cost of order $O(K \log K)$, while at each iteration of the algorithm proposed by Wang *et al.* (2018b) an optimization problem needs to be solved by employing the ellipsoid method which involves matrix inversion with the cost of order $O(K^3)$. (As suggested by Wang *et al.* (2018b), for the ellipsoid method we employed the approach of Boyd (2018), and we chose a termination criterion of $\epsilon = 10^{-3}$.)

In our next numerical experiment for the binary offloading case, we examine the total energy consumption as the number of users increases. In this experiment we

consider a scenario in which all the users have equal problem sizes and the same latency constraints. In particular, we set $B_k = 6 \times 10^6$ bits and $L_k = 2$s. As in the previous experiment, the "randomized rounding" scheme refers to the selection of the best solution from offloading sets that are generated by a deterministic rounding of the partial offloading solution and $(K-1)$ randomized roundings.

In Fig. 3.2 we present the average energy consumption versus the number of users. In this setting all of the considered methods provide a significant reduction in the energy consumption over the No Offloading case. In the case that a FullMA scheme is employed, it can be seen that since the latency constraints of all the users are equal, the algorithm of Wang *et al.* (2018b) can achieve the same performance as our proposed algorithm, for both greedy search and rounding approaches. However, Fig. 3.3 indicates that the computational cost of the proposed algorithm is significantly less than that of the algorithm of Wang *et al.* (2018b). Fig. 3.2 also shows that by using the full capabilities of the channel, a FullMA scheme together with the proposed greedy search method can reduce the total energy consumption compared to the TDMA scheme with the same greedy approach.

3.6.2 Partial Computation Offloading

In the second phase of our numerical analysis we consider partial offloading of "data-partitionable" divisible computational tasks for which the (optimal) local energy consumption can be modeled as a function of number of bits, see (3.6), with $M_k = 10^{-19}$; (Zhang *et al.*, 2013; Wang *et al.*, 2016). To make fair comparisons with the conventional local computational architecture considered by You *et al.* (2017), we consider

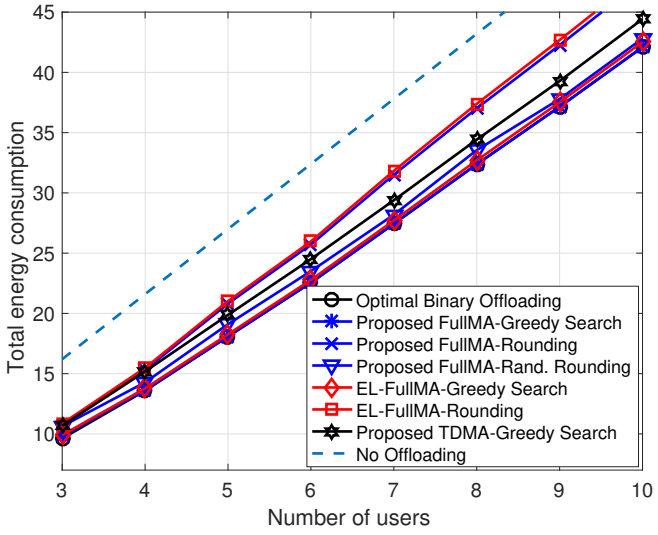


Figure 3.2: Average energy consumption of a binary offloading system, in which the users' tasks have the same latency constraints, for different number of users. EL-FullMA denotes the equal-latency FullMA-based approach proposed by Wang *et al.* (2018b).

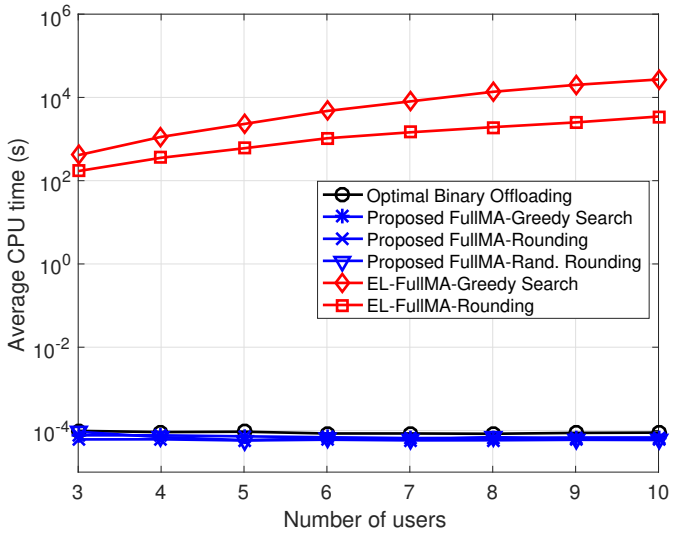


Figure 3.3: Average CPU time required for the proposed algorithm and the EL-FullMA algorithm proposed by Wang *et al.* (2018b) for different number of users when a full multiple access scheme is employed in binary offloading case.

problems that require 1,000 computational cycles per bit, and we set the local computing energy per cycle for each user in such a way that that user is able to complete its computational task locally within its latency constraint. We first examine the energy consumption of a four-user system analogous to that in Section 3.6.1, in which the latencies are $[L_1, L_2, L_3, L_4] = [1.2, 1.5, 1.8, 2.5]$s and the description lengths grow as $[B_1, B_2, B_3, B_4] = \zeta \times [2, 1, 3, 4] \times 10^6$ bits. Fig. 3.4 plots the total energy consumption as the problem sizes grow. It can be seen that our proposed algorithms, which benefit from the maximum available latency of each user, achieve substantially lower energy consumption than the existing techniques. Indeed, it can be seen that in the TDMA case, the energy consumption of the proposed algorithm is lower than that of the algorithm of You *et al.* (2017), and the performance gap increases as the number of bits increases. That is because in the proposed algorithm the users not only utilize their maximum available deadline to complete their tasks, they also employ dynamic voltage scaling which minimizes the local energy consumption. The energy consumptions in Fig. 3.4 and the computational costs in Table 3.2 indicate that in the FullMA case the proposed algorithm can achieve significantly lower energy consumption than the algorithm of Wang *et al.* (2018b), and does so at much lower computational cost. Fig. 3.4 also exhibits the impact of the multiple access scheme. Using a FullMA scheme substantially reduces the total energy consumption over TDMA.

In our final numerical experiment we examine the energy consumption as the number of users increases for a partial offloading system with equal problem sizes and the same latency constraints. We set $B_k = 4 \times 10^6$ bits and $L_k = 2$s. Fig. 3.5, like Fig. 3.4, shows that using the full capabilities of the channel enables the users to complete their computational tasks with significantly less energy consumption

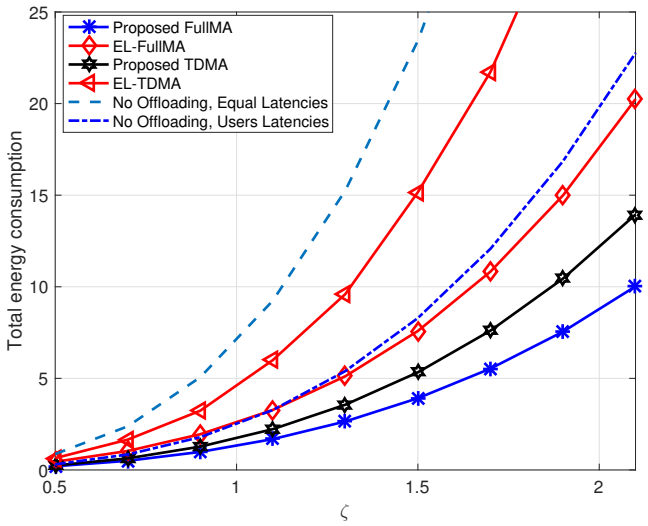


Figure 3.4: Average energy consumption of a four-user partial offloading system with different latency constraints versus the coefficient that defines the description length of the tasks, where EL-FullMA is the equal-latency FullMA-based approach proposed by Wang *et al.* (2018b), and EL-TDMA is the equal-latency TDMA-based approach of You *et al.* (2017).

compared to TDMA. In the FullMA case, it can be seen in Fig. 3.5 that because the latencies of the users are equal, the algorithm of Wang *et al.* (2018b) can achieve the same performance as our proposed algorithm. However, as it can be seen in Fig. 3.6 the computational cost of the proposed algorithm is much lower. We can also see in Fig. 3.5 that when TDMA is employed, the proposed algorithm achieves noticeably lower energy consumption than that of You *et al.* (2017) despite the fact that the latencies are equal in this scenario. The reason for this is that the proposed algorithm is for systems with dynamic voltage scaling, which enables the users to minimize the energy that they expend on the portion of the task that is computed locally.

Table 3.2: Average CPU times for the proposed algorithm and the equal-latency algorithm of (Wang *et al.*, 2018b) for a four-user FullMA partial offloading system.

Algorithm	**Average CPU time (sec)**
Proposed FullMA	4.1×10^{-3}
EL-FullMA (Wang *et al.*, 2018b)	1.9×10^{2}

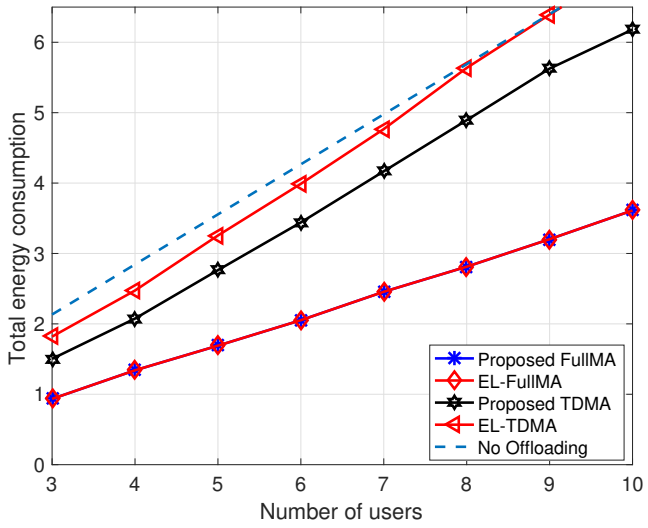


Figure 3.5: Average energy consumption of a partial offloading system, in which the users' tasks have the same latency constraints, for different number of users. EL-FullMA denotes the equal-latency FullMA-based approach proposed by Wang *et al.* (2018b), and EL-TDMA denotes the equal-latency TDMA-based approach proposed by You *et al.* (2017).

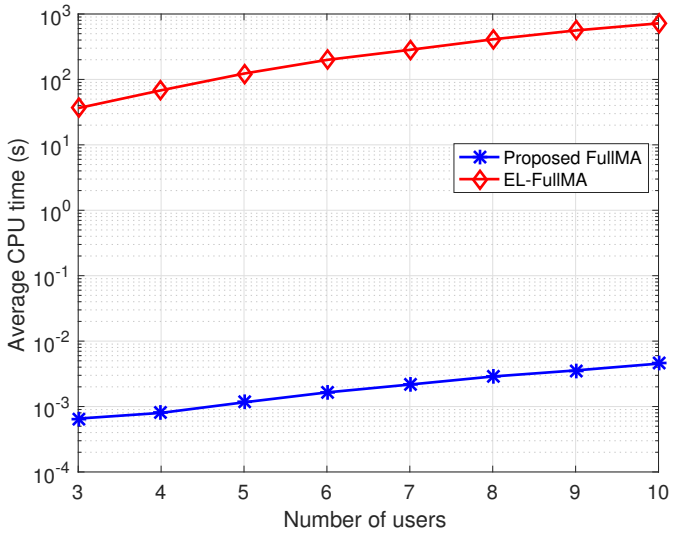


Figure 3.6: Average CPU time required for the proposed algorithm and the EL-FullMA algorithm of Wang *et al.* (2018b) for different number of users when the full multiple access scheme is employed in partial offloading case.

3.7 Conclusion

In this work, we have considered the problem of optimal uplink resource allocation in a K-user offloading system. In the binary offloading case, we obtained the optimal energy consumption of a given set of offloading users under a full multiple access scheme and under the TDMA scheme, and then we proposed a customized greedy search algorithm to find a set of offloading users with close-to-optimal energy consumption. In the partial offloading case, the energy minimization problem was tackled by proposing a low-complexity algorithm for a stationary solution in the full multiple access case and by finding the optimal solution of a convex optimization problem when TDMA is employed. Our strategy to decompose the optimization problem and to find the optimal values of some variables in terms of the others enabled us to significantly reduce the computational cost of our proposed algorithms compared to the existing

algorithms in this area.

While the proposed resource allocation algorithms have significant advantages over the existing algorithms, like the existing algorithms they have been based on a single time slot for communication. The work on the two-user case in Chapter 2 (see also Salmani and Davidson, 2018b) suggests that a further reduction in the energy consumption can be obtained by adopting a time-slotted structure in which different groups of users transmit in each time slot. One avenue for future work is the development of efficient resource allocation algorithms for the time-slotted structure. That avenue will be pursued in the next chapter.

3.A Exploiting the Polymatroid Structure of the Power Feasibility Region

Given the definition of class-ℓ constraints in Section 3.4.2, we will show that the candidate vertices are the result of simultaneous satisfaction with equality of a set of K constraints in (3.11d) such that there is at most one constraint from each class. To do so, let us assume that there are two constraints in (3.11d) that are satisfied with equality, namely C_1 and C_2, both of which belong to class-c. If $\mathcal{M}_{\text{com}}$ denotes the set of users that are present in both C_1 and C_2, and if $\mathcal{M}_{c_1}$ and $\mathcal{M}_{c_2}$ denote the set of users that are participating only in C_1 and C_2 respectively, we can write

$$C_1: \quad 2^{\left(\sum_{i\in\mathcal{M}_{\text{com}}} R_i+\sum_{j\in\mathcal{M}_{c_1}} R_j\right)} = 1+\sum_{i\in\mathcal{M}_{\text{com}}} \alpha_i P_i+\sum_{j\in\mathcal{M}_{c_1}} \alpha_j P_j,$$

$$C_2: \quad 2^{\left(\sum_{i\in\mathcal{M}_{\text{com}}} R_i+\sum_{k\in\mathcal{M}_{c_2}} R_k\right)} = 1+\sum_{i\in\mathcal{M}_{\text{com}}} \alpha_i P_i+\sum_{k\in\mathcal{M}_{c_2}} \alpha_k P_k.$$

By adding the above two equations we have that

$$\left(2^{\sum_{i\in\mathcal{M}_{\text{com}}} R_i}\right)\left(2^{\sum_{j\in\mathcal{M}_{c_1}} R_j} + 2^{\sum_{k\in\mathcal{M}_{c_2}} R_k}\right) - 1 - \sum_{i\in\mathcal{M}_{\text{com}}} \alpha_i P_i = 1 + \sum_{i\in\mathcal{M}_{\text{com}}} \alpha_i P_i + \sum_{j\in\mathcal{M}_{c_1}} \alpha_j P_j + \sum_{k\in\mathcal{M}_{c_2}} \alpha_k P_k. \quad (3.36)$$

In addition, there is a rate region constraint that includes all the users in $\mathcal{M}_{\text{com}} \cup \mathcal{M}_{c_1} \cup \mathcal{M}_{c_2}$,

$$2^{\left(\sum_{i\in\mathcal{M}_{\text{com}}} R_i + \sum_{j\in\mathcal{M}_{c_1}} R_j + \sum_{k\in\mathcal{M}_{c_2}} R_k\right)} \leq 1 + \sum_{i\in\mathcal{M}_{\text{com}}} \alpha_i P_i + \sum_{j\in\mathcal{M}_{c_1}} \alpha_j P_j + \sum_{k\in\mathcal{M}_{c_2}} \alpha_k P_k. \quad (3.37)$$

The right hand side of (3.37) can be replaced by its equivalent term given on the left hand side of (3.36). That results in

$$2^{\left(\sum_{i\in\mathcal{M}_{\text{com}}} R_i + \sum_{j\in\mathcal{M}_{c_1}} R_j + \sum_{k\in\mathcal{M}_{c_2}} R_k\right)} \leq \left(2^{\sum_{i\in\mathcal{M}_{\text{com}}} R_i}\right)\left(2^{\sum_{j\in\mathcal{M}_{c_1}} R_j} + 2^{\sum_{k\in\mathcal{M}_{c_2}} R_k}\right) - 1 - \sum_{i\in\mathcal{M}_{\text{com}}} \alpha_i P_i \leq \left(2^{\sum_{i\in\mathcal{M}_{\text{com}}} R_i}\right)\left(2^{\sum_{j\in\mathcal{M}_{c_1}} R_j} + 2^{\sum_{k\in\mathcal{M}_{c_2}} R_k} + 1\right), \quad (3.38)$$

where the second inequality in (3.38) is obtained from the rate region constraint $2^{\sum_{i\in\mathcal{M}_{\text{com}}} R_i} \leq 1 + \sum_{i\in\mathcal{M}_{\text{com}}} \alpha_i P_i$. By factoring out the term $2^{\sum_{i\in\mathcal{M}_{\text{com}}} R_i}$, we obtain

$$\left(2^{\sum_{j\in\mathcal{M}_{c_1}} R_j + \sum_{k\in\mathcal{M}_{c_2}} R_k}\right) \leq 2^{\sum_{j\in\mathcal{M}_{c_1}} R_j} + 2^{\sum_{k\in\mathcal{M}_{c_2}} R_k} - 1 \Rightarrow 0 \leq \left(2^{\sum_{j\in\mathcal{M}_{c_1}} R_j} - 1\right)\left(1 - 2^{\sum_{k\in\mathcal{M}_{c_2}} R_k}\right),$$

which is a contradiction, because of the fact that $0 \leq R_i$ and hence $0 \leq 2^{\sum_{j\in\mathcal{M}_{c_1}} R_j} - 1$ for any j. Therefore, at an optimal vertex of the inner problem in (3.12) no more than one constraint from any class can hold with equality.

3.B Quasi-convexity of the Objective Function in (3.33)

A function f is quasi-convex if at least one of the following conditions holds (Boyd and Vandenberghe, 2004): (a) f is non-increasing; (b) f is non-decreasing; (c) there is a (turning) point, c, such that for any $x \leq c$ the function $f(x)$ is non-increasing and for any $x \geq c$ the function $f(x)$ is non-decreasing. We will show that for each R_k, when the other transmission rates are constant, the objective function in (3.33) will satisfy either condition (b) or condition (c). We begin by rewriting that objective as

$$f_k = \Lambda_k\left(\frac{2^{R_k}-1}{T_s+\delta_c R_k}\right) + \Omega_k 2^{R_k} + \frac{M_k}{L_k^2}\left(B_k - \frac{\bar{L}_k R_k}{T_s+\delta_c R_k}\right)^3, \tag{3.39}$$

where $\Lambda_k = \frac{\bar{L}_k}{\alpha_k} 2^{\sum_{j=1}^{k-1} R_j}$ and $\Omega_k = \sum_{i=k+1}^{K} \frac{\bar{L}_i}{\alpha_i}\left(\frac{2^{R_i}-1}{T_s+\delta_c R_i}\right) 2^{\sum_{j\neq k}^{i-1} R_j}$ are always positive. The derivative of f_k with respect to R_k can be then written as $\frac{df_k}{dR_k} = \frac{F_r}{(T_s+\delta_c R_k)^2}$, where

$$\begin{aligned} F_r = \; & \Lambda_k\big(\ln 2\ (T_s+\delta_c R_k)\ 2^{R_k} - \delta_c(2^{R_k}-1)\big) + \Omega_k \ln 2\ (T_s+\delta_c R_k)^2\ 2^{R_k} \\ & - 3\bar{L}_k T_s \frac{M_k}{L_k^2}\left(B_k - \frac{\bar{L}_k R_k}{T_s+\delta_c R_k}\right)^2. \end{aligned}$$

As $\frac{1}{(T_s+\delta_c R_k)^2}$ is always positive, to show that either condition (b) or condition (c) holds, it is sufficient to show that F_r is non-decreasing. In order to show that, we will show that the derivative of F_r with respect to R_k is always non-negative. The

derivative is

$$\frac{dF_r}{dR_k} = \Lambda_k\big(\ln^2 2\ (T_s+\delta_c R_k)\ 2^{R_k}\big) + \Omega_k \ln 2\big(\ln 2\ (T_s+\delta_c R_k)^2 + 2\delta_c(T_s+\delta_c R_k)\big)\ 2^{R_k} + 6\bar{L}_k T_s \frac{M_k}{L_k^2}\big(B_k - \frac{\bar{L}_k R_k}{T_s+\delta_c R_k}\big)\big(\frac{\bar{L}_k T_s}{(T_s+\delta_c R_k)^2}\big).$$

Considering the constraint in (3.33b), $\frac{dF_r}{dR_k}$ is a summation of non-negative terms. Hence, $\frac{dF_r}{dR_k}$ is non-negative, and hence F_r is non-decreasing.

3.C Joint Convexity of the Objective Function in (3.35)

In order to show that the objective function in (3.35) is jointly convex in terms of $\{B'_k\}$ and $\{t_k\}$ we will show that the Hessain matrix of the objective is positive semidefinite. The first and the second derivatives of the objective, $f(\cdot)$, with respect to each of the B'_k's and t_k's are

$$\frac{\partial f}{\partial t_k} = \frac{1}{\alpha_k}\Big(2^{B'_k/t_k} - 1 - \ln 2\ \big(\frac{B'_k}{t_k}\big)\ 2^{B'_k/t_k}\Big), \quad \frac{\partial f}{\partial B'_k} = \frac{\ln 2}{\alpha_k}\ 2^{B'_k/t_k} - \frac{3M_k}{L_k^2}(B_k - B'_k)^2, \tag{3.42a}$$

$$\frac{\partial^2 f}{\partial t_k^2} = \frac{1}{\alpha_k}\Big(\ln^2 2\ \big(\frac{{B'_k}^2}{{t_k}^3}\big)\ 2^{B'_k/t_k}\Big), \quad \frac{\partial^2 f}{\partial {B'_k}^2} = \frac{\ln^2 2}{\alpha_k t_k}\ 2^{B'_k/t_k} + \frac{6M_k}{L_k^2}(B_k - B'_k), \tag{3.42b}$$

$$\frac{\partial^2 f}{\partial t_k \partial B'_k} = \frac{1}{\alpha_k}\Big(-\ln^2 2\ \big(\frac{B'_k}{t_k^2}\big)\ 2^{B'_k/t_k}\Big), \tag{3.42c}$$

and for $j \neq k$, $\frac{\partial^2 f}{\partial t_j \partial t_k}$, $\frac{\partial^2 f}{\partial B'_j \partial B'_k}$, and $\frac{\partial^2 f}{\partial t_j \partial B'_k}$ are all zero. The Hessian matrix $H \in \mathbb{R}^{2K\times 2K}$ can be constructed as the block matrix $H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix}$, where H_{11}, H_{22}, and H_{12} are diagonal matrices with i^{th} diagonal elements, $\frac{\partial^2 f}{\partial {t_i}^2}$, $\frac{\partial^2 f}{\partial {B'_i}^2}$, and $\frac{\partial^2 f}{\partial t_i \partial B'_i}$, respectively.

It can be seen from (3.42b) that all the elements of H_{11} are positive, and hence $H_{11} \succ 0$. Moreover, using (3.42b) and (3.42c), we can show that the inequality $\frac{\partial^2 f}{\partial {B'_k}^2} \times \frac{\partial^2 f}{\partial t_k^2} - (\frac{\partial^2 f}{\partial t_k \partial B'_k})^2 \geq 0$ holds for each user. These inequalities, together with the fact that the sub-blocks of the matrix H are diagonal matrices, illustrate that the Schur complement of the matrix H_{11} in H is positive semidefinite; i.e., $H_{22} - H_{12}^T H_{11}^{-1} H_{12} \succeq 0$. Hence, the matrix H is positive semidefinite (Boyd and Vandenberghe, 2004) and the objective function is jointly convex.

Chapter 4

Energy-Optimal Time-Slotted Multiple Access Computation Offloading

4.1 Abstract

Computation offloading enables energy-limited devices to expand the scope of the computational tasks that they can complete within specified latency requirements. However, when multiple devices seek to offload, effective allocation of resources becomes crucial. In this chapter, we seek an energy-optimal communication resource allocation for a K-user offloading system. We first introduce a time-slotted signalling structure that exploits the differences between the users' latencies by taking advantage of the interference reduction that arises when a device completes its offloading. We then obtain an optimized time-slotted structure for a multiple access scheme that exploits the full capabilities of the channel (FullMA). The optimized signalling

structure enables us to substantially reduce the dimension of the resource allocation problem, and to develop efficient algorithms to tackle that problem for both the binary and partial offloading cases. Our numerical experiments illustrate that the proposed time-slotted FullMA signalling structure significantly reduces the energy consumption of the devices compared to some existing methods that employ orthogonal multiple access schemes, such as TDMA, and to those with FullMA, but sub-optimal single-time-slot signalling structures.

4.2 Introduction

While the widespread adoption of mobile computing devices and wireless communication networks is beginning to realize the vision of ubiquitous computing, the modest computational, storage, memory, and energy resources of certain classes of devices remain significant constraints on the scope of the latency-sensitive applications and services that can be offered. The Mobile Edge Computing (MEC) framework addresses those constraints by offering the opportunity for the devices to offload (a portion of) their computational tasks to a shared computation resource, e.g., a computationally-equipped access point. In essence, it provides a potential computation parallelism between the devices and the access point, which facilitates the completion of low-latency computational tasks and can reduce the energy that the devices must expend (Dinh *et al.*, 2013; Hu *et al.*, 2015; Wang *et al.*, 2017b; Mao *et al.*, 2017b; Mach and Becvar, 2017).

In order to fully exploit the opportunities provided by the MEC framework, the available resources must be optimally allocated to the devices. Doing so involves an assessment of the energy and time that each device would expend offloading (a portion

of) its task or executing it locally, subject to constraints on the communication and computational infrastructure at the devices and the access point, and the latency and memory constraints of each device's task (Barbera *et al.*, 2013; Lei *et al.*, 2013; Mao *et al.*, 2017b; Liu *et al.*, 2013; Sardellitti *et al.*, 2015; Chen *et al.*, 2015; Muñoz *et al.*, 2015, 2014; Salmani and Davidson, 2016; Mao *et al.*, 2017a; Wang *et al.*, 2018a; Salmani and Davidson, 2018b, and Chapters 2, 3). The resource allocation also depends on the nature of the task itself. For example, if the components of the task are tightly coupled (and hence the task is indivisible), the device must either completely offload the task or execute the whole task locally; i.e., the offloading decisions are binary. In contrast, when the components of a task are independent (and hence the task is divisible), it can be "partially" offloaded, with a portion of the task being sent to the access point, while the remainder is executed in the mobile device.

In an offloading system in which multiple devices seek to offload (a portion of) their tasks, the way in which the multiple access scheme manages the interference between the offloading devices has a significant impact on the offloading system; see Chapter 2 and (Salmani and Davidson, 2018b). Most of the existing work considers the case in which orthogonal channels (in time or frequency) are assigned to the users (Chen *et al.*, 2018; Chen *et al.*, 2016b; You *et al.*, 2017; Mao *et al.*, 2017a). However, as shown in Chapter 2 (see also Salmani and Davidson, 2016, 2018b) for a two-user offloading system, using a multiple access scheme that can exploit the full capabilities of the channel (FullMA)[1] can significantly reduce the energy consumption compared

[1]That is, a multiple access scheme that can provide reliable operation at rates that approach the boundary of the capacity region. Optimally-ordered sequential decoding with time sharing is one example (Cover and Thomas, 2012; El-Gamal and Cover, 1980). FullMA will denote any such "capacity-region-approaching" multiple access scheme.

to some sub-optimal multiple access schemes, such as Time Division Multiple Access (TDMA), independent decoding, and sequential decoding without time sharing. In Chapter 3 (see also Salmani and Davidson, 2018c), it has been shown that the energy reduction advantages of a FullMA scheme over orthogonal schemes can be extended to a (simplified) system with K offloading users. That work also showed that exploiting the maximum allowable latency of each user, regardless of the choice of the multiple access scheme, can substantially reduce the total energy consumption as compared to some existing methods (e.g., You *et al.*, 2017; Wang *et al.*, 2018b) that impose an assumption of equal latency constraints on the users.

The main goal of this chapter is to further reduce the total user energy consumption of an offloading system by developing an optimal signalling structure that exploits not only the maximum allowable latency of each user, but also the differences between the latency constraints of the users. In particular, we seek to exploit the reduction in interference that arises when a user finishes offloading. Our focus is on a FullMA-based system with K users, each of which wishes to complete its computational task within its own specific latency constraint, either locally or by offloading (a portion of) the task to an access point with substantial computational resources. Our first step is to fully exploit the differences between the latency constraints by proposing a "time-slotted" signalling structure. In each time slot, different sets of users are offloading, and hence, a different power–rate pair is assigned to each user in each time slot. This is in contrast to the single-time-slot structure introduced in Chapter 3 (see also Salmani and Davidson, 2018c) and (Wang *et al.*, 2018b), in which a single power–rate pair is assigned to each user over its whole transmission time. The proposed time-slotted structure can lead to a significant energy consumption

reduction over the single-time-slot structure, but the number of slots in the generic time-slotted structure grows exponentially in the number of users, K. In our next step, we derive an optimized time-slotted structure for FullMA schemes that requires only K time slots and, we analytically determine the set of users that should transmit in each time slot, and the optimal duration of each slot.

To complete design of the proposed K-user time-slotted FullMA signalling structure we tackle the problem of allocating transmission powers and transmission rates to each user in each time slot, and determining the fraction of the problem description that should be offloaded by each user in each slot. We develop efficient algorithms for that problem in both the case of binary offloading (for indivisible computational tasks), and the case of partial offloading for the class of "data-partitionable" divisible tasks (Wang *et al.*, 2016).

In the binary offloading case, the energy minimization problem has a combinatorial structure imposed by the binary decisions of whether each user offloads its task or completes the task locally. To tackle that non-convex problem, we partition it into an inner problem of finding the resource allocation for a given set of offloading decisions (a problem that we will call the "complete offloading" problem), and an outer (tree-structured) search for the offloading decisions. In the solution of the complete offloading problem, the proposed time-slotted structure, along with the polymatroid structure of the capacity region of the multiple access channel (Tse and Hanly, 1998), enables us to obtain closed-form expressions for a large fraction of the design variables. Exploiting the structure of the remaining reduced-dimension energy minimization problem enables us to employ a block coordinate descent algorithm that is guaranteed to find a stationary solution of the remaining problem. Each

subproblem in that block coordinate descent algorithm is solved using a generalized water filling algorithm (e.g., Xing *et al.*, 2018). The solutions to the (inner) complete offloading problem form the core of a tailored greedy search algorithm for solutions to the binary offloading problem.

In the case of partial offloading, in addition to the communication resource allocation, the fraction of computational task to be offloaded by each user must be optimized. Following a similar approach to that for the complete offloading problem, we obtain closed-form expressions for a large number of variables, including the offloaded fraction for each user. We show that the objective function of the remaining reduced-dimension problem is convex with respect to the set of variables of each user when the variables of other users are fixed, and that the constraints on each user's set of variables are decoupled from the other users. That enables us to employ a block coordinate descent algorithm to find a stationary solution of the resource allocation. Interestingly, we show that each of the subproblems in the block coordinate descent algorithm, which are nominally over each user's set of variables, can be further simplified to a single-variable optimization problem that can be solved using a line search strategy.

In our numerical experiments, we compare the performance of the proposed time-slotted FullMA signalling structure to that of three exiting approaches to multiple access in the offloading context, namely, TDMA (e.g., You *et al.*, 2017; Salmani and Davidson, 2018c, and Chapter 3), which inherently incorporates a time-slotted structure, the single-time-slot FullMA (STS-FullMA) approach of Chapter 3 (see also Salmani and Davidson, 2018c), and the equal-latency FullMA (EL-FullMA) approach of (Wang *et al.*, 2018b). Our comparisons with TDMA indicate that exploiting the

full capabilities of the multiple access channel can significantly reduce the energy consumption. The comparisons with the STS-FullMA and EL-FullMA approaches illustrate the considerable advantages of fully exploiting the differences between the latency constraints of the devices. Although the number of design variables in the proposed time-slotted structure is larger than the number of design variables in those methods, the computational cost of the proposed algorithm is significantly lower than that of the EL-FullMA algorithm (Wang *et al.*, 2018b). That is mainly because in the proposed algorithm we have obtained closed-form (and quasi-closed-form) expressions for a large fraction of the design variables.

The remainder of this chapter is organized as follows. The system model is set up in Section 4.3. In Section 4.4.1, we introduce the generic time-slotted signalling structure and we formulate the energy minimization problem for that structure; see (4.7). The optimized signalling structure for FullMA schemes, including the constituent time slot durations, are derived in Section 4.4.2. We tackle the remaining resource allocation problem in the binary offloading case in Section 4.5, by first finding a solution to the minimum energy consumption problem for a given set of offloading users (Section 4.5.1), and then searching for a good set of offloading users (Section 4.5.2). In Section 4.6, we consider the case of partial offloading and we obtain quasi-closed-form solution to the resource allocation problem for that case. Our numerical results for both the cases of binary and partial offloading are provided in Section 4.7.

4.3 System Model

Let us consider an offloading system with a set $\mathcal{S} = \{1, 2, \cdots, K\}$ of K single-antenna users that are considering the possibility of offloading their latency-constrained computational tasks to a single-antenna access point. The channel between each user and the access point is assumed to be frequency-flat and quasi-static, and we adopt the conventional discrete-time baseband equivalent model with symbol interval T_s. Therefore, if $s_k[n]$ denotes the transmitted signal by the k^{th} user at the n^{th} channel use, and if h_k denotes the channel from the k^{th} user to the access point, then the signal received at the access point at the n^{th} channel use is $y[n] = \sum_{k=1}^{K} h_k s_k[n] + v[n]$, where $v[n]$ is an additive circular zero mean white Gaussian noise of variance σ^2. For later convenience we define $\alpha_k = \frac{|h_k|^2}{\sigma^2}$.

We consider a computational model in which the access point begins executing (its portion of) a computational task only after receiving the full description of (that portion of) the task, and it sends the results back to the user only after fully executing (its portion of) the task. Accordingly, the offloading time of the k^{th} user, t_{off_k}, contains the time it takes for the k^{th} user to offload its task, t_{UL_k}, the time it takes for the access point to execute that task, t_{exe_k}, and the time it takes to send the result back to the user k, t_{DL_k}. Therefore, the latency constraint of the k^{th} user can then be written as

$$t_{\text{off}_k} = t_{\text{UL}_k} + t_{\text{exe}_k} + t_{\text{DL}_k} \leq L_k, \tag{4.1}$$

where L_k denotes the maximum allowable latency for user k. The time t_{DL_k} depends on the description length of the results of the (partially) offloaded task, which is often considerably shorter than the description length of the task itself. It also depends on

the downlink signalling scheme chosen by the access point, and the energy that the access point expends on the downlink, rather than the energy consumption at each user. Accordingly, we will model t_{DL_k} as a (possibly different) constant for each user (see e.g., Sardellitti *et al.*, 2015; Wang *et al.*, 2018a; You *et al.*, 2017). Moreover, under the assumption of sufficiently large computational resources at the access point, we can assume that the time t_{exe_k} is constant (e.g., Wang *et al.*, 2018a; You *et al.*, 2017). Hence, we consider $T_k = t_{\mathrm{exe}_k} + t_{\mathrm{DL}_k}$ to be a (different) constant for each user. In the case that the mobile device contributes to the completion of the task (or completes it entirely), the latency constraint must also be applied to the time it takes for that user to (locally) execute its component of the task, t_{loc_k}; i.e., $t_{\mathrm{loc}_k} \leq L_k$.

In a generic computation offloading scenario, the users seek to complete different computational tasks with different latency requirements. In some of the existing work in this area (e.g., Wang *et al.*, 2018b; You *et al.*, 2017), the users are assumed to have equal latency constraints. While that simplifies the resource allocation problem, it impedes those methods from obtaining the optimal, or a close-to-optimal, offloading energy consumption. In Chapter 3 (see also Salmani and Davidson, 2018c), a single-time-slotted signalling structure for FullMA schemes was introduced in which a single power–rate pair is assigned to each user for transmission. While that scheme enables each user to exploit its maximum latency constraint, L_k, the total energy consumption would be further reduced if we could design a signalling structure that can also exploit the differences between the users' latencies. In Section 4.4.1, we will propose an energy-optimal "time-slotted" signalling structure that realizes this goal. The users not only benefit from their maximum allowable latencies, they also benefit

from the reduction in the interference that arises when a user completes its transmission. Furthermore, we will develop an optimized time-slotted structure for a system that employs a "full" multiple access scheme; see Section 4.4.2.

4.4 Time-Slotted Signalling Structure

As discussed above, the energy consumption of the users in an offloading system can be reduced if the users can take the advantage of the interference reduction that results when a user completes its offloading. That can be achieved by a system with multiple time slots, rather than a single time slot, in which different sets of offloading users, with different transmission rates and powers, are assigned to each time slot. For a K-user system the generic time-slotted structure has $(2^K - 1)$ different time slots. That is quite a contrast to the single-time-slot signalling structure in which a single rate and a single power are assigned to each user, and hence it will be important to exploit the structure of the problem to develop an efficient algorithm for allocating the available resources.

4.4.1 Generic Time-Slotted Signalling Structure

In order to formulate the generic problem of minimizing the total user energy consumption of a time-slotted offloading system, we let $\mathcal{S}_i \subseteq \mathcal{S}$ denote the subset of users that are transmitting in the i^{th} time slot, and let τ_i denote the length (in channel uses) of that time slot. Let P_{ki} and R_{ki} denote the transmission power and transmission rate (in units per channel use) for the k^{th} user in time slot i. Then the offloading

energy consumption of user k in time slot i can then be written as

$$E_{\text{off}_{ki}} = \tau_i P_{ki}, \tag{4.2}$$

where $P_{ki} = 0, \forall k \notin \mathcal{S}_i$. For both the binary and partial offloading cases, we will let γ_k denote the fraction of the description of the task that is offloaded. In the binary offloading case $\gamma_k \in \{0, 1\}$, whereas in the partial offloading case $\gamma_k \in [0, 1]$. The total user energy consumption is then

$$E_{\text{total}} = \sum_k \big(\sum_i E_{\text{off}_{ki}} + E_{\text{loc}_k}(\gamma_k) \big), \tag{4.3}$$

where the function $E_{\text{loc}_k}(\gamma_k)$ describes the energy required to complete the fraction $(1 - \gamma_k)$ of the task that is retained by the user. In the case of binary offloading, the function $E_{\text{loc}_k}(\gamma_k)$ can simply be written as $(1 - \gamma_k)\underline{E}_{\text{loc}_k}$, where $\underline{E}_{\text{loc}_k}$ is the energy required to complete the k^{th} user's task on the local computational architecture of that user. In the case of partial offloading of data-partitionable tasks, the form of $E_{\text{loc}_k}(\gamma_k)$ depends on the nature of the local computational architecture that is employed by the users. In our design in Section 4.6, we will consider users with the dynamic voltage scaling architecture (Wang *et al.*, 2016). In that case, the local energy required to complete the fraction $(1 - \gamma_k)$ of the task that the user retains within the its latency constraint, L_k, is (Wang *et al.*, 2016)

$$E_{\text{loc}_k}(\gamma_k) = \tfrac{M_k}{L_k^2} \big((1 - \gamma_k) B_k \big)^3, \tag{4.4}$$

where M_k depends on the chip characteristics of user k.

In both the binary offloading and partial offloading cases, the fraction of the description of the task that user k offloads in time slot i, $\gamma_{ki} \in [0, 1]$, is an implicit design variable of the energy minimization problem. (Note that $\sum_i \gamma_{ki} = \gamma_k$.) If B_k denotes the description length of that task (in bits), then the number of bits that user k offloads in time slot i is

$$\gamma_{ki} B_k = \tau_i R_{ki}. \tag{4.5}$$

(Here, we have assumed that the descriptions are long enough that we can treat $\gamma_{ki} B_k$ as a continuous quantity.) If $\bar{\imath}_k$ denotes the index of the last time slot in which user k is offloading, then the transmission time for each user, which is the summation of the lengths of the time slots during which the user still has fractions of bits to offload, is

$$t_{\mathrm{UL}_k} = T_s \sum_{i=1}^{\bar{\imath}_k} \tau_i. \tag{4.6}$$

To complete the formulation, we recall that the set $\mathcal{S}_i$ contains the indices of the users that are transmitting in the i^{th} time slot, and we let $\{R_{ki}\}_{k \in \mathcal{S}_i}$ and $\{P_{ki}\}_{k \in \mathcal{S}_i}$ denote the sets of transmission rates and transmission powers in that slot. For simplicity, we let $\{R_{ki}\}$ and $\{P_{ki}\}$ denote the sets of all rates and powers over all time slots. Similarly, we let $\{\tau_i\}$ denote the set of all time slot lengths (in channel uses), and let $\{\mathcal{S}_i\}$ denote the set of all transmitting sets $\mathcal{S}_i$. The energy minimization

problem can then be written as

$$\min_{\substack{\{R_{ki}\},\{P_{ki}\},\{\gamma_k\},\\ \{\tau_i\},\{\mathcal{S}_i\}}} \sum_k \left(\left(\sum_i \tau_i P_{ki} \right) + E_{\mathrm{loc}_k}(\gamma_k) \right) \tag{4.7a}$$

$$\text{s.t.} \quad \gamma_k \in \{0,1\} \text{ or } \gamma_k \in [0,1], \quad \forall k, \tag{4.7b}$$

$$\tau_i \geq 0, \quad \forall i, \tag{4.7c}$$

$$\sum_i \tau_i R_{ki} = \gamma_k B_k, \quad \forall k, \tag{4.7d}$$

$$T_s \sum_{i=1}^{\bar{\imath}_k} \tau_i \leq \tilde{L}_k, \quad \forall k, \tag{4.7e}$$

$$t_{\mathrm{loc}_k} \leq L_k, \quad \forall k, \tag{4.7f}$$

$$P_{ki} \geq 0, \quad \forall k \in \mathcal{S}_i, \forall i, \tag{4.7g}$$

$$P_{ki} = 0, \quad \forall k \notin \mathcal{S}_i, \forall i \tag{4.7h}$$

$$\{R_{ki}\}_{k \in \mathcal{S}_i} \in \mathcal{R}_{\mathcal{S}_i}\left(\{P_{ki}\}_{k \in \mathcal{S}_i}\right), \quad \forall i, \tag{4.7i}$$

where $\tilde{L}_k = L_k - t_{\mathrm{exe}} - t_{\mathrm{DL}}$ is the latency constraint on the uplink, $\mathcal{R}_{\mathcal{S}_i}\left(\{P_{ki}\}_{k \in \mathcal{S}_i}\right)$ denotes the achievable rate region of the multiple access scheme used in the i^{th} time slot (which ensures that the rates of the offloading users are non-negative), the choice of the constraints in (4.7b) depends on whether binary or partial offloading is performed, and the combination of (4.7h) and (4.7i) implicitly sets $R_{ki} = 0, \forall k \notin \mathcal{S}_i$.

In the problem in (4.7) there are $(2^K - 1)$ different time slots, and hence there are $(2^K - 1)!$ choices for the set of transmitting sets $\{\mathcal{S}_i\}$. For each choice, there are $\binom{K}{m}$ transmitting sets that contain m users, and hence there are $K + \sum_{m=1}^{K} \binom{K}{m}(2m+1)$ remaining design variables in (4.7). While those numbers may appear to preclude an efficient solution, we will show that for a FullMA scheme the problem in (4.7) can be reduced to a K-dimensional problem with a structure that is amenable to efficient

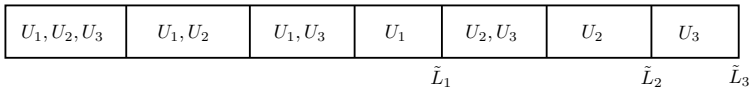


Figure 4.1: Latency-sorted time-slotted structure for a 3-user system.

algorithms. As the first step in that direction, in the following section we will derive an optimized time-slotted structure for a FullMA scheme and we will show that only K time slots are required to obtain the minimum energy consumption in that scheme. We will also determine the optimal transmitting sets $\mathcal{S}_i$ and the optimal durations for each time slot.

4.4.2 Optimized Time-slotted Structure for a FullMA Scheme

Without loss of generality, let us order the users according to their latency constraints, so that

$$\tilde{L}_1 \leq \tilde{L}_2 \leq \cdots \leq \tilde{L}_K, \tag{4.8}$$

where $\tilde{L}_k$ was defined after (4.7). In order for the latency constraint of user k to be met, all the time slots in which that user transmits must occur before its latency limit, $\tilde{L}_k$. Other than that, those time slots can be arranged in an arbitrary order. One particular order is that in which all the time slots involving the first user are grouped at the beginning, and within that group the slots are ordered in non-increasing order of the number of transmitting users. The remaining time slots that involve user 2 are then grouped together and ordered analogously. This procedure is continued until there is a single remaining time slot for user K. This ordering is illustrated in Fig. 4.1 for the case of $K = 3$ users.

Although the generic structure has $(2^K - 1)$ time slots, we will now show that if a FullMA scheme is used in each time slot, the number of time slots can be reduced

to K without loss of generality. To do so, we first observe that if we assume that the time slot lengths are long enough, the achievable rate region for a FullMA scheme approaches the capacity region.[2] If $\mathcal{N}$ denotes an arbitrary subset of $\mathcal{S}_i$, the capacity region corresponding to the i^{th} time slot is the region bounded by constraints of the form (El-Gamal and Cover, 1980)

$$0 \leq \sum_{k\in\mathcal{N}} R_{ki} \leq \log\big(1+\sum_{k\in\mathcal{N}} \alpha_k P_{ki}\big), \quad \forall \mathcal{N} \subseteq \mathcal{S}_i. \tag{4.9}$$

In the case of $K = 2$ users, we were able to exploit the structure of this region to reduce the number of time slots for the FullMA case from the generic 3 to 2 (Salmani and Davidson, 2018b), and in the case of $K = 3$ users we were able to reduce the number of slots from 7 to 3 (Salmani and Davidson, 2019a). To extend those results to the K-user case, we will consider the arrangement of the time slots described in the previous paragraph and illustrated in Fig. 4.1 for $K = 3$. In Appendix 4.A, we show that any two time slots in which there are m users common to both slots and one user different in each slot can be reduced to a single time slot of length equal to the sum of the lengths, over which all $(m + 2)$ users transmit, while the total energy consumption remains the same. By applying that time slot reduction in a sequential manner to the aforementioned time slot arrangement, we can show that under a FullMA scheme the optimal energy consumption obtained using the generic number of time slots, $(2^K - 1)$, can be obtained using the optimized time-slotted structure with only K time slots that is illustrated in Fig. 4.2. For later convenience we will denote the transmitting set for the i^{th} of these K time slots as $\mathcal{S}_i^\star, i = 1, 2, \ldots, K$;

[2] Adaptations to the finite block length regime can be based on MolavianJazi and Laneman (2012, 2013).

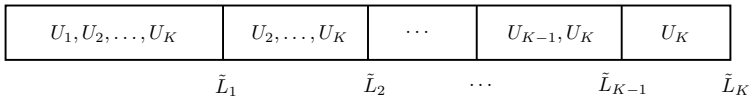


Figure 4.2: Optimized time-slotted structure for a K-user FullMA system.

i.e., $\mathcal{S}_i^\star = \{1, 2, \ldots, K-i+1\}$.

Using similar insights to those used in Appendix 4.A, it can be shown that the optimal energy consumption of each user in a FullMA scheme is obtained when it utilizes its maximum allowable latency. Accordingly, the optimal slot durations are

$$\tau_i^\star = \frac{\tilde{L}_i - \tilde{L}_{i-1}}{T_s}, \quad \text{where} \quad \tilde{L}_0 = 0. \tag{4.10}$$

4.5 Binary Offloading

In this section, we will tackle the problem of minimizing the total energy consumption of a K-user system when the computational tasks of the users are indivisible, i.e., each task must be either totally offloaded to the access point or locally executed by the user. Since the offloading decisions $\gamma_k \in \{0, 1\}$ are binary, the objective of the energy minimization problem in (4.7) can be written as

$$\sum_k \big(\gamma_k (\sum_i E_{\text{off}_{ki}}) + (1-\gamma_k) \underline{E}_{\text{loc}_k}\big), \tag{4.11}$$

where $E_{\text{off}_{ki}}$ is given in (4.2). This expression emphasizes the "mixed" structure of the problem. It consists of the joint optimization over the binary offloading decisions, γ_k, and the continuous rate and power allocations, $\{R_{ki}\}$ and $\{P_{ki}\}$. There are two basic classes of algorithmic strategies that can be taken to address the combinatorial structure of this problem. The first class involves a decomposition of the problem into

an outer problem over the offloading decisions and an inner problem over the rates and powers for a given set of offloading decisions. We will call that inner problem the "complete offloading" problem and we will develop an efficient algorithm for that problem in the next section. (That efficient algorithm is based on further decompositions of the complete offloading problem.) The outer problem over the offloading decisions has a tree structure, and hence is amenable to a variety of optimal and lower-complexity tree search algorithms. In Section 4.5.2 we will develop a tailored pruned greedy search algorithm to find the set of offloading users that results in a close-to-optimal energy consumption of the system.

The second class of algorithmic strategies to address the combinatorial structure of the problem is to relax the binary constraints to $\gamma_k \in [0, 1]$, solve the resulting partial offloading problem, round (or randomized round) the fractional decisions to $\{0, 1\}$, and then solve the complete offloading problem for those decisions (see Wang *et al.*, 2018b; Salmani and Davidson, 2018c; Bi and Zhang, 2018, and also Chapter 3). In Chapter 3 (see also Salmani and Davidson, 2018c), it was shown that for the single-time-slot case the decomposition-tree-search strategy produced solutions with significantly lower energy consumption than the relaxation-rounding strategy and did so at significantly lower computational cost. For that reason we will focus on the decomposition-tree-search strategy in this chapter. We will derive an efficient algorithm for the (inner) complete offloading problem in the next section, and we will describe the (outer) customized tree search algorithm in Section 4.5.2.

4.5.1 Complete Computation Offloading

For a given set of offloading decisions, let $\mathcal{S}_{\text{off}} \subseteq \mathcal{S}$, with $|\mathcal{S}_{\text{off}}| = K_{\text{o}}$, denote the subset of users that is scheduled to fully offload their tasks, (i.e., $\gamma_k = 1, \forall k \in \mathcal{S}_{\text{off}}$ and $\gamma_k = 0, \forall k \notin \mathcal{S}_{\text{off}}$). Since we are employing a FullMA scheme, the results in Section 4.4.2 show that only K_{o} time slots need to be considered. For notational convenience we will renumber the users so that the offloading users are indexed by $k = 1, 2, \ldots, K_{\text{o}}$, in the order indicated by (4.8). From Section 4.4.2, we have an optimal set of transmitting sets $\{\mathcal{S}_i^\star\}_{i=1}^{K_{\text{o}}}$ and an optimal set of slot durations $\{\tau_i^\star\}_{i=1}^{K_{\text{o}}}$. Hence, the fraction of the task description that user k offloads in time slot i is determined by the choice of the rate as $\gamma_{ki}^\star = \frac{\tau_i^\star R_{ki}}{B_k}$, see (4.5). Therefore, the reduced-dimension optimization problem of minimizing the total energy consumption of a given set of offloading users in the complete offloading case can be written as

$$\min_{\{R_{ki}\},\{P_{ki}\}} \sum_{k=1}^{K_o} \sum_{i=1}^{k} \tau_i^\star P_{ki} \tag{4.12a}$$

$$\text{s.t.} \quad \sum_{i=1}^{k} \tau_i^\star R_{ki} = B_k, \quad \forall k \in \mathcal{S}_{\text{off}}, \tag{4.12b}$$

$$0 \leq \sum_{k\in\mathcal{N}} R_{ki} \leq \log_2\big(1 + \sum_{k\in\mathcal{N}} \alpha_k P_{ki}\big), \quad \forall \mathcal{N} \subseteq \mathcal{S}_i^\star, \forall i, \tag{4.12c}$$

where the constraints $P_{ki} \geq 0$ in (4.7g) are implicit in (4.12c), and we have left the constraint $P_{ki} = 0,\ \forall k \notin \mathcal{S}_i^\star, \forall i$ implicit, as the optimal solution to (4.12) will satisfy that constraint.

By employing the insights developed in Chapters 2 and 3 (see also Salmani and Davidson, 2018b,c), our first step in solving (4.12) is to decompose the problem into an inner optimization problem over the transmission powers and an outer rate allocation

problem, as follows

$$\begin{array}{ll} \min\limits_{\{R_{ki}\}} & \min\limits_{\{P_{ki}\}} \ (4.12a) \\ \text{s.t.} \quad (4.12b), (4.12c\text{-LHS}) & \text{s.t.} \quad (4.12c\text{-RHS}), \end{array} \tag{4.13}$$

where we have partitioned the left hand side (LHS) and the right hand side (RHS) of (4.12c). For a given set of transmission rates, the inner optimization problem is a linear programme over $\{P_{ki}\}$, and hence has a polyhedral feasibility region. An optimal solution could be obtained by searching over the vertices of the feasibility region. However, we will now show that by exploiting the polymatroid structure of the constraints in (4.12c) (e.g., Tse and Hanly, 1998), the number of candidate vertices can significantly be reduced. In fact, closed-form solutions for the powers can be obtained.

Closed-form Solutions for Transmission Powers

For a given set of rates, the inner optimization problem in (4.13) is

$$\min_{\{P_{ki}\}_{k=1,i=1}^{K_o,\ k}} \quad \textstyle\sum_{k=1}^{K_o} \sum_{i=1}^{k} \tau_i^\star P_{ki} \tag{4.14a}$$

$$\text{s.t.} \quad \textstyle\sum_{k\in\mathcal{N}} R_{ki} \leq \log_2\big(1+\sum_{k\in\mathcal{N}} \alpha_k P_{ki}\big), \forall \mathcal{N} \subseteq \mathcal{S}_i^\star, \forall i, \tag{4.14b}$$

where we have made the set of active powers explicit. The achievable rate region constraints in (4.14b) imply that for a given set of transmission rates, the transmission power of each user in time slot i depends only on the rates of the users that are transmitting in that time slot. That implies that the problem in (4.14) consists of K_{o}

decoupled problems, each one over one of the K_o separate time slots. Accordingly, the energy minimization problem in the ℓ^{th} time slot is

$$\min_{\{P_{k\ell}\}_{k=1}^{K_o-\ell+1}} \quad \sum_{k=1}^{K_o-\ell+1} \tau_\ell^\star P_{k\ell} \tag{4.15a}$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{N}} R_{k\ell} \leq \log_2\big(1+\sum_{k\in\mathcal{N}} \alpha_k P_{k\ell}\big), \forall \mathcal{N} \subseteq \mathcal{S}_\ell^\star. \tag{4.15b}$$

The problem in (4.15) has a similar structure to the corresponding problem for the single-time-slot signalling structure, for which a closed-form expression for the optimal powers was obtained in Chapter 3 (see also Salmani and Davidson, 2018a,c). To adapt that solution to (4.15), the offloading users in time slot ℓ are sorted according to their channel attenuations $\tilde{\rho}_k = \frac{1}{\alpha_k}$. Let $\pi(\cdot)$ denote a permutation of the users such that

$$\tilde{\rho}_{\pi(1)} \geq \tilde{\rho}_{\pi(2)} \geq \cdots \geq \tilde{\rho}_{\pi(K_o-1)} \geq \tilde{\rho}_{\pi(K_o)}, \tag{4.16}$$

and let $\pi^{-1}(\cdot)$ denote the inverse permutation; that is for user k_o there are $\pi^{-1}(k_o)-1$ users with values of $\tilde{\rho}_k$ that are larger than $\tilde{\rho}_{k_o}$. Exploiting the polymatroid structure of the feasibility region constraints in (4.15b), the closed-form optimal solutions for the power of each user in terms of its rate and the rates of the other users in time slot ℓ can be obtained as (Salmani and Davidson, 2018a,c, and Chapter 3)

$$P_{k\ell} = \big(\tfrac{2^{R_{k\ell}}-1}{\alpha_k}\big) 2^{\left(\sum_{j=1}^{\pi^{-1}(k)-1} R_{\pi(j)\ell}\right)}. \tag{4.17}$$

Solutions for Transmission Rates

Now by substituting these closed-form expressions for the powers in terms of the rates into (4.13), and hence reducing the problem dimension, the outer problem in (4.13)

becomes

$$\min_{\{R_{ki}\}_{k=1,i=1}^{K_o,\ k}} \quad \sum_{k=1}^{K_o}\sum_{i=1}^{k}\tau_i^\star\big(\tfrac{2^{R_{ki}}-1}{\alpha_k}\big)2^{\left(\sum_{j=1}^{\pi^{-1}(k)-1}R_{\pi(j)\ell}\right)} \tag{4.18a}$$

$$\text{s.t.} \qquad R_{ki}\geq 0 \quad \forall k,i, \tag{4.18b}$$

$$\sum_{i=1}^{k}\tau_i^\star R_{ki}=B_k, \quad \forall k. \tag{4.18c}$$

The problem in (4.18) is difficult to tackle directly because of the non-convex objective function and the large number of variables. (There are $K_o(K_o+1)/2$ rates to be determined.) To develop a strategy for solving (4.18) we observe that the constraints on the transmission rates of different users are separable, see (4.18b) and (4.18c). Moreover, we can show that the objective function is convex in the rates of a given user, say user n, when the rates of all other users in all time slots are fixed. These observations suggest that the problem in (4.18) could be tackled using a block coordinate descent approach (e.g., Hong *et al.*, 2016), in which at each step the rates of one user are updated. The subproblem of finding the optimal rates of user n, when the rates of all the other users are given is

$$\min_{\{R_{ni}\}_{i=1}^{n}} \quad \sum_{i=1}^{n}\tau_i^\star A_{ni}\ 2^{R_{ni}} \tag{4.19a}$$

$$\text{s.t.} \qquad R_{ni}\geq 0, \quad \forall i, \tag{4.19b}$$

$$\sum_{i=1}^{n}\tau_i^\star R_{ni}=B_n, \tag{4.19c}$$

where the permutation π was defined in (4.16), and

$$A_{ni}=\big(\tfrac{1}{\alpha_n}\big)\ 2^{\left(\sum_{j=1}^{\pi^{-1}(n)-1}R_{\pi(j)i}\right)}+\sum_{j=\pi^{-1}(n)+1}^{K}\big(\tfrac{2^{R_{\pi(j)i}}-1}{\alpha_{\pi(j)}}\big)\ 2^{\left(\sum_{m=1,\ m\neq n}^{j}R_{\pi(m)i}\right)}. \tag{4.20}$$

The problem in (4.19) is not only convex, it has a structure that makes it amenable to very efficient algorithms. In particular, the optimal solution can be efficiently obtained by applying a generalized water-filling approach, such as that of Xing *et al.* (2018). Furthermore, the convexity of (4.19) guarantees that the block coordinate descent algorithm will obtain a stationary solution of the problem in (4.18) (see Hong *et al.*, 2016). Once the transmission rates of all the users are obtained, the corresponding values for the transmission powers can be achieved by substituting the obtained solutions for the transmission rates in (4.17).

4.5.2 Binary Computation Offloading

In Section 4.5.1 we have obtained a solution for the resource allocation problem for a given set of binary offloading users. Now, we seek to find the set of offloading users that minimizes the total user energy consumption while the constraints on the latencies and achievable rate region are met. The tree structure of that combinatorial optimization problem, which has a search space of 2^K possibilities, suggests that tree-search strategies can be employed to find the optimal (or a good suboptimal) set of offloading users. We have developed a low-complexity pruned greedy search algorithm in Chapter 3 (see also Salmani and Davidson, 2018c) to tackle the binary offloading problem for a single-time-slot binary offloading system. Since the signalling structure only affects the structure of the complete offloading problem, and not the searching strategy, the principles of that algorithm can be applied to find an appropriate set of offloading users in the time-slotted binary offloading system, as well. That algorithm begins with an initial state in which all the users that cannot complete their whole tasks locally within their latency constraints are scheduled to offload their tasks,

and the other users are scheduled to execute their tasks locally. At each iteration, the algorithm seeks to add one user to the offloading set in a greedy manner so that by adding that user to the offloading set, the total energy consumption of the users at that iteration is reduced by the largest amount (see Step 5 in Algorithm 5). Furthermore, in order to reduce the search complexity of the algorithm, all of those users for which offloading their tasks (in the current iteration) would impose a larger energy consumption to the set of offloading users as a whole than executing their tasks locally, are pruned from the search space for the next iteration (see Steps 2 and 3 in Algorithm 5). The rationale for pruning the tree in this way is that if offloading cannot help a user to reduce its energy consumption at the current iteration, it won't help that user to do so in the subsequent iterations, where more users are scheduled to offload and hence there is higher interference in the system. The steps of the resulting algorithm are summarized in Algorithm 5.

4.6 Partial offloading

If the computational tasks can be divided into components that can be completed separately, the offloading system can benefit from the computational parallelism between the access point and the devices to reduce the total energy consumption. In this section, we will focus on "data-partitionable" tasks (Wang *et al.*, 2016) in which multiple blocks of data are processed independently, using a simple-to-describe computational operation. In that case, the number of computational operations required to complete (a fraction of) the task can be modeled as being a function of the description length (Wang *et al.*, 2016; Muñoz *et al.*, 2015; Zhang *et al.*, 2013).

In the partial offloading case, the energy that each user expends to complete its

Algorithm 5 : Binary Offloading Solution

Input data: $\{B_k\}$, $\{L_k\}$, $\{\tilde{L}_k\}$, $\{\alpha_k\}$, $\{\underline{E}_{\text{loc}_k}\}$, $\{t_{\text{loc}_k}\}$, T_s.
Step 1: Initialize the offloading set to $\mathcal{S}_{\text{off}} = \{k | t_{\text{loc}_k} > L_k\}$, and the undecided set to $\mathcal{U} = \{1, 2, \ldots, K\} \setminus \mathcal{S}_{\text{off}}$. Set $j = 0$.
Step 2: Solve the complete offloading problem in (4.12) for $\mathcal{S}_{\text{off}}$ and let $E_{\text{off}}^{(0)}$ denote the corresponding total user energy consumption.
Step 2: Set the pruning set $\mathcal{V} = \emptyset$ and $j \leftarrow j + 1$.
for each $k \in \mathcal{U}$ **do**
Obtain the energy consumption of the system when user k is added to the set of offloading users, $E_{\text{total}_k}^{(j)}$, by solving (4.12).
if $E_{\text{off}}^{(j-1)} + \underline{E}_{\text{loc}_k} \leq E_{\text{total}_k}^{(j)}$ **then**
Add user k to the set of users to be pruned; $\mathcal{V} \leftarrow \mathcal{V} \cup \{k\}$
end if
end for
Step 3: Prune the selected users from the tree; $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{V}$.
Step 4: If $\mathcal{U} = \emptyset$, terminate the algorithm.
Step 5: Select the "best" user by choosing
$k^\star = \arg\max_{k \in \mathcal{U}} \big(E_{\text{off}}^{(j-1)} + \underline{E}_{\text{loc}_k} - E_{\text{total}_k}^{(j)}\big)$.
Step 6: Update the offloading set and the undecided set; $\mathcal{S}_{\text{off}} \leftarrow \mathcal{S}_{\text{off}} \cup \{k^\star\}$ and $\mathcal{U} \leftarrow \mathcal{U} \setminus \{k^\star\}$.
Step 7: Update the energy consumption; $E_{\text{off}}^{(j)} = E_{\text{total}_{k^\star}}^{(j)}$.
Step 8: If $\mathcal{U} = \emptyset$, stop. If not go to Step 3.

computational task is the sum of the energy used to offload a portion of the task to the access point and the energy used to execute the remaining part of the task locally. It was shown by Wang *et al.* (2016) that the local computation energy in a mobile device can be reduced by employing the dynamic voltage scaling architecture. In that case, the energy that user k expends to locally process the $(1-\gamma_k)B_k$ bits that it retains, within its latency constraint, L_k, was given in (4.4). In order to formulate the problem of minimizing the total user energy consumption of a partial offloading system with K users, we will consider the optimized time-slotted structure depicted in Fig. 4.2, with offloading sets $\{\mathcal{S}_i^\star\}_{i=1}^K$ and slot durations $\{\tau_i^\star\}_{i=1}^K$. By employing the closed-form expressions in (4.5), we obtain the optimal offloading fraction in each time slot as a function of the rate, $\gamma_{ki}^\star = \frac{\tau_i^\star R_{ki}}{B_k}$. Hence, we can show that the optimal number of bits to be offloaded satisfies $\gamma_k^\star B_k = \sum_i \gamma_{ki}^\star B_k = \sum_i \tau_i^\star R_{ki}$. Using that expression, the remaining optimization problem is to allocate the rates and powers to each user in each time slot, namely,

$$\min_{\{R_{ki}\},\{P_{ki}\}} \sum_{k=1}^K \Big(\sum_{i=1}^k \tau_i^\star P_{ki}\Big) + \frac{M_k}{L_k^2}\Big(B_k - \sum_i \tau_i^\star R_{ki}\Big)^3 \tag{4.21a}$$

$$\text{s.t.} \quad 0 \le \sum_i \tau_i^\star R_{ki} \le B_k, \quad \forall k, \tag{4.21b}$$

$$0 \le \sum_{k\in\mathcal{N}} R_{ki} \le \log_2\big(1+\sum_{k\in\mathcal{N}} \alpha_k P_{ki}\big), \forall \mathcal{N} \subseteq \mathcal{S}_i^\star, \forall i, \tag{4.21c}$$

where, as in (4.12), we have left the constraint $P_{ki} = 0, \forall k \notin \mathcal{S}_i^\star, \forall i$, implicit.

As in the binary offloading case, in the first step of solving the problem in (4.21),

we perform the decomposition

$$\begin{array}{ll} \min\limits_{\{R_{ki}\}} & \min\limits_{\{P_{ki}\}} \quad (4.21a) \\ \text{s.t.} \quad (4.21b), (4.21c\text{-LHS}), & \text{s.t.} \quad (4.21c\text{-RHS}), \end{array} \tag{4.22}$$

and since the inner problem has a similar structure to the problem in (4.14), we can obtain closed-form expressions for the transmission powers in terms of the transmission rates.

4.6.1 Closed-form Solutions for Transmission Powers

The achievable rate region constraints applied to the users transmitting in time slot ℓ imply that for a given set of transmission rates, the power of each user in that time slot depends only on the rates of the set of users that are offloading in that specific time slot, $\mathcal{S}_\ell^\star$. Hence, as was the case for the complete offloading, the inner problem in (4.22) can be addressed by solving K optimization problems, each of which minimizes the energy consumption of the offloading users in a given time slot. For time slot ℓ, the optimization problem is

$$\min_{\{P_{k\ell}\}_{k=1}^{K-\ell+1}} \quad \textstyle\sum_{k=1}^{K-\ell+1} \tau_\ell^\star P_{k\ell} \tag{4.23a}$$

$$\text{s.t.} \quad \textstyle\sum_{k\in\mathcal{N}} R_{k\ell} \leq \log_2\big(1+\sum_{k\in\mathcal{N}} \alpha_k P_{k\ell}\big), \forall \mathcal{N} \subseteq \mathcal{S}_\ell^\star. \tag{4.23b}$$

Analogous to the complete offloading case, by exploiting the polymatroid structure of the capacity region of a FullMA scheme and sorting the users according to their

channel attenuations (see (4.16)), a closed-form expression for the optimal transmission power of each user in terms of the rates of all the other users in that time slot can be obtained; see (4.17).

4.6.2 Solutions for Transmission Rates

Having derived the optimal powers, the outer problem is

$$\min_{\{R_{ki}\}_{k=1,i=1}^{K_o,\,k}} \quad \sum_{k=1}^{K}\sum_{i=1}^{k}\tau_i^\star\big(\tfrac{2^{R_{ki}}-1}{\alpha_k}\big)2^{\left(\sum_{j=1}^{\pi^{-1}(k)-1}R_{\pi(j)i}\right)}+\tfrac{M_k}{L_k^2}\big(B_k-\sum_i\tau_i^\star R_{ki}\big)^3 \tag{4.24a}$$

$$\text{s.t.} \quad R_{ki}\geq 0, \quad \forall k,i, \tag{4.24b}$$

$$\sum_{i=1}^{k}\tau_i^\star R_{ki}\leq B_k, \quad \forall k. \tag{4.24c}$$

It can be seen from (4.24b) and (4.24c) that the decomposition that we have chosen has decoupled the constraints on the rates of each user from the rates of the other users. That suggests the use of the block coordinate descent technique. Indeed, if we fix the set of rates of all the users except user n and employ the definition of A_{ni} in (4.20), the transmission rates of user n in its offloading time slots can be obtained by solving the following problem

$$\min_{\{R_{ni}\}_{i=1}^{n}} \quad \sum_{i=1}^{n}\tau_i^\star A_{ni}\,2^{R_{ni}}+\tfrac{M_n}{L_n^2}\big(B_n-\sum_{i=1}^{n}\tau_i^\star R_{ni}\big)^3 \tag{4.25a}$$

$$\text{s.t.} \quad R_{ni}\geq 0, \quad \forall i, \tag{4.25b}$$

$$\sum_{i=1}^{n}\tau_i^\star R_{ni}\leq B_n. \tag{4.25c}$$

Single-variable problem to find transmission rate of each user

The structure of the problem in (4.25) is similar to the corresponding problem in the complete offloading case, c.f. (4.19), except for the fact that there is an extra local energy consumption term in the objective function in (4.25), and the constraint in (4.25c) is an inequality constraint. In general, due to the local energy consumption term, the objective function in (4.25) is not convex. However, we can tackle that problem by showing that the objective function is convex when the fraction of offloaded bits is fixed. To do so, we decompose the problem in (4.25) into an inner optimization problem that determines the transmission rates of user n required to offload a given fraction of offloading bits, $\tilde{B}_n$, and an outer optimization problem that obtains the optimal value of $\tilde{B}_n$, namely,

$$\begin{array}{ll} \min\limits_{\tilde{B}_n} & \min\limits_{\{R_{ni}\}_{i=1}^{n}} \ \sum_{i=1}^{n} \tau_i^\star A_{ni}\ 2^{R_{ni}} \\ \text{s.t.} \quad \tilde{B}_n \leq B_n & \text{s.t.} \ \ R_{ni} \geq 0,\ \forall i,\ \sum_{i=1}^{n} \tau_i^\star R_{ni} = \tilde{B}_n. \end{array} \tag{4.26}$$

The inner optimization problem in (4.26) is a convex problem for which Slater's condition (e.g., Boyd and Vandenberghe, 2004) holds. Accordingly, if we write the Lagrangian function of that problem as

$$\mathcal{L} = \textstyle\sum_{i=1}^{n} \tau_i^\star A_{ni}\ 2^{R_{ni}} - \lambda_n R_{ni} + \nu\big(\sum_{i=1}^{n} \tau_i^\star R_{ni} - \tilde{B}_n\big),$$

the KKT conditions indicate that if R_{ni} is the optimal solution, then $\lambda_n^\star R_{ni}^\star = 0$ and

$$\tfrac{\partial \mathcal{L}}{\partial R_{ni}} = \ln(2)\tau_i^\star A_{ni}\ 2^{R_{ni}} - \lambda_n + \nu\tau_i^\star = 0. \tag{4.27}$$

Hence, the optimal $R_{\pi(n)i}$ can be obtained as

$$\begin{cases} R^{\star}_{ni} = \log_2\big(\frac{-\nu}{\ln(2)A_{ni}}\big), & \text{if } \lambda^{\star}_n = 0, \\ R^{\star}_{ni} = 0, & \text{if } \lambda^{\star}_n > 0. \end{cases} \tag{4.28}$$

By writing these optimal values of the rates as $R^{\star}_{ni} = \big[\log_2\big(\frac{-\nu}{\ln(2)A_{ni}}\big)\big]^+$, where $[x]^+ = \max\{0, x\}$, and substituting them into the problem in (4.25), that problem is reduced to the following single-variable optimization problem

$$\min_{\nu} \quad \textstyle\sum_{i=1}^{n} \tau^{\star}_i A_{ni}\, 2^{\left[\log_2\left(\frac{-\nu}{\ln(2)A_{ni}}\right)\right]^+} + \frac{M_n}{L_n^2}\big(B_n - \sum_{i=1}^{n} \tau^{\star}_i \big[\log_2(\frac{-\nu}{\ln(2)A_{ni}})\big]^+\big)^3 \tag{4.29a}$$

$$\text{s.t.} \quad \textstyle\sum_{i=1}^{n} \tau^{\star}_i \big[\log_2(\frac{-\nu}{\ln(2)A_{ni}})\big]^+ \leq B_n. \tag{4.29b}$$

To gain insight into the structure of the problem in (4.29) we observe that the domain of the logarithm requires that ν be negative. We also observe that as ν decreases from 0^-, the transmission rate (of user n) in the i^{th} slot becomes positive when $\nu < -A_{ni}\ln(2)$. Furthermore, between those threshold values of ν both the objective and the constraint in (4.29) are continuous. To simplify our analysis of those piecewise continuous functions, we will reorder the n time slots over which user n may transmit, using the permutation $\tilde{\pi}$, according to the values of A_{ni} so that $A_{n\tilde{\pi}(1)} \leq A_{n\tilde{\pi}(2)} \leq \cdots \leq A_{n\tilde{\pi}(n)}$. For simplicity, we let $\tilde{A}_{n\ell}$ denote the sorted values of $A_{n\tilde{\pi}(\ell)}$, i.e., $\tilde{A}_{n1} \leq \tilde{A}_{n2} \leq \cdots \leq \tilde{A}_{nn}$. We let $\tilde{\tau}^{\star}_{\ell}$ denote the corresponding sorting of the time slots. With those observations and definitions in place, we can rewrite the

problem in (4.29) in the form of $\min_\nu f(\nu)$ subject to (4.29b), where

$$f(\nu) = \begin{cases} \sum_{j=1}^{n} \tilde{\tau}_j^\star \tilde{A}_{nj} + \frac{M_n}{L_n^2} B_n^3, & \nu \in (-\ln(2)\tilde{A}_{n1}, 0], \\ f_\ell(\nu), & \nu \in \Gamma_\ell,\ \ell = 1, \ldots, n-1, \\ f_n(\nu), & \nu \in (-\infty, -\ln(2)\tilde{A}_{nn}], \end{cases} \tag{4.30}$$

where $\Gamma_\ell = (-\ln(2)\tilde{A}_{n(\ell+1)}, -\ln(2)\tilde{A}_{n\ell}]$, and

$$f_\ell(\nu) = \tfrac{-\nu}{\ln(2)} \textstyle\sum_{j=1}^{\ell} \tilde{\tau}_j^\star + \frac{M_n}{L_n^2} \big(B_n - \sum_{j=1}^{\ell} \tilde{\tau}_j^\star \log_2(\frac{-\nu}{\ln(2)\tilde{A}_{nj}})\big)^3.$$

Although the constraint in (4.29b) and each $f_\ell(\nu)$ are convex, in the general case $f(\nu)$ is not convex. However, in Appendix 4.B we show that there is at most one value of ℓ for which the point(s) at which the derivative of $f_\ell(\nu)$ is equal to zero lie(s) within the relevant sub-interval in (4.30), and in Appendix 4.C we show that for $\ell = 1, 2, \cdots, n-1$, the derivatives of $f_\ell(\nu)$ and $f_{\ell+1}(\nu)$ have the same sign on either side of $\ln(2)\tilde{A}_\ell$. Those two facts are sufficient to show that $f(\nu)$ is quasi-convex (cf. Boyd and Vandenberghe, 2004), and hence the problem in (4.29) can be solved using one of a number of line search strategies; (e.g., Antoniou and Lu, 2007, Chapter 4). As an alternative, in the next subsection we will exploit the structure of each $f_\ell(\nu)$ to develop a quasi-closed-form expression for the optimal value for ν, and hence the transmission rates of user n given the rates of the other users.

Quasi-closed-form solutions for the transmission rates of each user

To develop a quasi-closed-form solution to (4.29), we use the observation established in Appendix 4.B that there is at most one sub-interval in which the derivative of the

piecewise convex function $f(\nu)$ goes to zero. As explained below, that enables us to sequentially test the points at which $\frac{df_\ell(\nu)}{d\nu}$ goes to zero, for $\ell = 1, 2, \cdots, n$, until we find an ℓ for which a stationary point lies in the relevant sub-interval, $\nu \in \Gamma_\ell$. If there is no such value for ℓ, then one of the end points for ν must be optimal. (The observation in Appendix 4.C regarding the signs of $\frac{df_\ell(\nu)}{d\nu}$ and $\frac{df_{\ell+1}(\nu)}{d\nu}$ on either side of $-\ln(2)\tilde{A}_{n(\ell+1)}$ and the convexity of $f_\ell(\nu)$ can be used to reduce the number of sub-intervals that need to be examined.)

The key observation in this technique is that the stationary points of $f_\ell(\nu)$ can be found analytically. Indeed, the points at which $\frac{df_\ell(\nu)}{d\nu} = 0$ satisfy

$$-\nu_\ell = \big(\beta_1 - \beta_2 \log_2(-\nu_\ell)\big)^2, \tag{4.31}$$

where $\beta_1 = \sqrt{\frac{3M_n}{L_n^2}}\big(B_n + \sum_{j=1}^{\ell} \tilde{\tau}_j^\star \log_2(\ln(2)\tilde{A}_{nj})\big)$, and $\beta_2 = \sqrt{\frac{3M_n}{L_n^2}}\big(\sum_{j=1}^{\ell} \tilde{\tau}_j^\star\big)$. A closed-form expression for that value of ν can then be obtained using the Lambert function $W(\cdot)$ (Corless *et al.*, 1996),

$$\nu_\ell = \frac{2W\big(\frac{\ln(2)}{2\beta_2}\, 2^{\frac{\beta_1}{2\beta_2}}\big)}{\ln(2)} + \frac{\beta_1}{\beta_2}. \tag{4.32}$$

If the obtained ν_ℓ lies within the sub-interval corresponding to $f_\ell(\nu)$ and if it satisfies the constraint in (4.29b), then ν_ℓ is the optimal solution of the problem in (4.29). Otherwise, we will examine $f_{\ell+1}(\nu)$ to determine if the value of ν at which its derivative goes to zero can be found within the corresponding sub-interval. If, after examining over all the sub-intervals we cannot find such a ν, then the optimal value for ν is either $\nu = 0$ or the most negative ν for which the inequality in (4.29b) holds with

equality, namely,

$$\nu_{\text{low}} = -2^{\frac{B_n - \sum_{j=1}^{n} \tilde{\tau}_j^{\star} \log_2(\frac{1}{\ln(2)\tilde{A}_{nj}})}{\sum_{j=1}^{n} \tilde{\tau}_j^{\star}}}. \tag{4.33}$$

The steps to find a closed-form expression for the optimal ν are summarized in Algorithm 6. Having obtained the optimal solution for ν, we can find a closed-form solution for the transmission rate of user n in each time slot, given all the other rates, by substituting the optimal ν in (4.28).

Stationary solutions for the transmission rates of all users

Now that we have developed an algorithm to find the rates of each of the users, when the rates of all other rates are fixed, we can apply a coordinate descent algorithm. Since the objective function of each of the constituent sub-problems is quasi-convex in terms of its corresponding ν (see (4.29)), by considering the KKT conditions in (4.28), we can show that the block coordinate descent algorithm is guaranteed to obtain a stationary solution for (4.24); (cf. Hong *et al.*, 2016, Theorem 1). The transmission powers of each user in each time slot can then be obtained using (4.17).

4.7 Numerical Results

We now illustrate the performance of the proposed time-slotted FullMA signalling structure in both binary offloading and partial offloading scenarios. We compare its performance with several existing approaches from the literature. To highlight the impact of exploiting the full capabilities of the channel on the user energy consumption, we make comparisons with the TDMA-based algorithm in Chapter 3 (see also Salmani and Davidson, 2018c), which is inherently time-slotted. To assess the energy

Algorithm 6 : The optimal solution of the problem in (4.29)

Input data: values of B_n, L_n, M_n, $\{\tilde{\tau}_i^\star\}_{i=1}^n$, $\{\tilde{A}_{ni}\}_{i=1}^n$.
Step 1: Set $\ell = 1$.
while $\ell \leq n$ **do**
  Step 2: Calculate ν_ℓ using (4.32).
  if ν_ℓ satisfies (4.29b), and $-\ln(2)\tilde{A}_{n(\ell+1)} \leq \nu < -\ln(2)\tilde{A}_{n\ell}$ **then**
    $\nu^\star = \nu_\ell$, and terminate the algorithm.
  else $\ell = \ell + 1$.
  end if
end while
if $\ell > n$ **then**
  Calculate (4.29a) for $\nu = 0$, and $\nu = \nu_{\text{low}}$ in (4.33) and choose the one with the minimum objective value as $\nu^\star$.
end if

reduction that results from exploiting the maximum allowable latency of each user and exploiting the differences between the latency constraints of the users, we compare the proposed time-slotted FullMA (TS-FullMA) signalling scheme with single-time-slot FullMA (STS-FullMA) in Chapter 3 (see also Salmani and Davidson, 2018c) and equal-latency FullMA (EL-FullMA) (Wang *et al.*, 2018b). The computation efficiency of the proposed TS-FullMA resource allocation algorithms is also illustrated via a comparison with those FullMA-based algorithms. We consider an offloading system with K users that are uniformly distributed over a cell of radius 1,000m. The channel between each user and the access point is modeled as a slowly fading channel with a path-loss exponent of 3.7 and independent identically distributed Rayleigh small-scale fading. The symbol interval of the channel is $T_s = 10^{-6}$s, and the receiver noise variance is $\sigma^2 = 10^{-13}$. The energy consumption in each experiment is averaged over 100 channel realizations. We assume that the sum of the execution time and the time it takes to download the results to the mobile users is equal for all the users, $t_{\text{exe}_k} + t_{\text{DL}_k} = 0.2$s. As explained in Section 4.4.1, in the case of partial offloading, we

consider data-partitionable computational tasks for which the (optimal) local energy consumption can be modeled as a function of number of bits; see (4.4). In order to be consistent with the measurements in (Miettinen and Nurminen, 2010), the constant M_k in the expression in (4.4) is set to be $M_k = 10^{-19},\ \forall k$, (Zhang *et al.*, 2013; Wang *et al.*, 2016). For the case of binary offloading, we simply set $\underline{E}_{\text{loc}_k} = \frac{M_k}{L_k^2} B_k^3$ in (4.11), and we assume that each user can complete its task locally within its latency constraint; i.e., $t_{\text{loc}_k} \leq L_k,\ \forall k$.

4.7.1 Binary Computation Offloading

In this section, we consider indivisible tasks, and hence each user either offloads its whole task or completes it locally. For our first experiment, we consider an offloading system with four users with the latencies $[L_1, L_2, L_3, L_4] = [1.1, 1.2, 2.1, 2.2]$s, and we examine the energy consumption as the (different) description lengths of the tasks grow (in proportion) $[B_1, B_2, B_3, B_4] = \zeta \times [2, 1, 3, 4] \times 10^6$ bits.

Fig. 4.3 plots the average energy consumption of the users as ζ increases, and it can be seen that the proposed TS-FullMA structure consumes less energy than the existing structures. (For TS-FullMA and STS-FullMA we have included results for an exhaustive search over the offloading decisions, in addition to the customized greedy search outlined in Algorithm 5.) Fig. 4.3 highlights the fact that exploiting the interference reduction that results when a user completes its offloading, enables TS-FullMA to significantly reduce the energy consumption as compared to STS-FullMA in Chapter 3 (see also Salmani and Davidson, 2018c). Moreover, exploiting the full capabilities of the channel by employing a FullMA scheme significantly reduces the

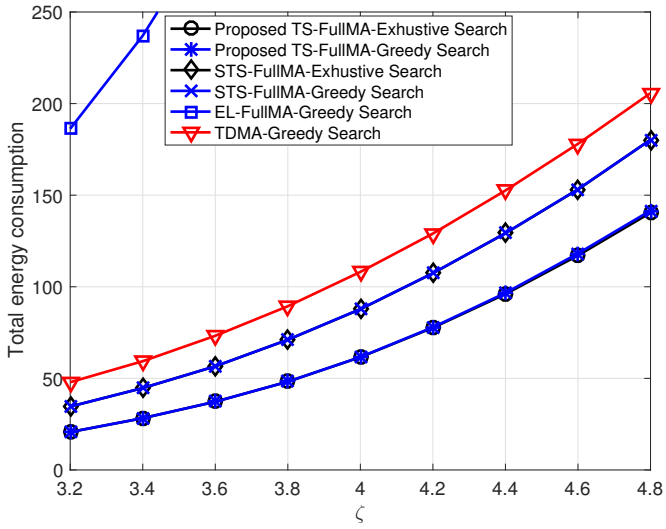


Figure 4.3: Average energy consumption of a binary offloading system with four users with different latency constraints versus the parameter that defines the number of bits required to describe the users' tasks, where STS-FullMA is the single-time-slotted approach introduced in Chapter 3, and EL-FullMA is the equal-latency FullMA-based approach proposed by Wang *et al.* (2018b).

energy consumption over TDMA. Finally, Fig. 4.3 shows that the energy consumptions of the schemes that utilize the maximum allowable latencies of the users (the proposed TS-FullMA scheme, the TDMA and STS-FullMA schemes in Chapter 3 (see also Salmani and Davidson, 2018c)) are significantly less than that of the EL-FullMA scheme in which all the users are constrained to operate as if they have equal latency constraints (Wang *et al.*, 2018b). (The energy consumption of the equal-latency TDMA system of You *et al.* (2017) is, quite naturally, larger than that of the EL-FullMA system.)

In order to compare the computational cost of the proposed TS-FullMA resource allocation algorithm with the algorithms for the existing FullMA systems, we provide their computational costs (in terms of the average CPU times) in Table 4.1. Since

those CPU times are (essentially) independent from the description length of the tasks, we provide the average CPU time over the values of ζ. All the algorithms were coded in MATLAB, with similar diligence paid to the efficiency of the programs. The convex optimization subproblems in the EL-FullMA system (Wang *et al.*, 2018b) were solved using SDPT3 (Toh *et al.*, 1999) through the CVX interface (Grant *et al.*, 2008). (As suggested by Wang *et al.* (2018b), for the ellipsoid method we employed the approach of Boyd (2018), and we chose a termination criterion of $\epsilon = 10^{-3}$.) The CPU times were evaluated on a Mac mini with a Core i7 processor running at 3.2GHz, and 16GB of RAM. It can be seen that the computational cost of the proposed TS-FullMA algorithm is significantly lower than that for EL-FullMA. That is due to the closed-form solutions that we obtained for a large fraction of the variables in our proposed algorithm. The algorithm for EL-FullMA employs a variant of the ellipsoidal algorithm. Table 4.1 also shows that the computational cost of the proposed TS-FullMA algorithm is somewhat higher than that for STS-FullMA. The main reason is that in the proposed structure, there are multiple time slots and in each time slot a different transmission rate and transmission power are assigned to each user. Accordingly, the number of variables to be designed is larger than that of the single-time slot structure. The results in Fig. 4.3 and Table 4.1 indicate that the proposed TS-FullMA system outperforms EL-FullMA in terms of the energy consumption and the computational cost, while there is a performance-cost trade-off between the proposed TS-FullMA and STS-FullMA.

In our second numerical experiment for the binary offloading case, we examine the total user energy consumption as the number of users increases. We consider the scenario in which the description length of user k's task is $B_k = (3 + 0.5k) \times 10^6$

Table 4.1: Average CPU time required for resource allocations in the proposed TS-FullMA system, STS-FullMA in Chapter 3 (see also Salmani and Davidson, 2018c), and EL-FullMA (Wang *et al.*, 2018b), for a four-user binary offloading system.

Algorithm	**Average CPU time (sec)**
Proposed TS-FullMA-Greedy Search	4.3×10^{-2}
STS-FullMA (Chapter 3 and (Salmani and Davidson, 2018c))-Greedy Search	1.5×10^{-5}
EL-FullMA (Wang *et al.*, 2018b)-Greedy Search	2.7×10^{5}

bits and its maximum allowable latency is $L_k = 1 + 0.4k$ s. In Figs 4.4 and 4.5, we illustrate the average energy consumption and the average computational costs of the proposed TS-FullMA structure and the other considered methods, respectively. Those figures show that TS-FullMA can substantially reduce the energy consumption as compared to EL-FullMA, and it does so at a much lower computational cost. The results in those figures also indicate that exploiting the differences between the latency constraints of the users and the interference reduction that results when a user completes its offloading (as the proposed system does) can significantly reduce the energy consumption compared to STS-FullMA. However, that energy reduction, which is obtained by allocating a different rate and power to each user in each time slot, imposes a somewhat higher computational cost. Hence, there is, once again, a performance-cost trade-off when it comes to choosing between the proposed TS-FullMA and STS-FullMA.

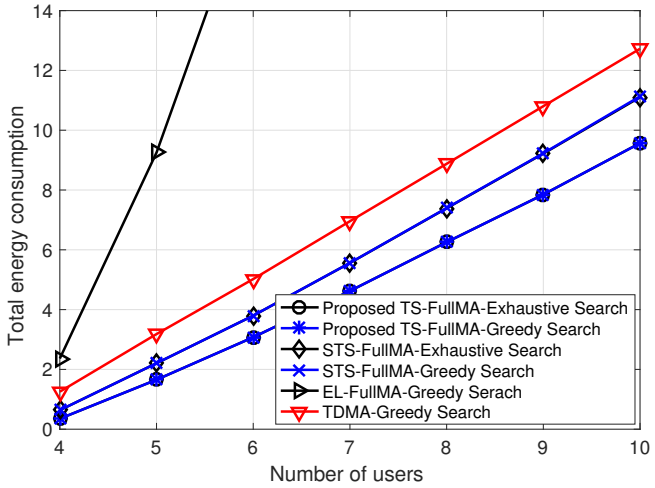


Figure 4.4: Average energy consumption of a binary offloading system. STS-FullMA denotes the single-time-slotted approach introduced in Chapter 3, and EL-FullMA denotes the equal-latency FullMA-based approach proposed by Wang *et al.* (2018b).

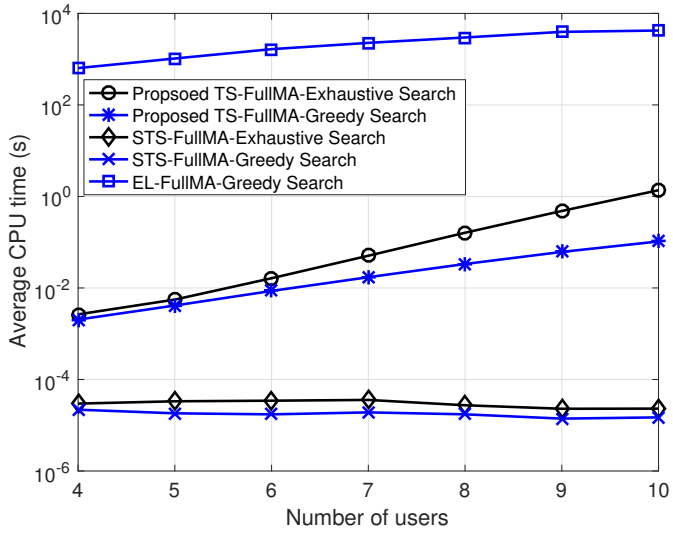


Figure 4.5: Average CPU time required for the proposed TS-FullMA algorithm and the algorithms for STS-FullMA, introduced in Chapter 3, and EL-FullMA, proposed by Wang *et al.* (2018b), for different number of users in binary offloading case.

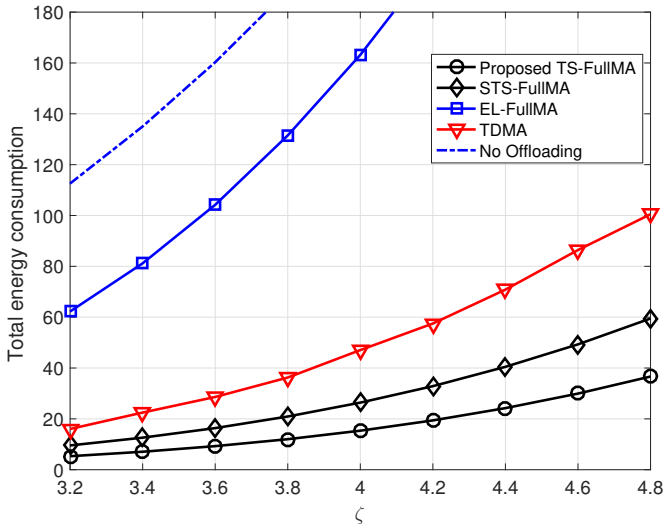


Figure 4.6: Average energy consumption of a four-user partial offloading system with different latency constraints versus the parameter that defines the number of bits required to describe the users' tasks, where STS-FullMA is the single-time-slotted approach introduced in Chapter 3, and EL-FullMA is the equal-latency FullMA-based approach proposed by Wang *et al.* (2018b).

4.7.2 Partial Computation Offloading

In this section, we consider computational tasks that are divisible, and hence each user can offload a portion of its task to the access point while the remaining part is processed locally. We begin with a 4-user offloading system with the same parameters as in the first experiment in Section 4.7.1, where the computational tasks are now assumed to be divisible.

In Fig. 4.6, we present the total energy consumption versus ζ which is proportional to the description lengths of the users' tasks. Similar to the binary offloading case, that figure shows that the energy consumption of TS-FullMA is lower than that of STS-FullMA and TDMA, and significantly lower than that of EL-FullMA. The computational costs of the FullMA systems for the case of the partial offloading are

Table 4.2: Average CPU time required for resource allocations in the proposed TS-FullMA system, STS-FullMA Chapter 3 (see also Salmani and Davidson, 2018c), and EL-FullMA (Wang *et al.*, 2018b), for a four-user partial offloading system.

Algorithm	**Average CPU time (sec)**
Proposed TS-FullMA	1.5
STS-FullMA (Chapter 3 and (Salmani and Davidson, 2018c))	3.4×10^{-3}
EL-FullMA (Wang *et al.*, 2018b)	5.8×10^{2}

shown in Table 4.2. It can be seen that, similar to the case of binary offloading, the computational cost of TS-FullMA is significantly lower than that of EL-FullMA, while there is a performance-cost trade-off in choosing between TS-FullMA and STS-FullMA.

In our final numerical experiment we examine the energy consumption of a partial offloading system as the number of users increases. We consider the scenario in which the description length of user k's task is $B_k = (6 + 0.5k) \times 10^6$ bits and its maximum allowable latency is $L_k = 1 + 0.5k$ s. Fig. 4.7 shows that, as in the binary offloading case, exploiting the full capabilities of the channel together with utilizing the differences between the latency constraints of the users enables TS-FullMA to substantially reduce the total energy consumption as compared to STS-FullMA, EL-FullMA, and TDMA. Fig. 4.7 also illustrates that as the number of users increases, the total energy consumption of "no offloading" scenario, in which the users execute their tasks locally within their individual latency constraints, becomes less than that of EL-FullMA. That is because we assume that all the users employ the dynamic voltage scaling approach which minimizes the local energy consumption of

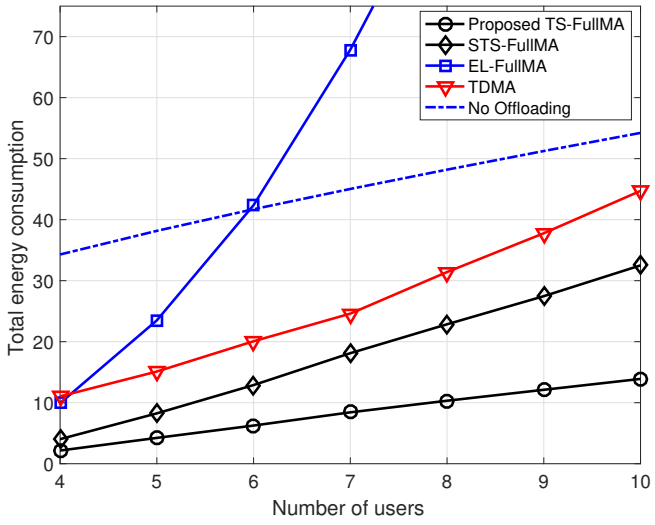


Figure 4.7: Average energy consumption of a partial offloading system. STS-FullMA denotes the approach introduced in Chapter 3, and EL-FullMA denotes the equal-latency FullMA-based approach proposed by Wang *et al.* (2018b).

each user with the maximum allowable latency of each user being taken into account. Figs 4.7 and 4.8 indicate that the proposed TS-FullMA scheme substantially reduces the energy consumption of the partial offloading system as compared to EL-FullMA, and it does so at a significantly lower computational cost. Moreover, they show that the significant energy reduction of TS-FullMA over STS-FullMA comes with a somewhat higher computational cost.

4.8 Conclusion

In this work we have proposed a time-slotted signalling structure for K-user computational offloading systems. This structure enables the system to take advantage of both the maximum allowable latency of each individual user and the differences between the latency constraints. Furthermore, we have shown that when the system employs

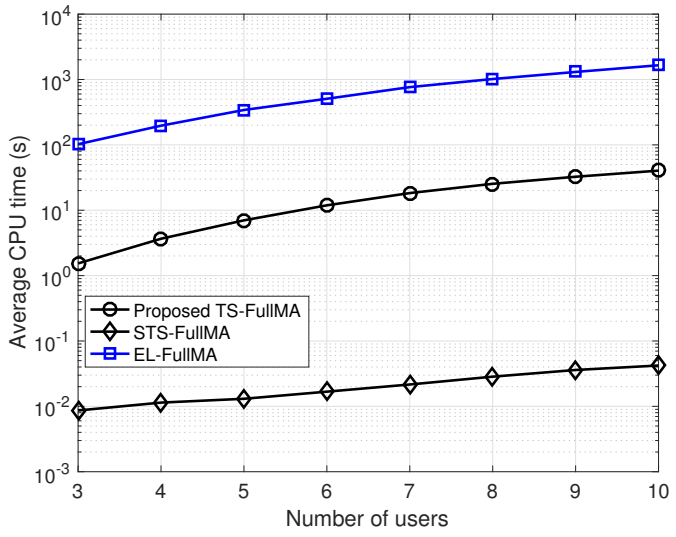


Figure 4.8: Average CPU time required for the proposed TS-FullMA algorithm and the algorithms for STS-FullMA introduced in Chapter 3, and EL-FullMA, proposed by Wang *et al.* (2018b), for different number of users in partial offloading case.

a multiple access scheme that exploits the full capabilities of the channel (FullMA), the generic time-slotted structure can be reduced to K slots, with simple expressions for the optimal slot durations and for the set of users that offload in each time slot. Furthermore, for both binary offloading and partial offloading we obtain closed-form expressions for a large fraction of design variables in the underlying communication resource allocation problems, and we exploit the structure of the remaining optimization problems to obtain effective algorithms to find the remaining design variables. Our numerical experiments demonstrate significant performance improvements and computational cost reduction over a competing equal-latency FullMA scheme (Wang *et al.*, 2018b) and a performance-computational cost trade-off with an existing single-time-slot FullMA scheme,Chapter 3 and (Salmani and Davidson, 2018c).

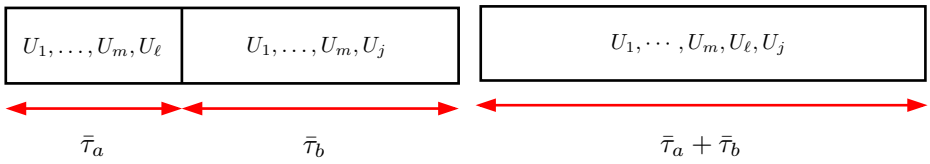


Figure 4.9: Equivalent time slots in a K-user FullMA offloading system.

4.A Reduced Number of Slots in a FullMA Scheme

Consider the two signalling structures in Fig 4.9. Let a and b denote the indices of the two time slots on the left, and let c denote the index of the single slot on the right, which is of duration $\bar{\tau}_c = \bar{\tau}_a + \bar{\tau}_b$. Let $\bar{\mathcal{S}} = \{1, 2, \cdots, m\}$ denote the users that are common to slots a and b. Hence, the offloading sets are $\mathcal{S}_a = \bar{\mathcal{S}} \cup \{\ell\}$, $\mathcal{S}_b = \bar{\mathcal{S}} \cup \{j\}$ and $\mathcal{S}_c = \bar{\mathcal{S}} \cup \{\ell, j\}$. Moreover, let R_{kn} and P_{kn} denote the rate and power of user $k \in \mathcal{S}_n$ in time slot $n \in \{a, b, c\}$.

In order for the two structures in Fig. 4.9 to be equivalent, the energy and the number of transmitted bits of each user should be equal, i.e.,

$$\bar{\tau}_a P_{ka} + \bar{\tau}_b P_{kb} = (\bar{\tau}_a + \bar{\tau}_b) P_{kc}, \quad \forall k \in \mathcal{S}_c,$$

$$\bar{\tau}_a P_{\ell a} = (\bar{\tau}_a + \bar{\tau}_b) P_{\ell c}, \quad \bar{\tau}_b P_{jb} = (\bar{\tau}_a + \bar{\tau}_b) P_{jc},$$

$$\bar{\tau}_a R_{ka} + \bar{\tau}_b R_{kb} = (\bar{\tau}_a + \bar{\tau}_b) R_{kc}, \quad \forall k \in \mathcal{S}_c,$$

$$\bar{\tau}_a R_{\ell a} = (\bar{\tau}_a + \bar{\tau}_b) R_{\ell c}, \quad \bar{\tau}_b R_{jb} = (\bar{\tau}_a + \bar{\tau}_b) R_{jc}.$$

The solution of that set of linear equations is

$$P_{kc} = \frac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b} P_{ka} + \frac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b} P_{kb}, \quad \forall k \in \mathcal{S}_c, \tag{4.35a}$$

$$P_{\ell c} = \frac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b} P_{\ell a}, \quad P_{jc} = \frac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b} P_{jb}, \tag{4.35b}$$

$$R_{kc} = \frac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b} R_{ka} + \frac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b} R_{kb}, \quad \forall k \in \mathcal{S}_c, \tag{4.35c}$$

$$R_{\ell c} = \frac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b} R_{\ell a}, \quad R_{jc} = \frac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b} R_{ja}. \tag{4.35d}$$

We assume that the powers and rates satisfy the rate region constraints for slots a and b, and what remains is to show that the powers and rates in (4.35) satisfy the rate region constraint for the slot c. Based on that assumption we need only examine the rate constraints for slot c that include both users ℓ and j:

$$\textstyle\sum_{k\in\bar{\mathcal{N}}} R_{kc} + R_{\ell c} + R_{jc} \leq \log_2(1 + \sum_{k\in\bar{\mathcal{N}}} \alpha_k P_{kc} + \alpha_\ell P_{\ell c} + \alpha_j P_{jc}), \quad \forall \bar{\mathcal{N}} \subseteq \bar{\mathcal{S}}. \tag{4.36}$$

Rewriting the left hand side of (4.36) in terms of the rates of the two time slots on the left hand side of Fig. 4.9, we have

$$\begin{aligned}\textstyle\sum_{k\in\bar{\mathcal{N}}} R_{kc} + R_{\ell c} + R_{jc} &= \frac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b}\big(\textstyle\sum_{k\in\bar{\mathcal{N}}} R_{ka} + R_{\ell a}\big) + \frac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b}\big(\textstyle\sum_{k\in\bar{\mathcal{N}}} R_{kb} + R_{jb}\big) \\ &\leq \frac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b}\big(\log_2(1 + \textstyle\sum_{k\in\bar{\mathcal{N}}} \alpha_k P_{ka} + \alpha_\ell P_{\ell a})\big) + \frac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b}\big(\log_2(1 + \textstyle\sum_{k\in\bar{\mathcal{N}}} \alpha_k P_{kb} + \alpha_j P_{jb})\big),\end{aligned}$$

where the inequality results from the fact that the constraints on the achievable rates are met in slots a and b. The proof is completed by using the concavity of the log

function along with the expressions in (4.35) as follows,

$$\begin{aligned}&\textstyle\sum_{k\in\bar{\mathcal{N}}} R_{kc} + R_{\ell c} + R_{jc}\\ &\quad\le \log_2\Big(1 + \tfrac{\bar{\tau}_a}{\bar{\tau}_a+\bar{\tau}_b}\big(\textstyle\sum_{k\in\bar{\mathcal{N}}} \alpha_k P_{ka} + \alpha_\ell P_{\ell a}\big) + \tfrac{\bar{\tau}_b}{\bar{\tau}_a+\bar{\tau}_b}\big(\textstyle\sum_{k\in\bar{\mathcal{N}}} \alpha_k P_{kb} + \alpha_j P_{jb}\big)\Big)\\ &\quad= \log_2\Big(1 + \textstyle\sum_{k\in\bar{\mathcal{N}}} \alpha_k P_{kc} + \alpha_\ell P_{\ell c} + \alpha_j P_{jc}\Big).\end{aligned}$$

This proof remains valid for the case in which there is only one user that is not common between slots a and b.

4.B At Most One Interval with Zero Derivative of (4.29a)

We will prove by contradiction that there is at most one sub-interval in the feasibility interval for ν that contains the point(s) at which the derivative of the objective function in (4.29a) is equal to zero. We begin with the fact that in each sub-interval of ν, (4.29a) is a convex function (its second derivative in each sub-interval is non-negative). Hence, in each sub-interval there is at most one (contiguous set of) point(s) at which the derivative of (4.29a) is equal to zero.

Now, let us assume that there are (sets of) points in two different intervals, m and n, at which the derivative of (4.29a) is zero. If we let $\nu_m \in (-\ln(2)\tilde{A}_{m+1}, -\ln(2)\tilde{A}_m]$ in sub-interval m denote a point at which the derivative of (4.29a) is zero, and if we define ν_n analogously, with $-\nu_m > -\nu_n$, then

$$\tfrac{df}{d\nu}\big|_{\substack{\nu=\nu_i,\\ i=n,m}} = 0 \Rightarrow -\nu_i = \tfrac{3M_k}{L_k^2}\big(B_k - \textstyle\sum_{\ell=1}^{i} \tau_\ell^\star \log_2(\tfrac{-\nu_i}{\ln(2)\tilde{A}_\ell})\big)^2.$$

Since $-\nu_m > -\nu_n$, we can write

$$\sum_{\ell=1}^{m} \tau_\ell^\star \log_2\big(\tfrac{-\nu_m}{\ln(2)\tilde{A}_\ell}\big) < \sum_{\ell=1}^{n} \tau_\ell^\star \log_2\big(\tfrac{-\nu_n}{\ln(2)\tilde{A}_\ell}\big). \tag{4.40}$$

On the other hand, $-\nu_m > -\nu_n$ implies that $m > n$. Hence,

$$\sum_{\ell=1}^{n} \tau_\ell^\star \log_2\big(\tfrac{-\nu_m}{\ln(2)\tilde{A}_\ell}\big) < \sum_{\ell=1}^{m} \tau_\ell^\star \log_2\big(\tfrac{-\nu_m}{\ln(2)\tilde{A}_\ell}\big) \tag{4.41a}$$

$$\Rightarrow \sum_{\ell=1}^{n} \tau_\ell^\star \log_2\big(\tfrac{-\nu_m}{\ln(2)\tilde{A}_\ell}\big) < \sum_{\ell=1}^{n} \tau_\ell^\star \log_2\big(\tfrac{-\nu_n}{\ln(2)\tilde{A}_\ell}\big) \tag{4.41b}$$

$$\Rightarrow -\nu_m < -\nu_n, \tag{4.41c}$$

which is a contradiction. So, there is at most one sub-interval that contains the (set of) point(s) at which the derivative of the objective function in (4.29a) is equal to zero.

4.C No Sign Change in Derivative of (4.29a) at Transitions

Here we will show that the sign of the derivative of the objective function in (4.29a) does not change at the transition points of the sub-intervals, i.e., $\nu = -\ln(2)\tilde{A}_\ell$. To do so, we observe that the derivative of (4.29a) in a generic sub-interval n, with $\nu \in (-\ln(2)\tilde{A}_{n+1}, -\ln(2)\tilde{A}_n]$, can be written as

$$\sum_{\ell=1}^{n} \tfrac{\tau_\ell^\star}{\ln(2)} \Big(-1 + \tfrac{3M_k}{L_k^2}\big(B_k - \sum_{j=1}^{n} \tau_j^\star \log_2(\tfrac{-\nu}{\ln(2)\tilde{A}_j})\big)^2 (\tfrac{-1}{\nu})\Big).$$

The values of derivatives on either side of the transition point between interval n and $n+1$, namely, $\nu = -\ln(2)\tilde{A}_{n+1}$, are

$$\sum_{\ell=1}^{n} \frac{\tau_\ell^\star}{\ln(2)}\Big(-1+\zeta_k\big(B_k-\sum_{j=1}^{n}\tau_j^\star \log_2(\tfrac{\tilde{A}_{n+1}}{\tilde{A}_j})\big)^2\Big), \tag{4.42}$$

and

$$\sum_{\ell=1}^{n+1} \frac{\tau_\ell^\star}{\ln(2)}\Big(-1+\zeta_k\big(B_k-\sum_{j=1}^{n}\tau_j^\star \log_2(\tfrac{\tilde{A}_{n+1}}{\tilde{A}_j})\big)^2\Big), \tag{4.43}$$

where $\zeta_k = \frac{3M_k}{L_k^2 \ln(2)\tilde{A}_{n+1}}$. It can be seen that the only difference between (4.42) and (4.43) is the extra term in the summation in (4.43) that results when $\ell = n+1$. Since the sign of that extra term is the same as the signs of the other terms, for $\ell = \{1, 2, \cdots, n\}$, the sign of derivative remains the same across the transition point.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis has focused on finding the optimal communication resource allocation in a computation offloading system. In doing so, it has captured the total energy consumption of users in the offloading system, and it has considered the two main offloading cases that result from the nature of the computational tasks, namely, the binary offloading and the partial offloading cases. As the first step, this thesis explored the impact of the multiple access scheme that is employed by the offloading system on the total energy consumption of the users by considering four different multiple access schemes in the two-user case. It has been shown that, in both the binary offloading and partial offloading cases, a capacity-approaching multiple access scheme, which can exploit the full capabilities of the channel, can significantly reduce the total energy consumption in a two-user offloading system as compared to the other multiple access schemes, i.e., independent decoding, sequential decoding without time sharing, and the TDMA scheme. By using the closed-form optimal solutions that have been

obtained for the resource allocation problem in those schemes, this thesis has also been able to show that i) the optimal energy consumption of a system that employs independent decoding can be achieved when that system operates in the TDMA scheme (for the scenarios in which the TDMA scheme is feasible), and ii) when the channel gains of the two users are equal, the optimized TDMA scheme achieves the optimal energy consumption.

As the next step, the energy minimization problem for a K-user offloading system, in which all the users are served over a single time slot, is addressed under a FullMA scheme and the TDMA scheme. It has been shown that in order to achieve the minimum energy consumption, the maximum allowable latency constraint of each individual user must be exploited. In the binary offloading case, the polymatroid structure of the capacity region is used to obtain a closed-form optimal solution for the case in which the offloading decisions are given (the "complete offloading problem"). That closed-form solution enables the development of a customized pruned greedy search algorithm that can achieve a close-to-optimal solution for the binary offloading case, while incurring a computational cost that is significantly lower than that of the existing approaches in the literature. In the partial offloading case, the proposed decomposition approach proposed by this thesis leads to a low-complexity algorithm that can achieve a quasi-closed-form expression for a stationary solution to the resource allocation problem.

Finally, based on the insights extracted from the two-user case, a time-slotted signaling structure was proposed for the K-user offloading system. That time-slotted signaling structure not only exploits the maximum latency constraint of each individual user, it also exploits the differences between the latency constraints of the users

by considering the reduction in the interference that results when a user finishes its offloading. The proposed algorithms in both the binary and partial offloading cases that are based on the time-slotted signaling outperform the existing algorithms in terms of both the energy consumption and the computational cost.

In summary, it has been shown in this thesis that in order to achieve the optimal energy consumption in a K-user offloading system, a multiple access scheme that can exploit the full capabilities of the channel must be employed. Moreover, it has also been shown that the optimal resource allocation can be obtained when both the maximum allowable latency of each individual user and the differences between the latency constraints of the users are fully exploited. Accordingly, an optimized time-slotted FullMA-based signalling structure was proposed. That structure enables efficient communication resource allocation in a K-user offloading system in both the binary offloading and partial offloading cases.

5.2 Future Work

In this thesis, energy-efficient and computationally-efficient algorithms have been proposed to achieve good solutions for the communication resource allocation problem in both the binary offloading and partial offloading cases of a K-user offloading system. However, there are several additional directions to be considered as future works.

5.2.1 Computation Resource Allocation

In the development herein, it is assumed that there are sufficiently large computation resources at the access point and hence, there is no contention among the users

to utilize the available computation resources. In practice, the limitation in the available computation resources must be taken into account and accordingly, those resources must be efficiently allocated to the users. In that case the resource allocation problem becomes a joint optimization problem over the available communication and computation resources. A preliminary work on that joint optimization problem in a two-user offloading system analogous to that studied in Chapter 2 has already been published as a conference paper (Salmani and Davidson, 2019c). That work showed that the optimal solution in the computationally constrained case can be obtained by using similar techniques to the unconstrained case. Tackling the corresponding joint resource allocation problem for a generic K-user offloading system is of considerable interest in scenarios in which the computational resources of the access point are in high demand. The successful transferral of insights from the two-user case to the K-user case that was demonstrated in this thesis suggests that this will be fruitful line of investigation.

5.2.2 Finite-block-length Regime

The constraints on the available communication resources that are considered in this thesis are based on the rate region that is achievable when the data block lengths are long enough for the guidance from asymptotic information theoretic results to be valid. However, latency-sensitive computational tasks imply a natural limit on the block length and hence, an investigation on the achievable rate region in the finite-block-length scenario is a natural extension of the work in this thesis. The fundamental limits on communication over finite block lengths has recently been studied

in several works (e.g., Polyanskiy *et al.*, 2010; MolavianJazi and Laneman, 2012). Accordingly, the resource allocation problem can be formulated for low-latency scenarios based on the these characterizations of the limits on the achievable rate region for a given probability error in the finite-block-length regime. As a starting point for that investigation, one could consider the communication resource allocation in a two-user binary offloading system over the finite-block-length regime under certain multiple access schemes (Salmani and Davidson, 2019b).

5.2.3 Data-driven Binary Offloading Decision Making

The resource allocation problem in the binary offloading case has a mixture of binary and continuous variables. One approach to tackling such a mixed-binary problem is to decompose that problem into an inner energy minimization problem over the continuous variables when the binary offloading decisions are given, and an outer problem to find a good binary offloading decisions. In this thesis, a customized pruned greedy search algorithm has been proposed that can achieve a close-to-optimal binary offloading decisions at a modest computational cost. However, in practice, further reductions in the computational cost of binary decision making would be an advantage. One approach to doing so is to train a deep neural network so that it can provide the binary offloading decisions based on the system parameters. A preliminary work (Salmani *et al.*, 2019) shows that a set of close-to-optimal binary offloading decisions in a single-time-slot FullMA-based binary offloading system (like that considered in Chapter 3) can be achieved based on a deep neural network which is trained using the concept of reinforcement learning. The extension of those results to the time-slotted signalling structure that was introduced in Chapter 4 is a rational

candidate for future work.

Bibliography

Abbas, N., Zhang, Y., Taherkordi, A., and Skeie, T. (2018). Mobile Edge Computing: A survey. *IEEE Internet Things J.*, **5**(1), 450–465.

Akherfi, K., Gerndt, M., and Harroud, H. (2018). Mobile cloud computing for computation offloading: Issues and challenges. *Appl. Comput. Informat.*, **14**(1), 1–16.

Antoniou, A. and Lu, W.-S. (2007). *Practical Optimization: Algorithms and Engineering Applications*. Springer.

Barbera, M., Kosta, S., Mei, A., and Stefa, J. (2013). To offload or not to offload? The bandwidth and energy costs of mobile cloud computing. In *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, pages 1285–1293, Turin, Italy.

Bi, S. and Zhang, Y. J. (2018). Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading. *IEEE Trans. Wireless Commun.*, **17**(6), 4177–4190.

Boyd, S. (2018). Ellipsoid method. Notes for EE364B, Stanford Univ., Palo Alto, CA, USA. Available online at: `http://web.stanford.edu/class/ee364b/lectures/ellipsoid_method_notes.pdf`.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Chen, M., Dong, M., and Liang, B. (2018). Resource sharing of a computing access point for multi-user mobile cloud offloading with delay constraints. *IEEE Trans. Mobile Comput.*, **17**(12), 2868–2881.

Chen, M.-H., Liang, B., and Dong, M. (2015). A semidefinite relaxation approach to mobile cloud offloading with computing access point. In *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, pages 186–190, Stockholm, Sweden.

Chen, M.-H., Dong, M., and Liang, B. (2016a). Joint offloading decision and resource allocation for mobile cloud with computing access point. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, pages 3516–3520, Shanghai, China.

Chen, M.-H., Dong, M., and Liang, B. (2016b). Multi-user mobile cloud offloading game with computing access point. In *Proc. IEEE Int. Conf. Cloud Netw. (CloudNet)*, pages 64–69, Pisa, Italy.

Chen, M.-H., Liang, B., and Dong, M. (2018). Multi-user multi-task offloading and resource allocation in mobile cloud systems. *IEEE Trans. Wireless Commun.*, **17**(10), 6790–6805.

Chen, X., Jiao, L., Li, W., and Fu, X. (2016c). Efficient multi-user computation offloading for mobile-edge cloud computing. *IEEE Trans. Netw.*, **24**(5), 2795–2808.

Chun, B.-G., Ihm, S., Maniatis, P., Naik, M., and Patti, A. (2011). CloneCloud:

Elastic execution between mobile device and cloud. In *Proc. ACM Conf. Computer Syst.*, pages 301–314, Salzburg, Austria.

Cisco (2017). Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper. Available online at: `http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html`.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., and Knuth, D. E. (1996). On the lambert W function. *Adv. Comput. Maths.*, **5**(1), 329–359.

Cover, T. M. and Thomas, J. A. (2012). *Elements of Information Theory*. John Wiley & Sons, 2nd edition.

Cuervo, E., Balasubramanian, A., Cho, D.-k., Wolman, A., Saroiu, S., Chandra, R., and Bahl, P. (2010). MAUI: Making smartphones last longer with code offload. In *Proc. ACM Int. Conf. Mobile Syst. Appl. Serv.*, pages 49–62, San Francisco, CA, USA.

Dinh, H. T., Lee, C., Niyato, D., and Wang, P. (2013). A survey of mobile cloud computing: Architecture, applications, and approaches. *Wireless Commun. Mobile Comput.*, **13**(18), 1587–1611.

El-Gamal, A. and Cover, T. M. (1980). Multiple user information theory. *Proc. IEEE*, **68**(12), 1466–1483.

El-Gamal, A. and Kim, Y.-H. (2011). *Network Information Theory*. Cambridge University Press.

Fernando, N., Loke, S. W., and Rahayu, W. (2013). Mobile cloud computing: A survey. *Future Generation Computer Systems*, **29**(1), 84–106.

Grant, M., Boyd, S., and Ye, Y. (2008). *CVX: MATLAB software for disciplined convex programming*. `http://cvxr.com/cvx/`.

Hong, M., Razaviyayn, M., Luo, Z.-Q., and Pang, J.-S. (2016). A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. *IEEE Signal Process. Mag.*, **33**(1), 57–77.

Hu, Y. C., Patel, M., Sabella, D., Sprecher, N., and Young, V. (2015). Mobile edge computing—A key technology towards 5G. *ETSI white paper*, **11**(11), 1–16.

Khan, A. R., Othman, M., Madani, S. A., and Khan, S. U. (2014). A survey of mobile cloud computing application models. *IEEE Commun. Surveys Tuts.*, **16**(1), 393–413.

Khan, M. A. (2015). A survey of computation offloading strategies for performance improvement of applications running on mobile devices. *J. Netw. Comput. Appl.*, **56**, 28–40.

Kosta, S., Aucinas, A., Hui, P., Mortier, R., and Zhang, X. (2012). ThinkAir: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading. In *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, pages 945–953, Orlando, FL, USA.

Kumar, K. and Lu, Y.-H. (2010). Cloud computing for mobile users: Can offloading computation save energy? *IEEE Computer*, **43**(4), 51–56.

Kumar, K., Liu, J., Lu, Y.-H., and Bhargava, B. (2013). A survey of computation offloading for mobile systems. *Mobile Netw. Appl.*, **18**(1), 129–140.

Lawrence, C. T. and Tits, A. L. (2001). A computationally efficient feasible sequential quadratic programming algorithm. *SIAM J. Opt.*, **11**(4), 1092–1118.

Lei, L., Zhong, Z., Zheng, K., Chen, J., and Meng, H. (2013). Challenges on wireless heterogeneous networks for mobile cloud computing. *IEEE Wireless Commun.*, **20**(3), 34–44.

Liu, F., Shu, P., Jin, H., Ding, L., Yu, J., Niu, D., and Li, B. (2013). Gearing resource-poor mobile devices with powerful clouds: Architectures, challenges, and applications. *IEEE Wireless Commun.*, **20**(3), 14–22.

Mach, P. and Becvar, Z. (2017). Mobile edge computing: A survey on architecture and computation offloading. *IEEE Commun. Surveys Tuts.*, **19**(3), 1628–1656.

Mahmoodi, S. E., Uma, R., and Subbalakshmi, K. (2019). Optimal joint scheduling and cloud offloading for mobile applications. *IEEE Trans. Cloud Comput.*, **7**(2), 301–313.

Mao, Y., Zhang, J., Song, S. H., and Letaief, K. B. (2017a). Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems. *IEEE Trans. Wireless Commun.*, **16**(9), 5994–6009.

Mao, Y., You, C., Zhang, J., Huang, K., and Letaief, K. B. (2017b). A survey on mobile edge computing: The communication perspective. *IEEE Commun. Surveys Tuts.*, **19**(4), 2322–2358.

Miettinen, A. P. and Nurminen, J. K. (2010). Energy efficiency of mobile clients in cloud computing. In *Proc. USENIX Workshop Hot Topics Cloud Comput. (HotCloud)*, pages 4–11, Boston, USA.

MolavianJazi, E. and Laneman, J. N. (2012). Simpler achievable rate regions for multiaccess with finite blocklength. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 36–40, Cambridge, MA, USA.

MolavianJazi, E. and Laneman, J. N. (2013). A finite-blocklength perspective on Gaussian multi-access channels. Available online at: `https://arxiv.org/abs/1309.2343v1`.

Muñoz, O., Pascual Iserte, A., Vidal, J., and Molina, M. (2014). Energy-latency trade-off for multiuser wireless computation offloading. In *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, pages 29–33, Istanbul, Turkey.

Muñoz, O., Pascual-Iserte, A., and Vidal, J. (2015). Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading. *IEEE Trans. Veh. Technol.*, **64**(10), 4738–4755.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, 2nd edition.

Patel, M., Naughton, B., Chan, C., Sprecher, N., Abeta, S., Neal, A., *et al.* (2014). Mobile-edge computing: Introductory technical. *ETSI white paper*, pages 1–36.

Polyanskiy, Y., Poor, H. V., and Verdú, S. (2010). Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, **56**(5), 2307–2359.

Raghavan, P. and Tompson, C. D. (1987). Randomized rounding: A technique for provably good algorithms and algorithmic proofs. *Combinatorica*, **7**(4), 365–374.

Rahimi, M. R., Ren, J., Liu, C. H., Vasilakos, A. V., and Venkatasubramanian, N. (2014). Mobile cloud computing: A survey, state of art and future directions. *Mobile Netw. Appl.*, **19**(2), 133–143.

Razaviyayn, M., Hong, M., and Luo, Z.-Q. (2013). A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM J. Opt.*, **23**(2), 1126–1153.

Salmani, M. and Davidson, T. N. (2016). Multiple access computational offloading. In *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, pages 1–6, Edinburgh, Scotland.

Salmani, M. and Davidson, T. N. (2017a). Energy-optimal computational offloading for simplified multiple access schemes. In *Conf. Rec. 51st Asilomar Conf. Signals, Syst., Comput.*, pages 1847–1851, Pacific Grove, CA, USA.

Salmani, M. and Davidson, T. N. (2017b). Multiple access computational offloading with computational constraints. In *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, pages 1–6, Sapporo, Japan.

Salmani, M. and Davidson, T. N. (2017c). Multiple access partial computational offloading: Two-user case. In *Proc. 23rd Asia-Pacific Conf. Commun. (APCC)*, pages 1–6, Perth, Australia.

Salmani, M. and Davidson, T. N. (2018a). Multiple access binary computational offloading in the K-user case. In *Conf. Rec. 52nd Asilomar Conf. Signals, Syst. Comput.*, pages 1599–1603, Pacific Grove, CA, USA.

Salmani, M. and Davidson, T. N. (2018b). Multiple access computational offloading: Communication resource allocation in the two-user case (extended version). Available online at: `https://arxiv.org/abs/1805.04981v2`.

Salmani, M. and Davidson, T. N. (2018c). Uplink resource allocation for multiple access computational offloading (extended version). Available online at: `https://arxiv.org/abs/1809.07453v2`.

Salmani, M. and Davidson, T. N. (2019a). Energy minimization of multi-user latency-constrained binary computation offloading. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 4589–4593, Brighton, UK.

Salmani, M. and Davidson, T. N. (2019b). On multi-user binary computation offloading in the finite-block-length regime. Submitted to the *53rd Asilomar Conf. Signals, Syst. Comput.*

Salmani, M. and Davidson, T. N. (2019c). Time-slotted resource allocation in a two-user computationally-constrained offloading system. In *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, pages 1–6, Cannes, France.

Salmani, M., Sohrabi, F., Davidson, T. N., and Yu, W. (2019). Multiple access binary computation offloading via reinforcement learning. In *Proc. IEEE Can. Workshop Inf. Theory (CWIT)*, pages 93–98, Hamilton, Canada.

Sardellitti, S., Scutari, G., and Barbarossa, S. (2015). Joint optimization of radio and computational resources for multicell mobile-edge computing. *IEEE Trans. Signal Info. Process. over Network*, **1**(2), 89–103.

Satyanarayanan, M., Bahl, P., Caceres, R., and Davies, N. (2009). The case for VM-based cloudlets in mobile computing. *IEEE Pervasive Comput.*, **8**(4), 14–23.

Scutari, G., Facchinei, F., and Lampariello, L. (2017). Parallel and distributed methods for constrained nonconvex optimization—Part I: Theory. *IEEE Trans. Signal Process.*, **65**(8), 1929–1944.

Toh, K.-C., Todd, M. J., and Tütüncü, R. H. (1999). SDPT3— A MATLAB software package for semidefinite programming, version 1.3. *Optim. Methods Softw.*, **11**(1-4), 545–581.

Tse, D. N. C. and Hanly, S. V. (1998). Multiaccess fading channels. I. Polymatroid structure, optimal resource allocation and throughput capacities. *IEEE Trans. Inf. Theory*, **44**(7), 2796–2815.

Wang, F., Xu, J., and Ding, Z. (2017a). Optimized multiuser computation offloading with multi-antenna NOMA. In *Proc. IEEE Globecom Workshops*, pages 1–7, Singapore.

Wang, F., Xu, J., Wang, X., and Cui, S. (2018a). Joint offloading and computing optimization in wireless powered mobile-edge computing systems. *IEEE Trans. Wireless Commun.*, **17**(3), 1784–1797.

Wang, F., Xu, J., and Ding, Z. (2018b). Multi-antenna NOMA for computation offloading in multiuser mobile edge computing systems. Available online at: `https://arxiv.org/abs/1707.02486v3`.

Wang, S., Zhang, X., Zhang, Y., Wang, L., Yang, J., and Wang, W. (2017b). A survey

on mobile edge networks: Convergence of computing, caching and communications. *IEEE Access*, **5**, 6757–6779.

Wang, Y., Sheng, M., Wang, X., Wang, L., and Li, J. (2016). Mobile-edge computing: Partial computation offloading using dynamic voltage scaling. *IEEE Trans. Commun.*, **64**(10), 4268–4282.

Weiser, M. (1991). The computer for the 21st century. *Sci. Am.*, **265**(3), 94–104.

Wu, H., Wang, Q., and Wolter, K. (2013). Tradeoff between performance improvement and energy saving in mobile cloud offloading systems. In *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, pages 728–732, Budapest, Hungary.

Xing, C., Jing, Y., Wang, S., Ma, S., and Poor, H. V. (2018). New viewpoint and algorithms for water-filling solutions in wireless communications. Available online at: `https://arxiv.org/abs/1808.01707`.

You, C., Huang, K., Chae, H., and Kim, B.-H. (2017). Energy-efficient resource allocation for mobile-edge computation offloading. *IEEE Trans. Wireless Commun.*, **16**(3), 1397–1411.

Zhang, W., Wen, Y., Guan, K., Kilper, D., Luo, H., and Wu, D. O. (2013). Energy-optimal mobile cloud computing under stochastic wireless channel. *IEEE Trans. Wireless Commun.*, **12**(9), 4569–4581.

Zhang, Y., Liu, H., Jiao, L., and Fu, X. (2012). To offload or not to offload: An efficient code partition algorithm for mobile cloud computing. In *Proc. IEEE 1st Int. Conf. Cloud Netw. (CloudNet)*, pages 80–86, Paris, France.