The Dissertation Committee for Hongseok Kim
certifies that this is the approved version of the following dissertation:

**Exploring Tradeoffs in Wireless Networks under Flow-Level Traffic: Energy, Capacity and QoS**

Committee:

Gustavo de Veciana, Supervisor

Jeffrey G. Andrews

Robert W. Heath Jr.

Lili Qiu

Sanjay Shakkottai

# Exploring Tradeoffs in Wireless Networks under Flow-Level Traffic: Energy, Capacity and QoS

by

**Hongseok Kim, B.S.; M.S.**

**Dissertation**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

**The University of Texas at Austin**

December 2009

Dedicated to my wife, Eun-Hee for her support, encouragement and devotion.

# Acknowledgments

First of all I would like to express my deepest appreciation to my advisor Prof. Gustavo de Veciana. I am extremely fortunate to study under his supervision. He is the epitome of the great teacher and advisor. Not to mention of his insight, intuition, smartness and high expectation, he is really patient in advising students so that they can finally develop and improve themselves. He is truly my role model of the advisor, teacher and researcher. I also would like to thank my committee members, Prof. Jeff Andrews, Prof. Robert W. Heath Jr., Prof. Lili Qiu, and Prof. Sanjay Shakkottai for sharing their time and insights on my work.

I also want to thank my lab members Bilal, Yuchul, Balaji, Seung-June and Shailesh for many discussions and having a good time together. I also want to thank Xiangying, Muthu, Yang-Seok and Nageen for my pleasant and productive internship at Intel Corporation. I want to thank Qualcomm Flarion People including Saurabh, Aleks, Junyi and Rajiv for the enjoyable internship. I thank all of my Seoul Science High School Alumni in Austin, specially Sang-hyun, Goo and Taejoon, the class of 1994. I thank Chan-Byoung, Wonsoo, Taesoo, who came to UT with me and shared many memories during the years. I also want to thank my MS advisor Prof. Taejeong Kim in Seoul National University.

I sincerely appreciate the support and the prayer from my father and mother, who always encourage me to pursue my goal. I also thank my parents in law and my daughter Suann. Finally I cannot thank anyone more than my wife Eun-Hee, who delayed her study to support me. It would not have been possible without her understanding, encouragement and devotion. I will pay my debt to her throughout my life.

**Exploring Tradeoffs in Wireless Networks under Flow-Level Traffic: Energy, Capacity and QoS**

Publication No. _____

Hongseok Kim, Ph.D.
The University of Texas at Austin, 2009

Supervisor: Gustavo de Veciana

Wireless resources are scarce, shared and time-varying making resource allocation mechanisms, e.g., *scheduling*, a key and challenging element of wireless system design. In designing good schedulers, we consider three types of performance metrics: system capacity, quality of service (QoS) seen by users, and the energy expenditures (battery lifetimes) incurred by mobile terminals. In this dissertation we investigate the impact of scheduling policies on these performance metrics, their interactions, and/or tradeoffs, and we specifically focus on *flow-level* performance under stochastic traffic loads.

In the first part of the dissertation we evaluate interactions among flow-level performance metrics when integrating QoS and best effort flows in a wireless system using opportunistic scheduling. We introduce a simple flow-level model capturing the salient features of bandwidth sharing for an opportunistic scheduler which ensures a mean throughput to each QoS stream on every time slot. We show that the integration of QoS and best effort flows results in a loss of opportunism, which in turn results in a reduction of the stability region, degradation in system capacity, and increased file transfer delay.

In the second part of the dissertation we study several ways in which mobile terminals can backoff on their uplink transmit power (thus slow down their transmissions) in order to extend battery lifetimes. This is particularly effective when a wireless system is underloaded, so the degradation in the users' perceived performance can be negligible. The challenge, however, is developing a mechanism that achieves a good tradeoff among transmit power, idling/circuit power, and the performance customers will see. We consider systems with flow-level dynamics supporting either real-time or best effort (e.g., file transfers) sessions. We show that significant energy savings can be achieved by leveraging dynamic spare capacity. We then extend our study to the case where mobile terminals have multiple transmit antennas.

In the third part of the dissertation we develop a framework for user association in infrastructure-based wireless networks, specifically focused on adaptively balancing flow loads given spatially inhomogeneous traffic distributions. Our work encompasses several possible user association objective functions resulting in rate-optimal, throughput-optimal, delay-optimal, and load-equalizing policy, which we collectively denote $\alpha$-optimal user association. We prove that the optimal load vector that minimizes this function is the fixed point of a certain mapping. Based on this mapping we propose an iterative distributed user association policy and prove that it converges to the globally optimal decision in steady state. In addition we address admission control policies for the case where the system cannot be stabilized.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Wireless cellular systems are evolving towards supporting both best effort (data) and real time (voice and video) traffic so as to meet the growing demands of multiple services over a single platform. Since wireless channels are shared, limited and time-varying, resource allocation, e.g., *scheduling* users' transmissions, is a key and challenging element in the design of such systems. In designing good schedulers, we need to consider at least three key performance metrics: system capacity, quality of service (QoS) seen by users, and the energy expenditure (or battery lifetime) incurred by mobile terminals, and it is desirable to have a high degree of control over tradeoffs associated with these metrics. Fig.1.1 shows the design space in terms of performance metrics such as QoS, capacity and energy efficiency. Roughly speaking a specific scheduling policy is represented as a point in this design space.

The scope of this dissertation concerns the impact of scheduling policies on these performance metrics, their interactions, and/or associated tradeoffs. We focus on flow-level performance considering stochastic traffic loads. New flows, either real-time sessions and/or file transfer requests, are initiated at random and leave the system after being served. As a consequence the number of ongoing flows dynamically changes in time. This is usually referred to as the *flow-level dynamics*. Studying dynamic systems is helpful to better understand performance in real systems, but, in general, it is hard to do and has not been pursued as extensively as the static versions, i.e., with a fixed set of backlogged users.

Each flow is an abstraction of a stream of packets corresponding to a new file,

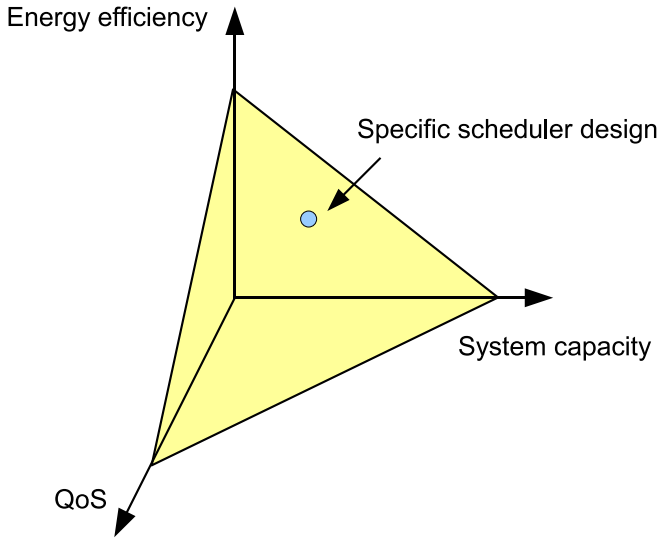



Figure 1.1: Design space of wireless systems

web page or real-time voice/video session. Poisson processes have generally been used to model flows generated by large population of (independent) customers, or new web page download patterns from populations of web browsing sessions, see e.g., [20, 21, 37, 45]. Recently, flow-level models have been considered in studying statistical bandwidth sharing in wired networks [20, 37, 45]. In the context of wireless systems, it was observed that throughput seen by dynamic user populations can be substantially different from that of a fixed number of users [21]. In this dissertation we will also encounter various cases where flow-level performance with a dynamic user population is very different from that of a static user population. Specifically, we mostly focus on capacity (or equivalently flow delay) and energy-efficiency. Below we give a brief overview of the tradeoffs considered in this dissertation.

## 1.1 QoS-capacity tradeoff in an opportunistic wireless system

Time-varying channels are a distinguishing characteristic of wireless systems relative to wired ones and can potentially result in poor system capacity. To overcome this difficulty, channel-aware scheduling (so called opportunistic scheduling) was proposed, where one chooses to serve users who currently see good channel conditions. Opportunistic scheduling has been shown to substantially increase system capacity and been implemented in current systems supporting data services [12]. However, unlike what is typically the case in wired systems, more capacity does not necessarily imply better user-perceived QoS in an opportunistic system. This is because the maximum achievable capacity in an opportunistic system is constrained by the individual users' QoS requirements. Specifically, in order to sustain a minimum bandwidth for users, one may need to sometimes compromise opportunism by scheduling users whose current capacity is not the highest. In Chapter 2 we investigate the tradeoff between QoS and system capacity when integrating services (QoS and best effort) in an opportunistic wireless system. We will see that the integration of QoS and best effort flows compromises the benefits of opportunism in crucial aspects; stability region reduction, system throughput degradation and increased file transfer delay. All of these negative impacts are referred to as a loss in opportunism due to integration.

## 1.2 Energy-conservation leveraging spare capacity

Not unlike most networking infrastructure (particularly that supporting data), wireless access networks are unlikely to be fully utilized all the time. Indeed as a result of time varying, non-stationary loads, or unpredictable bursty loads these networks are often overdesigned to be able to support a peak load condition, and so are often underutilized. Thus, if a system has *spare capacity*, which we will interpret as excess capacity relative to a desired user-perceived performance, one may consider slowing down transmissions (so called 'lazy

scheduling' [95]). This can be beneficial in terms of conserving mobile terminals' energy expenditures.

In Chapter 3 of this dissertation we address energy conservation techniques for real-time sessions and file transfers, respectively. In the case of real-time sessions, the sojourn time of the flows are independent of energy saving mechanism but the system must satisfy a minimum rate requirement. We show that the energy-optimal transmission strategy for real-time sessions is determined by solving a convex optimization. An iterative approach exhibiting superlinear convergence achieves a substantial energy savings, e.g., more than 50% for a system where session blocking probability is 0.1% or less.[1] The case of file transfers is more subtle because power backoff (and thus the slowing down of transmissions) changes the system flows' dynamics. For such "best effort" traffic we study energy-efficient transmission strategies to realize energy-delay tradeoff. The proposed mechanism achieves a 35–75% in energy savings depending on the traffic load and file transfer target throughput. A key insight, relative to previous work focusing on static scenarios, is that idling power has a significant impact on energy-efficiency, while circuit power has a limited impact as the load increases.

In Chapter 4 we further extend our energy saving approach to the case of multiple antenna systems. We propose a mechanism to switch between multiple-input multiple-output (MIMO) with two transmit antennas and single-input multiple-output (SIMO) to conserve mobile terminals' energy. The key idea is simple. When the system is underutilized, the MT operates in the SIMO mode at a low spectral efficiency to save energy, but when congested, the MT operates in the MIMO mode achieving high spectral efficiency to increase throughput. This is done in an adaptive way considering two aspects – the dynamics of

[1] Note the session blocking probability is an indicator of the amount of spare capacity in the system.

network traffic and the channel variations. Extensive flow-level simulations under dynamic loads confirm that the proposed technique can reduce the transmission energy by more than 50% and enables an effective tradeoff between file transfer delay and energy conservation.

## 1.3 $\alpha$-optimal user association and cell load balancing

While Chapters 2–4 focused on the operation of single cells, in Chapter 5 we consider a multiple cell scenario. One of the important problems in multi-cell data networks is properly associating mobile terminals with serving base stations. This problem is usually called *user association.* In Chapter 5 we develop a framework for user association specifically focused on flow-level cell load balancing under spatially inhomogeneous traffic distributions. Our work encompasses several different user association policies: rate-optimal, throughput-optimal, delay-optimal, and load-equalizing, which we collectively denote $\alpha$-optimal user association. Interestingly, this problem can be viewed as a simple routing problem among queues, but until our work had not been fully studied in the context of dynamic systems. The optimal load vector minimizing our objective function is shown to be a fixed point of a certain mapping, and the fixed point equation can be iteratively solved in a distributed manner. This leads to a simple adaptive approach to the user association problem. In addition we address admission control policies for the case where the system is overloaded and thus cannot be stabilized.

## 1.4 Summary

As mentioned earlier this dissertation covers a wide range of problems associated with wireless networks: QoS, energy conservation and capacity. The central theme is a focus on stochastic loads, flow-level performance and tradeoffs amongst these key metrics.

# Chapter 2

# Evaluating Service Integration in an Opportunistic Wireless Systems

## 2.1 Introduction

Wireless networks are evolving towards supporting multiple services, *e.g.,* both best effort and QoS streaming traffic. Since the integration of different services on a single platform is expected to generate new revenue and reduce network management cost, intensive research efforts have been devoted towards designing such networks. However, because wireless resources are limited and shared by users experiencing time-varying channels, service integration may be quite challenging. A key element in such systems is the traffic scheduler and complementary resource management component that can assure users appropriate QoS. In addition, such schedulers may be designed to be opportunistic, *i.e.,* serve users whose current channel capacities are high. Attempting to be opportunistic while meeting users' QoS requirements presents significant new challenges, see [12, 21, 53, 70, 71, 87–90, 106, 107].

Opportunistic scheduling schemes developed so far are mostly packet-level algorithms [12, 53, 70, 71, 87–89, 106, 107] focusing on the case where the user population is static and their queues are backlogged. This assumption is meaningful in short time scales where the user population does not change much. Under this assumption, [70, 71] propose strategies that maximize system throughput under temporal and utilitarian fairness criteria. The work in [87, 89] proposes a scheduling scheme based on a history of channel information. However, static approaches may not capture the flow-level dynamics in which new flows

come to the system randomly and leave after being served. In a real system, user population changes over time and it is of interest to know system performance such as average throughput and average file transfer delays. This problem was first addressed by Borst who proposed flow-level analysis using multi-class processor sharing model [21]. However, this work did not deal with the mixes of QoS and best effort traffic. Other attempts to integrate QoS and best effort in wireless opportunistic systems have been recently done by [89, 106]. But, these studies focused on packet-level performance only.

Hence, we are motivated to study a new model that addresses the interaction of heterogeneous traffic at the flow-level. In *wired* network case, there had been several studies on this topic [2, 10, 11, 19, 38, 54, 66, 78]. To analyze the interaction of elastic and streaming flows, researchers have used a 2-dimensional Markovian model [2, 54, 66]. The work of Key *et al.* suggests that the integration of heterogeneous traffic has a positive consequence, *i.e.,* stabilizing effect [54]. The work in [38] highlights that such systems are likely to see transient, or local, instability. Specifically when there are too many QoS sessions, the best effort flows may accumulate, but subsequently subside once more bandwidth becomes available, *i.e.,* QoS sessions leave the system. In [10, 11] the authors propose an integrated admission control for both of streaming and elastic traffic to guarantee QoS. Note that these studies were done at a higher level, *i.e.,* considering the integration of TCP and UDP and where the service rate of wired network is constant. The major difference that arises in wireless networks, however, lies in that the service rate of a wireless network is time-varying, and, furthermore, may be shared in an opportunistic manner making the analysis somewhat challenging. The main goal of this chapter is to model and study the flow-level characteristics for an opportunistic wireless system shared by a traffic mix of QoS streaming and best effort traffic.

To this end, we propose a new flow-level model for such a system. The scheduler is designed to guarantee a mean throughput to streaming media irrespective of the number of ongoing flows by borrowing/lending bandwidth from/to best effort flows. Thus, QoS flows have a fixed average throughput per slot which might be set to roughly meet their QoS requirements. Other QoS metrics such as delay or jitter are not considered here to keep the model simple. By contrast, the performance metric for best effort flows will be the *average delay* to finish a finite size of file transfer using HTTP or FTP. Our QoS definition is simple but it gives us an insight of evaluating service integration.

***Contributions:*** The following are the key contributions of this chapter.

1. To our knowledge, this work is the first to attempt to investigate the flow-level interaction between QoS and best effort traffic in systems exploiting opportunism. Our model is simplistic but significant in that it incorporates the essential ingredients of flow-level behaviors such as QoS requirements, opportunistic sharing, stability and evaluation of the stationary distribution. We also identify the necessary and sufficient stability condition of 2-dimensional Markov chain.

2. We show how the integration of QoS and best effort flows compromises the benefits of opportunism in crucial aspects; stability region reduction, system throughput degradation and increased file transfer delay. All of these negative impacts are called *loss in opportunism* of integration. Our analysis shows that, for example, introducing a single QoS user requesting 300kbps would degrade by 33% the maximum system capacity in the CDMA/HDR system described in [12]. We will show that such losses increase in proportion to the opportunistic gains, the number of QoS users and the guaranteed bandwidth, but is inversely proportional to SNR in Rayleigh fading channel model.

3. As in the wired case [38, 66], even if the system is stable, it may exhibit local instability for best effort users. However it appears to be a more crucial phenomenon in wireless systems exploiting opportunism. To circumvent this problem, we suggest that admission control of best effort flows is necessary.

This chapter is organized as follows. In Section 2.2 we describe our flow-level model for mixed traffic based on bandwidth borrowing and lending among traffic types. We also build up a compact model and analysis tool to investigate flow-level dynamics. Section 2.3 is devoted to the stability of the system. Section 2.4 evaluates the opportunistic losses of integration by quantifying the reduction in the stability region, throughput degradation and increase in delay. Section 2.5 deals with the performance impact of local instability and the necessity of call admission control, and is followed by conclusions in Section 2.6.

## 2.2 System model

### 2.2.1 Assumptions

We consider a wireless access point shared by multiple mobile users. Wireless access point is assumed to accommodate multiple users by Time Division Multiple Access (TDMA) scheme where time is divided into equal-sized slots and at most one user gets served per slot. We assume that channel capacity for each user is a stationary ergodic process and these processes are independent, identically distributed (*i.i.d.*) across users. This assumption allows us to adopt max rate scheduling as a basic scheduling policy. It maximizes total capacity or throughput of the system [63]. (We will use the term *capacity* or *throughput* interchangeably in the sequel.) The scheduling policy will be revised to model the need to meet QoS session requirements. For simplicity, we divide users into two groups: QoS and

best effort users.[1] The channel capacity for each user is independent across slots and remains constant during a slot, *i.e.,* we assume fast fading channel, with a slot time corresponding to coherence time. In the sequel we assume time slots are small relative to flow dynamics and time-scale, and we will model the system dynamics based on a continuous-time model.

A sustained throughput is one of the basic requirements for QoS, so we assume that QoS users are guaranteed a mean throughput $\bar{b}$ per time slot irrespective of the number of best effort users. Our notion of QoS, however, does not guarantee delay constraints. In fact, one might argue that this assumption is not realistic because real-time interactive applications such as voice or video communication will require delay constraints. Nevertheless, if every time slot, QoS users get an average throughput $\bar{b}$, it is very unlikely that QoS user is starved for long period of time slots, which implies decent delay performance. This simple model roughly captures the throughput loss in the system – it is likely to be worse if a more sophisticated scheduling scheme meeting QoS requirements is used. Furthermore, for non real-time one-way streaming media such as video on demand, our QoS notion makes sense because we can assume that end-user devices have buffer space to compensate delay or jitter occurred during transmission.

### 2.2.2 Flow-level model of mixed traffic

In our flow-level model, QoS and best effort flows arrive randomly and leave after being served. We assume that the arrivals of QoS flows follow a Poisson process with arrival rate $\lambda_q$ and have a holding time which is exponentially distributed with mean $\mu_q^{-1}$. The maximum number of QoS flows is limited to $n^*$ in order to guarantee a bandwidth $\bar{b}$. We also assume that the arrivals of best effort flows follow an independent Poisson process with

[1] Even under this assumption, each user can still receive both QoS and best effort services using TDMA.

arrival rate $\lambda_b$ and have file sizes which are exponentially distributed with mean $\mu_b^{-1}$.

Let $N_q(t)$ be the number of QoS flows and $N_b(t)$ be the number of best effort flows in the system. Then, $(N_q(t), N_b(t))$ is a 2-dimensional Markov process with state space $\{0, \ldots, n^*\} \times \mathbb{Z}^+$. Since our model is an opportunistic system, the available capacity for QoS and best effort flows depends on how they are scheduled. Let $g(n_q, n_b)$ denote the average system capacity given $(n_q, n_b)$. Then, the total capacity required for QoS flows is $n_q\bar{b}$ and the capacity available to best effort flows is $g_b(n_q, n_b) := g(n_q, n_b) - n_q\bar{b}$. The rate matrix for the chain is then given by:

$$\begin{aligned}
q\big((n_q, n_b), (n_q + 1, n_b)\big) &= \lambda_q \mathbf{1}_{\{n_q < n^*\}}; \\
q\big((n_q, n_b), (n_q, n_b + 1)\big) &= \lambda_b; \\
q\big((n_q, n_b), (n_q, n_b - 1)\big) &= g_b(n_q, n_b)\mu_b \mathbf{1}_{\{n_b \geq 1\}}; \\
q\big((n_q, n_b), (n_q - 1, n_b)\big) &= n_q \mu_q. \qquad (2.1)
\end{aligned}$$

Note that the number of QoS sessions follows an M/M/m/m-like system so the stationary distribution of QoS flows $\pi_q(n_q)$ is independent of $n_b$ and given by

$$\pi_q(n_q) = \pi_q(0)\rho_q^{n_q}\frac{1}{n_q!} \qquad (2.2)$$

where $\rho_q = \frac{\lambda_q}{\mu_q}$ and $\pi_q(0) = \big[\sum_{n_q=0}^{n^*} \rho_q^{n_q}\frac{1}{n_q!}\big]^{-1}$. The blocking probability of QoS flows is given by *Erlang-B formula* as $\pi_q(n^*)$ [14]. Meanwhile the dynamics of the number of best effort flows follow a processor sharing system with varying capacity.

### 2.2.3 Proposed opportunistic scheduling

Suppose that at a given time we have total number of flows $n$. The total number of flows is the sum of $n_q$ QoS flows and $n_b$ best effort flows. Let $X_i$, $i \in \{1, \ldots, n\}$ be a random variable representing channel capacity of user $i$. Since all users are symmetric, the

maximum system capacity is given by $g(n) := \mathbb{E}[X^{(n)}]$ where $X^{(n)} \triangleq \max[X_1, \cdots, X_n]$ and is shared equally among the users. So the bandwidth per user $h(n) := \frac{g(n)}{n}$ is decreasing in $n$ while $g(n)$ is increasing in $n$ as shown in Fig. 2.1. Now, for every time slot, we want to guarantee $\bar{b}$ to QoS flows. If $n \leq n^*$ where

$$n^* := \max\{n | h(n) \geq \bar{b}, n \in \mathbb{Z}^+\},$$

every user has a bandwidth of at least $\bar{b}$ and we satisfy the QoS requirement. We refer to $\{(n_q, n_b) | n_q + n_b \leq n^*\}$ as the *normal regime.*

However, if $n_q + n_b > n^*$, QoS users will not meet their requirement $\bar{b}$. How can we guarantee $\bar{b}$ to QoS flows in the *overloaded regime* $\{(n_q, n_b) | n_q + n_b > n^*\}$? To do this, we propose to use *bandwidth borrowing* as follows. If $h(n)$ is below $\bar{b}$, *i.e.,* $n > n^*$, then QoS flows borrow time slots from best effort flows. Similarly, if $h(n)$ is over $\bar{b}$, then QoS flows lend their time slots to best effort flows. As a consequence, the average throughput of QoS is $\bar{b}$ in every time slot. Under this model we still need admission control for QoS flows to ensure $n_q \leq n^*$. These borrowing and lending mechanisms are described in more detail below.

#### 2.2.3.1 Capacity balance equation in the overloaded regime

The balance equation is given by

$$h(n_q + n_b) + \frac{\alpha(n_q, n_b)}{n_q} \mathbb{E}[X^{(n_q)} | X^{(n_b)} > X^{(n_q)}] P_b = \bar{b} \tag{2.3}$$

where $P_b := P\big(X^{(n_b)} > X^{(n_q)}\big)$ and $\alpha(n_q, n_b)$ is the *borrowing* probability. The intuition for the equation is as follows. The amount of bandwidth each QoS flow must borrow is $\bar{b} - h(n_q + n_b)$. To compensate this deficiency, we will randomly, with probability $\alpha(n_q, n_b)$, give a best effort slot, *i.e.,* one where $X^{(n_b)} > X^{(n_q)}$, to the QoS user currently seeing the

Table 2.1: Notation Summary

| | |
|---|---|
| $n_q$ | number of QoS flows in a system |
| $n_b$ | number of best effort flows in a system |
| $n^*$ | maximum number of QoS flows |
| $\bar{b}$ | average throughput of QoS flows |
| $X_i$ | a random variable representing channel capacity of user $i$. |
| $X^{(n)}$ | $\max[X_1, \cdots, X_n]$ |
| $g(n)$ | $\mathbb{E}[X^{(n)}]$ |
| $h(n)$ | $\frac{g(n)}{n}$ |
| $g(n_q, n_b)$ | the system capacity for $n_q$ QoS and $n_b$ best effort flows. |
| $g_b(n_q, n_b)$ | $g(n_q, n_b) - n_q\bar{b}$, the capacity of best effort users. |
| $h_b(n_q, n_b)$ | $\frac{g_b(n_q, n_b)}{n_b}$, the individual capacity of best effort user. |
| $\bar{g}(n_q)$ | $\lim_{n_b \to \infty} g(n_q, n_b)$ |
| $\bar{g}_b(n_q)$ | $\lim_{n_b \to \infty} g_b(n_q, n_b)$ |
| $g_b^*(n_q, n_b)$ | $g(n_q + n_b) - n_q\bar{b}$ |
| $\alpha(n_q, n_b)$ | bandwidth borrowing probability |
| $\beta(n_q, n_b)$ | bandwidth lending probability |
| $\xi(\bar{b}, n_q)$ | the capacity gap at $\bar{b}$ and $n_q$ |
| $C$ | maximum system capacity |
| $\kappa$ | $\frac{C}{E[X]}$, opportunistic gain |
| $\eta$ | call blocking probability |

best channel. Thus, the total borrowed bandwidth is $\alpha(n_q, n_b)\mathbb{E}[X^{(n_q)}|X^{(n_b)} > X^{(n_q)}]P_b$ and it is shared by $n_q$ QoS flows to meet the guaranteed bandwidth $\bar{b}$.

Note that, in the overload regime, to maintain the average throughput $\bar{b}$, at some time slots, a best effort user which currently has the best channel amongst all users may have to give the time slot to QoS user. Thus, meeting QoS requirements inevitably degrades the overall system throughput. We will study this in detail in Section 2.4.

#### 2.2.3.2 Capacity balance equation in the normal regime

Assuming QoS users do not need more bandwidth than $\bar{b}$, we reallocate the excess bandwidth of QoS flows to best effort flows, and the capacity balance equation in the normal

regime is given by

$$h(n_q + n_b) - \frac{\beta(n_q, n_b)}{n_q}\mathbb{E}[X^{(n_q)}|X^{(n_b)} \leq X^{(n_q)}]P_q = \bar{b} \tag{2.4}$$

where $\beta(n_q, n_b)$ is the *lending* probability and $P_q := P\big(X^{(n_b)} \leq X^{(n_q)}\big)$. [2]

Solving (2.3) and (2.4) we can obtain the opportunistic system capacity of $g(n_q, n_b)$. Since QoS users always have average throughput $\bar{b}$, the capacity of best effort flows at $(n_q, n_b)$ state is $g_b(n_q, n_b) = g(n_q, n_b) - n_q\bar{b}$.

### 2.2.4 Capacity of best effort flows in overloaded regime

To solve the capacity balance equation of (2.3) and (2.4), we first compute the conditional capacity of QoS users $\mathbb{E}[X^{(n_q)}|X^{(n_b)} > X^{(n_q)}]$. We will do this as follows:

$$\mathbb{E}[X^{(n_q)}|X^{(n_b)} > X^{(n_q)}]P_b + \mathbb{E}[X^{(n_q)}|X^{(n_b)} \leq X^{(n_q)}]P_q = \mathbb{E}[X^{(n_q)}] \tag{2.5}$$

where $P_b = P(X^{(n_b)} > X^{(n_q)}) = \frac{n_b}{n_q+n_b}$ and $P_q = 1 - P_b$. Note that

$$\mathbb{E}[X^{(n_q)}|X^{(n_b)} \leq X^{(n_q)}] \;\; = \;\; \mathbb{E}[X^{(n_q+n_b)}]. \tag{2.6}$$

Combining (2.3), (2.5), (2.6), we obtain

$$h(n_q + n_b)(1 - \alpha(n_q, n_b)) + h(n_q)\alpha(n_q, n_b) = \bar{b}$$

which means that the system in overloaded regime is identical to one operating as if it had only $n_q$ flows with probability $\alpha(n_q, n_b)$ and one with $n_q + n_b$ flows with probability $1 - \alpha(n_q, n_b)$. So, the borrowing probability is

$$\alpha(n_q, n_b) \;\; = \;\; \frac{\bar{b} - h(n_q + n_b)}{h(n_q) - h(n_q + n_b)}.$$

[2] Since we assume that $X_i$ is a continuous random variable, the equality of $X^{(n_b)} = X^{(n_q)}$ can be placed either in $P_b$ or $P_q$. In a discrete case which is more likely in the case in practice, we need a tie-breaking rule.

The numerator of $\alpha(n_q, n_b)$ is bandwidth deficit for individual QoS flow. So, as $n_q$ increases to $n^*$, $\alpha(n_q, n_b)$ goes to 1 and QoS flows have to borrow more time slots while best effort flows are starved. If $n_b$ goes to $\infty$, $\alpha(n_q, n_b)$ goes to $\frac{\bar{b}}{h(n_q)}$ which is less than or equal to 1.

Since best effort flows lose their slots with probability $\alpha(n_q, n_b)$, the total capacity available to best effort flows is $g_b(n_q, n_b) = g(n_q + n_b)(1 - \alpha(n_q, n_b))P_b$, and so

$$g_b(n_q, n_b) = h(n_q + n_b)n_b \frac{h(n_q) - \bar{b}}{h(n_q) - h(n_q + n_b)}. \tag{2.7}$$

### 2.2.5 Capacity of best effort flows in normal regime

From (2.4), we can determine the lending probability $\beta(n_q, n_b)$:

$$\begin{aligned} \beta(n_q, n_b) &= \Big(h(n_q + n_b) - \bar{b}\Big) \frac{n_q + n_b}{\mathbb{E}[X^{(n_q)} | X^{(n_b)} \leq X^{(n_q)}]} \\ &= 1 - \frac{\bar{b}}{h(n_q + n_b)}. \end{aligned}$$

Rearranging the above formula yields $(1 - \beta(n_q, n_b))h(n_q + n_b) = \bar{b}$, which is intuitively correct; each QoS flow balances its bandwidth to exactly $\bar{b}$.

Then, the capacity of best effort flows in the normal regime is

$$g_b(n_q, n_b) = h(n_q + n_b)n_b + \beta(n_q, n_b)\mathbb{E}[X^{(n_b)} | X^{(n_b)} \leq X^{(n_q)}]P_q.$$

Using the same approach as in (2.5) and (2.6), we have

$$g_b(n_q, n_b) = g(n_q + n_b)P_b + (1 - \frac{\bar{b}}{h(n_q + n_b)}) \times \big(g(n_b) - P_b g(n_q + n_b)\big).$$

Now, given the total capacity of best effort flows in overloaded and normal regime we can analyze 2-D Markov chain given in (2.1). To do so we only need to characterize $g(n)$ (or $h(n)$).

**Example 1.** *In Rayleigh fading channel, signal strength has Rayleigh distribution, and random variable $Y$ representing channel SNR has exponential distribution with mean $\nu^{-1}$. Let $Z = Y^{(n)}$. Then, Shannon (or ergodic) capacity using max rate scheduling is calculated as*

$$\begin{aligned} \mathbb{E}[X^{(n)}] &= \mathbb{E}[log(1+Z)] \\ &= \int_0^\infty n(1-e^{-\nu z})^{n-1}\nu e^{-\nu z} log(1+z)dz. \end{aligned} \tag{2.8}$$

*Using the integral with incomplete gamma function $\Gamma(0,x)$,*

$$\int_0^\infty e^{-kz} log(1+z)dz = \frac{e^k \Gamma(0,k)}{k} \tag{2.9}$$

*where $\Gamma(0,x) = \int_x^\infty \frac{e^{-t}}{t}dt$ and using the binomial theorem, $g(n) = \mathbb{E}[X^{(n)}]$ is computed from (2.8) as*

$$g(n) = n\sum_{k=0}^{n-1}(-1)^k \binom{n-1}{k} \frac{e^{\nu(k+1)}\Gamma\big(0,\nu(k+1)\big)}{k+1}.$$

*Otherwise, $g(n)$ can be approximated by numerical calculation of (2.8). Fig. 2.1 shows an example of $g(n)$ and $h(n)$ in 0dB Rayleigh fading channel.*

## 2.3 Stability

In this section we address stability of our system model. First we discuss the stability in case the maximum capacity of system is unbounded, *i.e.*, $g(n) \to \infty$ as $n \to \infty$. Ideal Rayleigh fading channel falls into this category. We show that the system is stable in this case. Then, we deal with the system that has finite capacity $C$ and identify the necessary and sufficient condition of stability. Practical systems will of course fall in this second category.

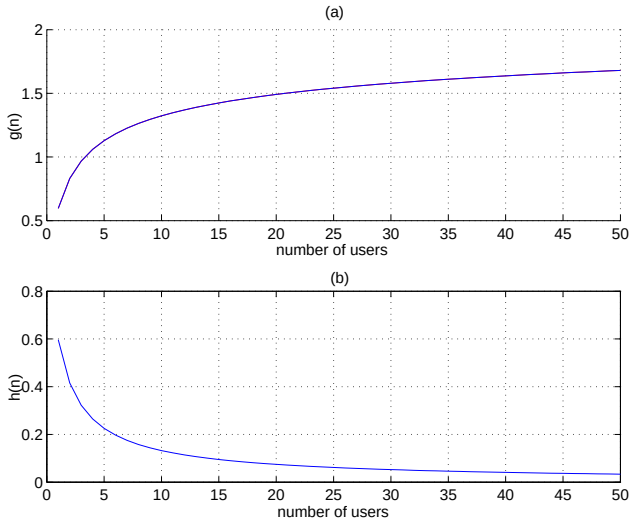



Figure 2.1: Plot of the total capacity $g(n)$ (a) and individual capacity $h(n)$ (b) of 0dB Rayleigh fading channel.

### 2.3.1 Unbounded case

**Theorem 1.** *If $g(n)$ is unbounded as $n$ goes to $\infty$,* e.g., *in ideal Rayleigh fading channel, then the system is stable for any offered load.*

***Proof of Theorem 1.*** Let $\chi_t = \big(N_q(t), N_b(t)\big)$ denote the state of our irreducible, aperiodic, continuous time 2-D Markov chain on $\mathbb{S} = \{(n_q, n_b) | n_q \in \{0, \cdots, n^*\}, n_b \in \mathbb{Z}^+\}$. Based on Foster theorem [41], since $g(n)$ is unbounded, for any offered load of best effort flows $\rho_b = \frac{\lambda_b}{\mu_b}$, we can find an $\exists l < \infty$ such that $g_b(n_q, n_b) > \rho$ for $\forall n_b > l$, which means the drift is negative with the corresponding Lyapunov function $\varphi(\chi_t) = n_b$. □

### 2.3.2 Bounded case

**Theorem 2.** *Suppose $C := \lim_{n\to\infty} g(n) < \infty$ and the maximum number of QoS users is limited to $n^*$. Then, the system is stable if and only if there exists $\exists l < \infty$ such that*

$$\mathbb{E}[g(N_q(t), l)] > \rho_q(1 - \eta_q)\bar{b} + \rho_b \tag{2.10}$$

*where* $\rho_q = \frac{\lambda_q}{\mu_q}$, $\rho_b = \frac{\lambda_b}{\mu_b}$ *and* $\eta_q = \pi_q(n^*)$, i.e., *blocking probability of QoS flows.*

***Proof of Theorem 2.*** Since the number of QoS flows is bounded we need only consider the stability of best effort flows. We model this system as 1-D queueing system where the service rate is a state dependent random process $g_b(N_q(t), n_b)$.

The necessary condition is clear in that the total influx rate should be less than the average service rate. To prove the sufficient condition, we shall use the *Saturation Rule* [6]. Let $T_{n_b}$ denote the time of the last departure from a system given it starts with $n_b$ customers at time 0 and there are no more arrivals thereafter. The Saturation Rule says that $T_{n_b}$ satisfies the strong law of large numbers, so $\lim_{n_b} \frac{n_b}{T_{n_b}}$ exists a.s. and the system is stable for the input process $\lambda_b$ if $\lim_{n_b \to \infty} \frac{n_b}{T_{n_b}} > \lambda_b$, *i.e.,* the departure rate for a saturated system exceeds the arrival rate. Let $T^l_{n_b}$ denote the stopping time from state $n_b > l$ to state $l < \infty$. Then, $T_{n_b} = T^l_{n_b} + T^0_l$. Since $T^0_l$ is finite, $\lim_{n_b} \frac{n_b}{T_{n_b}} = \lim_{n_b} \frac{n_b}{T^l_{n_b}}$, a.s. Let $s_i$ be the file size of best effort customer $i$. Since the total served bits on $[0, T^l_{n_b}]$ is less than $\sum_{i=1}^{n_b} s_i$,

$$\begin{aligned} \frac{n_b}{T^l_{n_b}} &\geq \frac{n_b \int_0^{T^l_{n_b}} g_b(N_q(t), l) dt}{T^l_{n_b} \sum_{i=1}^{n_b} s_i} \\ &= \frac{\frac{1}{T^l_{n_b}} \int_0^{T^l_{n_b}} g_b(N_q(t), l) dt}{\frac{1}{n_b} \sum_{i=1}^{n_b} s_i}. \end{aligned} \tag{2.11}$$

Then, $\lim_{n_b} \frac{1}{T^l_{n_b}} \int_0^{T^l_{n_b}} g_b(N_q(t), l) dt = \mathbb{E}[g_b(N_q(t), l)]$, a.s. since $g_b(N_q(t), l)$ is an ergodic process and $T^l_{n_b} \to \infty$, a.s. as $n_b \to \infty$ since service capacity is bounded by $C$. Also, by the strong law of large numbers, $\lim_{n_b} \frac{1}{n_b} \sum_{i=1}^{n_b} s_i = \mu_b^{-1}$, a.s. So, taking limit of (2.11) yields

$$\lim_{n_b} \frac{n_b}{T_{n_b}} = \lim_{n_b} \frac{n_b}{T^l_{n_b}} \geq \mathbb{E}[g_b(N_q(t), l)] \mu_b, \text{ a.s.}$$

Hence, if there exists $\exists l < \infty$ such that

$$\mathbb{E}[g_b(N_q(t), l)] > \rho_b, \tag{2.12}$$

then $\lim_{n_b} \frac{n_b}{T_{n_b}} > \lambda_b$. From (2.2), adding $\mathbb{E}[N_q]\bar{b} = \rho_q(1-\eta_q)\bar{b}$ to (2.12) makes (2.10) which completes the proof. □

**Corollary 1.** *Let $\bar{g}(n_q) := \lim_{n_b \to \infty} g(n_q, n_b)$. In the case where $g(n)$ is bounded, $\mathbb{E}[\bar{g}(N_q)]$ is less than $C$. So the maximum allowable influx rate into the system gets reduced by the integration of QoS and best effort flows.*

*Proof.* Corollary 1 is obvious from the *capacity gap* described in the next section. □

## 2.4 Loss in opportunism

In this section we address the negative impacts of integrating QoS and best effort traffic. We shall refer to these as the *loss in opportunism* of service integration. The fundamental reason for losing throughput from opportunism comes from balancing QoS requirements vs. opportunism. To maximize system capacity we need to schedule users with high channel rate all the time. However, if we need to meet QoS requirements, at some time slots we are forced to select sub-optimal users resulting in loss of opportunism. Hence, we have a trade-off between guaranteeing QoS and maximizing capacity.

### 2.4.1 Capacity gap

Suppose that the system has a maximum capacity $C$ and is supporting $n_q$ QoS flows. Then, best effort traffic might expect its capacity to be $C - n_q\bar{b}$. However, the maximum opportunistic capacity that best effort flows achieve is $\bar{g}_b(n_q) := \lim_{n_b \to \infty} g_b(n_q, n_b)$. This limit is determined from (2.7) as $C\big(1 - \frac{\bar{b}}{h(n_q)}\big)$. So, we have a gap between $C - n_q\bar{b}$ and $\bar{g}_b(n_q)$. We call this quantity the *capacity gap*. Given $n_q$ active QoS sessions each of which requires an average throughput of $\bar{b}$, the capacity gap $\xi(\bar{b}, n_q)$ is given by

$$\xi(\bar{b}, n_q) = \bar{b} n_q \Big( \frac{C}{g(n_q)} - 1 \Big) > 0, \;\; n_q = 1, \ldots, n^*. \tag{2.13}$$

The importance of investigating this gap lies in that it affects not only the stability region of the system as stated in Corollary 1, but also degrades the performance of best effort flows because active QoS sessions deprive best effort sessions of available capacity, more than what is expected, resulting in local instability for best effort traffic. Local instability means that conditioned on a fixed number of QoS streams $n_q$, the arrival rate for best effort traffic $\lambda_b$ exceeds the maximum service rate $\bar{g}_b(n_q)\mu_b$ and so traffic would temporarily accumulate. This usually happens when $n_q$ remains high, and as a consequence best effort flows would experience long delays. Performance implications of this will be addressed in the next section.

Let us consider some characteristics of the capacity gap. From (2.13) we see that $\xi(\bar{b}, n_q)$ is proportional to the guaranteed bandwidth per QoS flow, and as a corner case, if $\bar{b} = g(1)$, then $n^* = 1$ and $\xi(1) = C - \bar{b}$, which means the system can support only one QoS flow and no best effort flows. The shape of $n_q\big(\frac{C}{g(n_q)} - 1\big)$ depends on $g(n)$. In turn, $g(n)$ is determined by the probability density function of the channel capacity. For example, for uniformly distributed channel capacity, $g(n) = C\frac{n}{n+1}$ and $\xi(\bar{b}, n_q) = \bar{b}$, which means the system experiences a constant capacity gap, *i.e.,* independent of $n_q$. But usually the capacity gap will be an increasing function of $n_q$ in the domain of interest. [3]

**Example 2.** *This example will show that the capacity gap can be quite large in a real system. Let us define the opportunistic gain as the ratio of system capacity with and without opportunism,* i.e.,

$$\kappa := \lim_{n\to\infty} \frac{g(n)}{g(1)} = \frac{C}{E[X]}. \tag{2.14}$$

[3] If $\bar{b}$ is very small so $n^*$ can be large enough, then it can be shown not to be an increasing function, *i.e.,* eventually decreases in $n_q$.

*Then, the capacity gap of introducing a single QoS user is given by*

$$\xi(\bar{b}, 1) = \bar{b}(\kappa - 1).$$

*Suppose that we have sufficient number of best effort users so that system capacity is close to $C$. In the CDMA/HDR system described in [12], $C = 2457kbps$ and $E[X] = 659kbps$, so $\kappa = 3.75$. If we want to guarantee 300kbps per QoS flow, then capacity gap is $\xi(300, 1) = 300 \times (3.75 - 1) = 825kbps$, which is 33 % drop in capacity from 2457kbps. For $\bar{b} = 200kbps$, the gap is 550kbps corresponding to 22% drop. Fig. 2.2 shows capacity gaps for various $\bar{b}$ and $n_q$. Note that each plot has different range of $n_q$ for different $\bar{b}$ because the maximum number of QoS streams we can admit depends on $\bar{b}$. From the figure, we see that if $\bar{b} > 100kbps$, capacity gap increases as $n_q$ grows but the slopes are decreasing in $n_q$. Thus, the capacity gap impacts the system mostly when the first few QoS flows are admitted. However, if $\bar{b}$ is small such as $100kbps$, the capacity gap has a smooth peak and starts to decrease. This is because if $\bar{b}$ is small, the system can admit a sufficient number of QoS flows to generate the opportunistic capacity gain among the QoS flows. The next example illustrates the capacity gap of Rayleigh fading channels.*

**Example 3.** *Under a Rayleigh fading channel model, the impact of the capacity gap is more severe in low SNR than high SNR. Fig. 2.3 shows plots of capacity gap divided by $\bar{b}$ so as to only reflect the effect of $n_q$. We assume that a maximum of 1000 users can exist in the system and each SNR has a different bounded capacity. We see that the capacity gap of low SNR increases faster than that of high SNR. Considering the maximum capacity of low SNR is even less than high SNR, we see that the effect on capacity gap in the low SNR case is severe. This is consistent with what we see in the opportunistic gain from (2.14): $\kappa$ is 3.57 at 0dB, 2.14 at 10dB, 1.62 at 20dB, 1.41 at 30dB. Thus we see that high opportunistic gains*

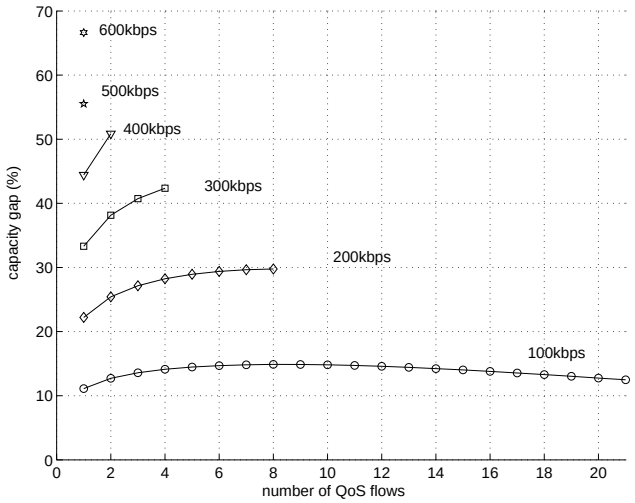



Figure 2.2: Percentage of capacity gap $\frac{\xi(\bar{b},n_q)}{C}$ for the CDMA/HDR model describe in [12]: $\bar{b} = 100, \ldots, 600$Kbps.

*will be associated with high loss in opportunism for integration. This is further reflected in the delay performance considered in next subsection.*

### 2.4.2 Delay increase

In this section, we investigate the effect of the capacity gap on the delay performance of best effort flows. In a mixed user system, the ideal capacity that best effort flows could see under an opportunistic scheduling scheme would be

$$g_b^*(n_q, n_b) = g(n_q + n_b) - n_q\bar{b},$$

*i.e.,* the overall maximum opportunistic capacity minus that given to QoS streams. However, the actual capacity seen in our model when attempting to meet QoS stream's requirements is $g_b(n_q, n_b)$, so the difference is $g_b^*(n_q, n_b) - g_b(n_q, n_b)$, which converges to the capacity gap as $n_b \to \infty$. In a dynamic system, as long as the best effort flows remain stable, offered load $\rho_b$ will be served, yet delay will depend on the character of $g_b(n_q, n_b)$.

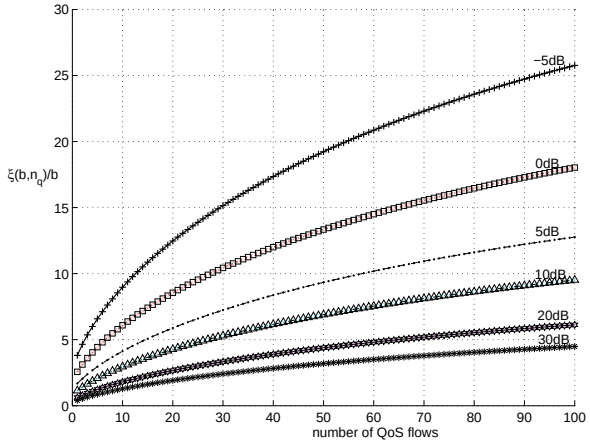



Figure 2.3: Normalized capacity gap $\frac{\xi(\bar{b},n_q)}{b}$ for various SNR under Rayleigh fading channel.

We will start, for simplicity, by evaluating the average delay of best effort users conditioned on a **fixed** number of active QoS sessions. In this case the distribution of best effort flows given $n_q$ active QoS sessions is given by

$$\begin{aligned} \pi_b(n_b|n_q) &= \pi_b(0|n_q)\Pi_{i=1}^{n_b}\frac{\rho_b}{g_b(n_q,i)} \\ E[D|n_q] &= \frac{1}{\lambda_b}\sum_{n_b=1}^{\infty} n_b\pi_b(n_b|n_q). \end{aligned} \tag{2.15}$$

As a baseline delay performance, we substitute $g_b(n_q,n_b)$ with $g_b^*(n_q,n_b)$, and we will compare the delay under various SNR, $\bar{b}$ and $n_q$. We shall only consider the case where best effort flows are locally stable, *i.e.,*

$$\bar{g}_b(n_q) > \rho_b. \tag{2.16}$$

Even though the channel is Rayleigh fading, we assume that the range of $X$ is finite since practical system can support only a finite number of users and they generate only finite opportunism. So, the delay of (2.15) is divided by $1-\eta_b$ where $\eta_b$ is blocking probability of best effort flows. We assume that up-to 1000 users can share the system.

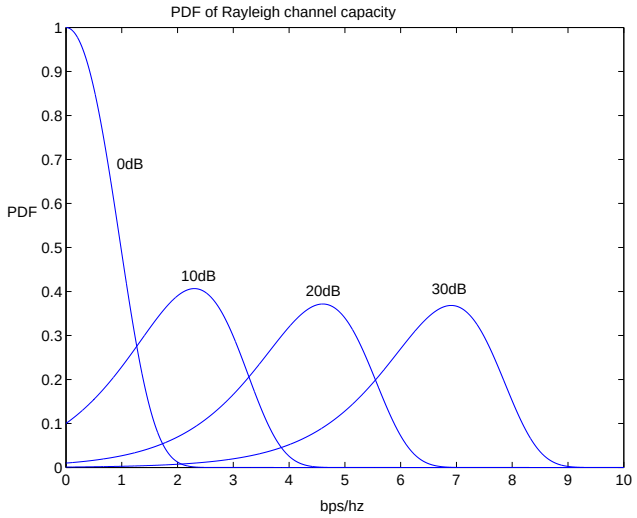



Figure 2.4: Probability density function of Channel capacity under Rayleigh fading channel(bps/hz) for 0dB, 10dB, 20dB, 30dB.

Fig. 2.5 to Fig. 2.7 show the average delay for various $\bar{b}$ and $n_q = 1$ and 5. The figures show two curves, *'real'* and baseline delay associated with $g_b(n_q, n_b)$ and $g_b^*(n_q, n_b)$, respectively. Here, $\lambda_b = 1/\text{sec}$ and $\mu_b^{-1} = 60\text{Kbytes}$ as in [21]. The baseline delay represents an ideal delay performance under mixed traffic to show the delay penalty of the integration of QoS and best effort flows, conditioned on a fixed number of active QoS sessions.

In Fig. 2.5 (a) we see that as $\bar{b}$ grows, real and baseline delays increase and after some point the system is unstable. In Fig. 2.5 (b) we present delay difference ratio of real and baseline delays. Even for $n_q = 1$, we see penalty in performance. For example, with SNR $= -3$dB, $C = 1240$kbps, $g(1) = 290$kbps and $\bar{b} = 120$kbps, we see that real average delay is over 10 sec where as the baseline delay is around 5 sec. This is more than 100% increase. A single QoS flow can substantially increase average delays of best effort flows. If $n_q = 5$, the delay difference ratio grows more rapidly.

Comparing Fig. 2.5 through Fig. 2.7, one can see that the delay difference ratio is

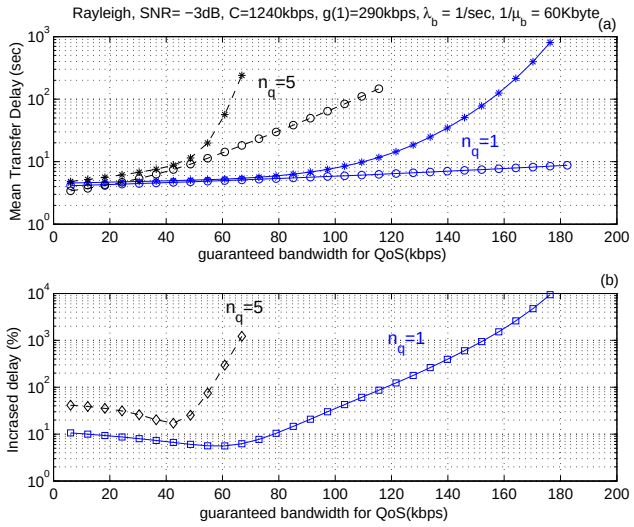



Figure 2.5: Delay comparison at -3dB, Rayleigh fading channel. (a) real delay ($*$) and baseline delay ($\circ$). (b) Delay difference ratio.

improved as the SNR increases. For example, if we compare SNR = -3dB, 0dB, 10dB at $n_q = 1$ and $\bar{b} = 0.4 \times g(1)$, then the ratios are 100% at -3dB, 15% at 0dB and 1% at 10dB.

*Remark* 2.4.1. From the above delay performance comparison, we can infer that guaranteeing bandwidth to a fixed number of QoS users will make the delay of best effort longer than one might expect. This becomes severe for lower SNR, higher $\bar{b}$ and large $n_q$.

## 2.5 Admission control for best effort flows

In this section we consider admission control for best effort flows to improve delay performance. We propose a simple admission control strategy to reduce the impact of local instability where the number of best effort flows might temporarily grow.

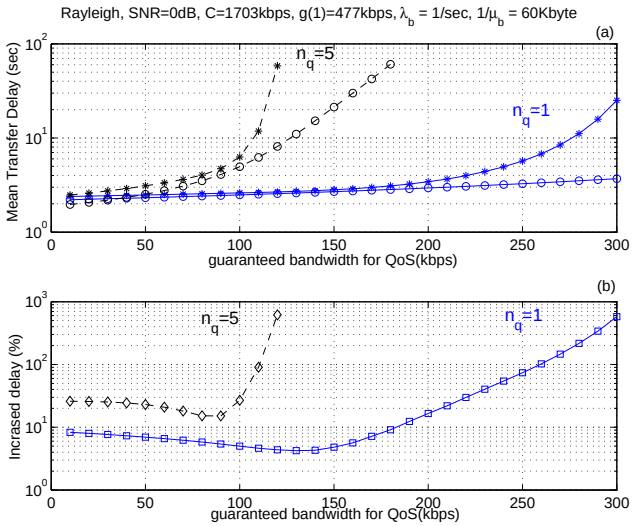



Figure 2.6: Delay comparison at 0dB, Rayleigh fading channel. (a) real delay ($*$) and baseline delay ($\circ$). (b) Delay difference ratio.

### 2.5.1 Delay and local instability

In the previous section we considered the delay performance of best effort flows given a **fixed** number of active QoS sessions and saw that it deteriorates quickly in the number of QoS flows. Delays get higher as $n_q$ increases and eventually may be locally unstable. Thus, if QoS flows remain in the system for a long time, best effort flows are not served much and their numbers may grow until best effort flows recover their capacity, *i.e.,* QoS sessions leave the system. As mentioned earlier this phenomenon is called *local instability* [38]. It is not unique to wireless network. It is common in bandwidth sharing systems where best effort flows are preempted by QoS flows. However, it is more serious in opportunistic wireless systems. Suppose that both of wired and wireless system have a maximum capacity $C$. For some given $n_q$, it is possible that a wired system does not experience local instability while wireless system does. This is because the wireless system needs a large number of flows to achieve opportunistic capacity. Clearly for every value of $n_b$, the capacity of best effort

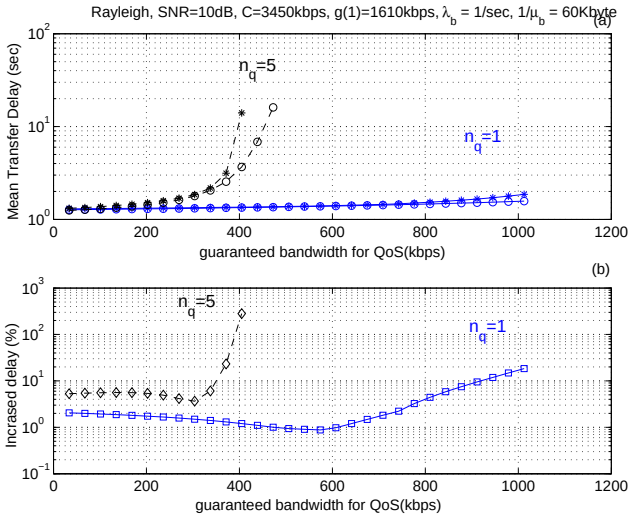



Figure 2.7: Delay comparison at 10dB, Rayleigh fading channel. (a) real delay ($*$) and baseline delay ($\circ$). (b) Delay difference ratio.

flows is smaller than that of a wired system. Furthermore, no matter how large $n_b$ is, we have a capacity gap $\xi(\bar{b}, n_q)$ that prevents $g_b(n_q, n_b)$ from reaching the capacity of the wired network.

Fig. 2.8 illustrates a typical example of local instability for a Rayleigh fading channel at offered load / $C = 0.79$ with $\bar{b} = 100$kbps, $C = 1.31$Mbps and $n^* = 10$. The joint distributions were computed numerically for the 2-dimensional Markov chain [85]. Since $n_b$ is finite in a real system, this example assumes that $n_b$ is limited by 30. Then, local instability results in accumulation of best effort flows at the boundary. In Fig. 2.8 (b) and (c) we see two peaks in the stationary distribution $\pi(n_q, n_b)$. The first peak at $(n_q, n_b) \simeq (2.5, 2.5)$ is preferred while the second peak at $(n_q, n_b) \simeq (7, 30)$ is problematic. For $\lambda_q = 0.023$/sec, $\mu_q^{-1} = 180$ sec, $\lambda_b = 1.3$/sec, $\mu_b^{-1} = 60$Kbytes, we see that the state $(n_q, n_b)$ is oscillating between two peaks along with drift arrow. If we decrease $\lambda_b$ or $\lambda_q$, then the first peak becomes dominant and the second peak diminishes. Conversely, if we increase $\lambda_b$ or

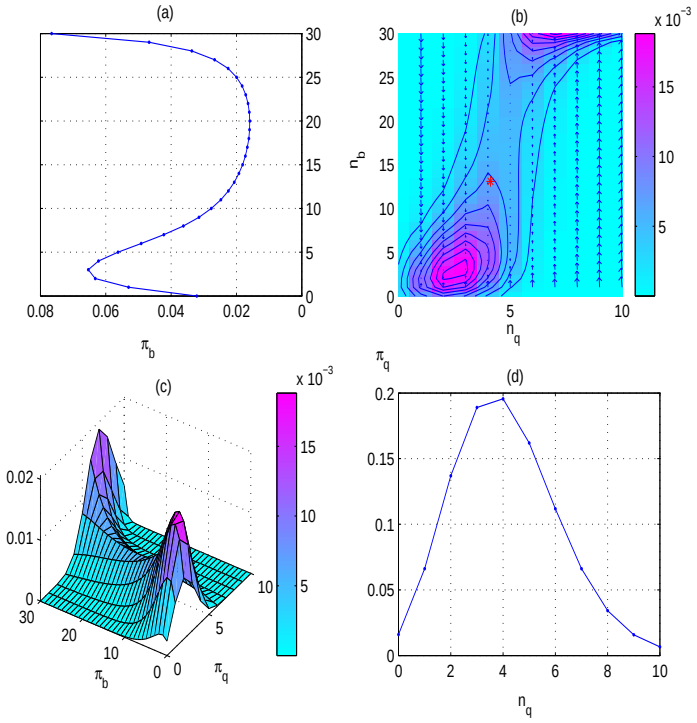



Figure 2.8: An example of local instability (a) $\pi_b(n_b)$ (b) $\pi(n_q, n_b)$ contour and drift vector (c) $\pi(n_q, n_b)$ in 3-D view (d) $\pi_q(n_q)$ : $\lambda_q = 0.023$/sec, $\mu_q^{-1} = 180$sec, $\lambda_b = 1.3$/sec, $\mu_b^{-1} =$ 60Kbyte, load ratio $= 0.79$, $\bar{b} = 100$kbps, $C = 1.31$Mbps, SNR 0dB Rayleigh fading channel.

$\lambda_q$, the first peak gradually disappears and the second peak dominates. This appears as to be a weak form of metastability, *i.e.,* where the system sees two operating regimes that are likely to jumps between them. This type of behavior is undesirable in practice, particularly if one of the modes corresponds to a poor performance, *e.g.,* long delay for best effort.

### 2.5.2 Admission control for best effort flows

One way to preclude such *metastable* behavior is to ensure that the capacity of best effort flows is always greater than the offered load, *i.e.,* (2.16). However, this approach is not preferred because we need to block QoS flows before $n_q$ reaches $n^*$. Another way is to apply admission control for best effort flows [10, 11]. For example, if the system state is

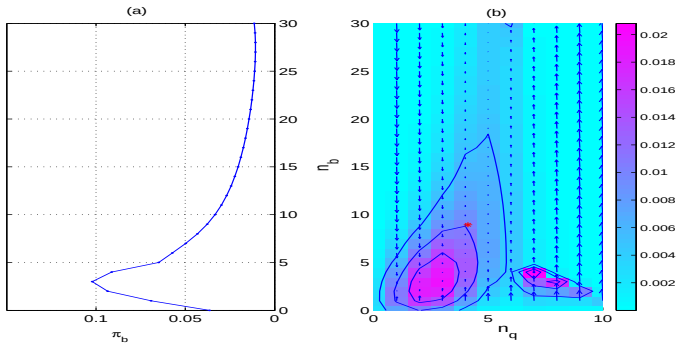



Figure 2.9: An example of admission control applied for Fig. 2.8 for best effort flows (a) $\pi_b(n_b)$ (b) $\pi(n_q, n_b)$ contour and drift vector: $\theta = 0.5\mu_b^{-1}$.

$(n_q, n_b)$, a new best effort flow might be blocked if

$$\rho_b > \theta + g_b(n_q, n_b + 1)$$

where $\theta$ is an admission control threshold. So, if the net influx rate of best effort flows is below some threshold, we admit all of them. As shown in Fig. 2.9 this *simple* admission control can alleviate the performance impact of local instability effectively – we see that the undesirable peak is eliminated.

Another importance of admission control is the expansion of capacity region. We define the capacity region as the amount of admitted load of best effort flow under a given average delay constraint. Fig. 2.10 exhibits the capacity region and call blocking probability of QoS flows for a fixed offered load of QoS traffic $\rho_q\bar{b}$. The maximum average delay constraint is 20 sec. We see that the capacity region under the delay constraint is reduced relative to the theoretical limit of $C-\rho_q\bar{b}(1-\eta_q)$ assuming no capacity gap. Nevertheless the capacity region expands considerably. In this plot, the maximum number of best effort flows is assumed to be 100. The expansion of capacity region is meaningful, in that the service provider can in principle achieve more throughput under the same delay constraint and thus generate more revenue. Note that admission control strategy enhances the capacity region

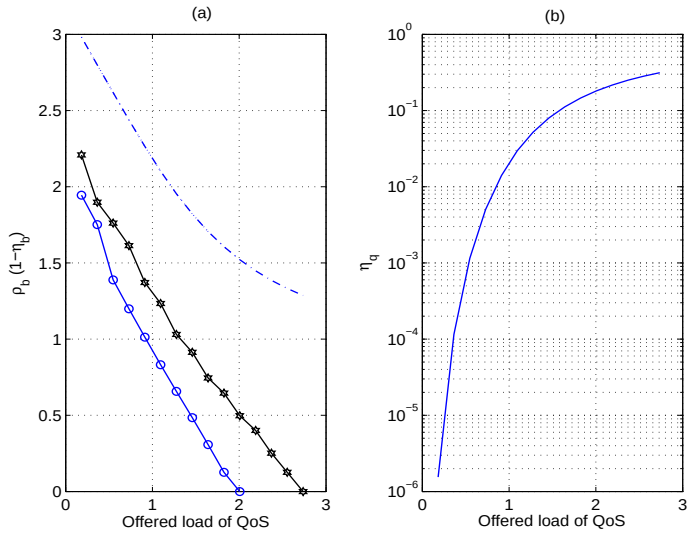



Figure 2.10: Capacity region expansion by admission control at delay constraint = 20 sec for 10dB Rayleigh fading channel.(a) Admitted load of best effort flows with and without admission control: Ideal upper bound with no opportunism and no delay constraint (dashed), with CAC ($*$) and without CAC ($\circ$) (b) Call blocking probability of QoS flows

with an increased blocking probability for best effort flows.

## 2.6 Conclusion

We have explored the flow-level dynamics and performance seen by a mixture of QoS and best effort flows sharing an opportunistic wireless system. In doing so, we proposed a new opportunistic scheduling scheme/model based on the concept of bandwidth borrowing/lending from/to best effort flows which enables the scheme to ensure a mean throughput in every time slot to QoS streams which is pertinent for the case where users see roughly homogenous channel variations. We evaluated the stability and the flow-level performance in this system. The results suggest that integrating QoS and best effort flows may degrade system performance in crucial aspects; reduction of the stability region, a gap in the capacity available to best effort traffic and increased file transfer delay. These nega-

tive impacts are referred to as *loss in opportunism*, and we found them to be proportional to the opportunistic gains, the guaranteed bandwidth, and to the number of QoS flows. We note, however, that these losses would be reduced in a system with a high SNR, assuming a Rayleigh fading channel model.

# Chapter 3

# Leveraging Dynamic Spare Capacity to Conserve Mobile Terminals' Energy

## 3.1 Introduction

In Chapter 2 we investigated tradeoffs between flow-level performance (QoS and capacity) in a wireless network integrating services. In this Chapter we consider tradeoffs between flow-level energy expenditure and capacity. Though future wireless systems promise to support higher capacity, this will be achieved, in most cases, at the expense of higher energy consumption resulting in shorter battery lifetimes for mobile terminals. So, work on energy conservation has become a critical and active research area. Unlike previous research on energy conservation in sensor and wireless local area networks (LAN) [34, 35, 39, 73, 93, 95, 97], we focus on energy saving techniques for broadband cellular systems, e.g., WiMAX or 3GPP-LTE. Specifically, we focus on reducing uplink RF transmission energy recognizing it is one of the main contributors to battery consumption (e.g., 60% in time division multiple access (TDMA) phones [97]). Other energy consumption such as display or microprocessor, etc., are not considered.

Not unlike most networking infrastructure (particularly that supporting data), wireless access networks are unlikely to be fully utilized all the time. Indeed as a result of time varying, non-stationary loads, or unpredictable bursty loads these networks are often overdesigned to be able to support a peak load condition, and so often underutilized. For example Internet service providers' networks see a long term utilization as low as 20% [64]. Similarly a substantial fraction of Wi-Fi hotspot capacity is unused [49]. More generally, due to the

high variations in capacity that a wireless access system can deliver to various locations in its coverage area, e.g., up to three orders of magnitude difference, one can also expect high variability in the system load [4, 98]. Furthermore in some cases, e.g., cellular networks, a substantial amount of bandwidth is set aside to ensure that calls are not dropped during handoffs; for example, a 0.5% of call dropping probability requires 30% of system capacity to be reserved [86]. This further contributes to underutilization of the system, even when the loads are heavy. The central premise of our work is that wireless access networks whose resources are occasionally underutilized can provide their users a better service/value by reducing mobile terminal energy consumption while causing a controlled or imperceptible impact on user's perceived quality of service (QoS).

The basic idea towards conserving energy is as follows. As a rough model for the relationship between power and capacity, consider Shannon's capacity formula

$$x = w \log\left(1 + \frac{p_{\text{out}}\, g}{\sigma^2}\right) \Leftrightarrow p_{\text{out}} = \left(\exp\left(\frac{x}{w}\right) - 1\right) \frac{\sigma^2}{g} \tag{3.1}$$

where $x$ is the transmission rate, $w$ is the spectral bandwidth, $p_{\text{out}}$ is the output power of the RF power amplifier, $g$ is the channel gain and $\sigma^2$ is the noise power. Note that the output power (defined as the power dissipated into the air) is an exponential function of the transmission rate. Thus a small back off in the transmission rate $x$ results in an exponential reduction in output power. The cost of doing so is a slow down in transmissions. So if users are insensitive to such slow downs a system can realize beneficial tradeoffs.

Users or applications are insensitive to slow downs if the expected quality of service is met. For real-time or streaming services this means meeting the required transmission rates. Thus when a wireless access point is underloaded one can back off from a user's *individual instantaneous* peak transmission rate without impacting the perceived performance. By contrast, for file transfers, reducing transmission rates will impact file transfer delays, yet

may still be desirable if noticeable energy savings can be achieved. Specifically, for the downlink, fast transmission may be critical to ensure users' satisfaction with web browsing applications or file download speeds. However, on the uplink, e.g., uploading of files such as pictures or emails, users may be quite delay-tolerant, so much so that transfers could be carried out as *background processes*. For best effort traffic it makes sense to set a target *average* throughput users might expect over a given time window. This recognizes the fact that file transfer delays depend on average throughput rather than instantaneous transmission rate. The time window reflects the time scales on which such averages make sense, e.g., seconds to minutes. The bigger the time windows the more flexibility a wireless system has in exploiting transient underloads to conserve energy.

In this work, we focus on *dynamic* user populations and traffic loads in a cellular system where new flows, either real-time sessions or file transfers, are initiated at random and leave the system after being served – these are referred to as flow-level dynamics [21], see Fig. 3.1. Dynamic systems are, in general, hard to analyze and have not been studied as extensively as the static versions, i.e., with a fixed set of backlogged users.

To better understand the challenges involved, consider a TDMA system supporting, a stationary dynamic load, of file transfer requests. If one slows down the uplink transmission rate to save energy then the number of users in the system may grow, resulting in excess power consumption associated with users that *idle* while awaiting transmission. Indeed although ideally idling users turn off their transmission chains, in practice they still consume power due to leakage current[1] [57, 94]. Hence, in a dynamic system, if the transmission rates are excessively reduced, the number of users that are idling may accumulate resulting in excessive overall *idling power* consumption. This makes tradeoffs between en-

[1]Idling power consumption depends on the specific power amplifier design. For example, power amplifier for WiMAX from Analog Devices consumes 2.5 to 25 mW during idling period [94].

ergy conservation and delay somewhat complex. Another challenge is to capture the power consumptions from several components in RF transmission chain of *active* users (as opposed to *idling* users). Even though the power amplifier is the main consumer of power, other analog devices such as mixers, filters, local oscillators, D/A converters, may also consume non-negligible power called *circuit power* [34, 57].

Earlier research on power control mainly focused on controlling interference rather than reducing energy consumption, i.e., sustaining a required signal to interference ratio (SIR) for reliable voice connections [8, 44, 121]. Energy-efficient power control was first explored in the context of sensor networks [39, 95]. The authors proposed 'lazy scheduling' where packets are transmitted as slowly as possible while meeting packet delay constraints. Lazy scheduling performs smoothing on arriving packets and thus makes output packet flows less 'bursty.' This leads to significant energy savings.

The work in [13, 30], [97] further explore energy-delay tradeoffs under various scenarios; they study minimizing the average transmit power subject to average buffer delay constraints under two state Gilbert-Elliot channels, fading channels, and additive white Gaussian noise (AWGN) channels, respectively. In fading environments, the use of opportunistic transmission to save energy was studied in [48, 68, 116, 124]; i.e., when the channel is good, transmit power is increased. However, the above work neglects circuit power, idling power and flow-level dynamics.

Recent results show that if circuit power is taken into account, circuit energy consumption increases monotonically as the transmission time grows [34, 76, 93, 123]. Thus, we cannot slow down the transmission rate arbitrarily, and indeed, there exists an energy-optimal transmission rate. In solving this optimization problem, the work in [34] focuses on the physical modulation techniques with a single sender and receiver pair for sensor

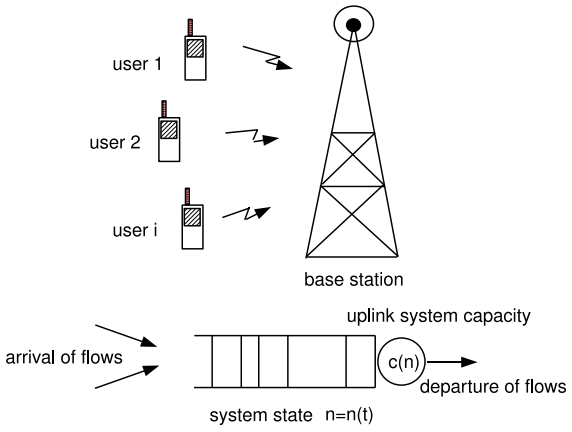



Figure 3.1: Flow-level model for uplink transmission in a dynamic system. One user corresponds to one flow.

networks. Cross-layer optimizations are also proposed with a view on capturing the physical and medium access control (MAC) layer in small scale sensor networks [35] and in wireless LANs [93], and further up-to the routing layer [73]. Energy-efficient transmission strategy for orthogonal frequency division multiple access (OFDMA) system considering circuit power was proposed in [80]. However, previous work has addressed static systems, not dynamic systems, and thus could not capture the coupling between power backoff and its impact on system dynamics. For example, idling power consumption may become huge when the number of users accumulate, e.g., 10–100, albeit only occasionally [57].

***Contributions.*** We highlight the contributions of this chapter as follows.

First, based on a detailed transmit power model, we show that idling power has a substantial impact on energy efficiency when reducing transmission rate changes the system dynamics, e.g., in the case of file transfers. Previous work has focused on static systems, thus only the impact of circuit power was exhibited. However, we show that, as the load increases, circuit power is asymptotically negligible in the case of dynamic systems. Nevertheless, circuit power remains important in the case of systems supporting real-time

sessions.

Second, we show how energy savings scale with the average load in a stationary system. Our flow-level queueing model captures the dynamic behavior of real systems and indicates that energy can be significantly saved when the system is underloaded. For example, in the case of real-time sessions, when the call blocking probability is less than 0.1%, more than 50% of energy can be saved without compromising user-perceived performance. In the case of file transfers, we demonstrate that 35–75% of the energy can be saved depending on the loads and target throughput.

Third, we propose two practical energy saving techniques for real-time sessions and file transfers, respectively. In the case of real-time sessions, we formulate the problem as a convex optimization and solve it in an iterative fashion exhibiting superlinear convergence. Our energy-optimal transmission policy minimizes the adverse impact of circuit power while reducing the output power level of mobile terminals at the cell edge, e.g., by 15 dB. This in turn can be beneficial in mitigating inter-cell interference. In the case of file transfers, we propose an energy-efficient algorithm that exploits energy-delay tradeoff considering users' preferences. The proposed algorithm addresses the possibly unfavorable impact of idling power.

Our work is significant in its wide applicability to future broadband wireless systems, which promise to support higher capacity but, in most cases, at the expense of much higher energy expenditures.

***Organization.*** This chapter is organized as follows. In Section 3.2, we describe our system model and assumptions. Section 3.3 is devoted to the optimization for energy-efficiency for real-time sessions. We address the energy savings for file transfers in a dynamic system in Section 3.4 and conclude the work with Section 3.5.

## 3.2 System model

### 3.2.1 Assumptions

We consider a centralized wireless communication system where a base station serves multiple mobile terminals, e.g., WiMAX or 3GPP-LTE. For simplicity, we assume that the system is shared via TDMA. Note, however, that the same approach is applicable in the context of frequency division multiple access (FDMA), and furthermore, already applied to multiple input multiple output (MIMO) systems [56]. We define a *time frame* as the fixed time period during which every user is scheduled once. We use $t$ to denote the time frame index and $s$ for continuous time. Since energy savings are more important at mobile terminals than at the base station, we focus on uplink transmissions as shown in Fig. 3.1. Our framework is also applicable to downlink transmissions to conserve the energy consumption at the base station.[2]

Our goal is to reduce the energy consumed in uplink RF transmission of mobile terminals. We assume that the transmission rate is continuous, and the power/rate mapping function is convex and differentiable.

For the channel models in this chapter, we shall use both static and fading models for different purposes. In particular, when we conduct a stationary analysis of the dynamic system we assume homogeneous users that have the same *static* channel gain for analytical tractability. This static channel assumption is in place for queuing analysis performed in Section 3.4.2 to Section 3.4.5 and is dropped in Section 3.4.6 where we describe the energy saving algorithms. In Section 3.4.8, the proposed algorithms are evaluated assuming i.i.d. Rayleigh fading channels. Note that Rayleigh fading channel is used for analytical

[2]However, saving energy at the base station may increase the energy consumption at the mobile terminals. This is because power consumption at reception is roughly independent of the receiving data rate, and fast transmission is more beneficial in saving energy at mobile terminals by reducing the circuit and idling energy consumption.

purpose, but it is not a good model for practical systems; Rayleigh fading model is valid only for narrowband channel with a sufficiently large number of multipath components, which make both in-phase and quadrature terms independent and identically distributed Gaussian random variables by central limit theorem. Then, the amplitude of the signal can be modeled as a random variable having Rayleigh distribution. Note that Rayleigh fading channel is an ideal model, and real systems usually have correlated channels.

### 3.2.2 Flow-level model for system dynamics

We will study a dynamic system where the number of ongoing users varies with time. User sessions/flows arrive to the system according to a Poisson process with rate $\lambda$ and leave after being served. Such models are traditionally used in modeling flow-level dynamics in communication networks, see [20, 21, 45, 59]. We will *separately* consider the case where a flow corresponds to real-time session or a file transfer, in Section 3.3 and Section 3.4 respectively. The system dynamics are captured by a *flow-level queueing model* shown in Fig. 3.1 which tracks the arrival and departure process of users (or flows), see e.g., [21]. We will assume each user corresponds to a single flow, and so user and flow are used interchangeably. We refer to the number of flows in the system $n$ as the system's state in the sequel.

### 3.2.3 Minimizing energy consumption in a stationary system

Our objective is to minimize the energy consumption of a *typical*[3] flow in a ***stationary*** system. Let $(F(s), s \in [0, T])$ be a random process modeling the power consumption of a typical flow, starting at 0 and whose typical sojourn time is modeled by a random variable

[3]For simplicity we define performance metrics for typical flows directly in terms of appropriate random variables rather than introducing Palm probabilities.

$T$. Letting $J$ denote the energy consumption of a typical flow, our goal will be to minimize

$$E[J] = E\left[\int_0^T F(s)ds\right] \tag{3.2}$$

subject to either sustaining minimum rate requirements for real-time sessions or achieving an average throughput for file transfers. Minimizing (3.2) is not straightforward because both $T$ and $F(s)$ may depend on system dynamics; in particular in the case of file transfers they are not independent, i.e., power backoff may reduce $F(\cdot)$ but increase $T$. However for a stationary system, minimizing the average energy consumption of a typical flow is equivalent to minimizing the average *system power* consumption. This is akin to Little's law and formally stated as follows.

**Theorem 3** (**Energy-power equivalence**). *Let $P$ be a random variable denoting the stationary system power consumption, $J$ be a random variable denoting the energy consumed to serve a* typical *user's flow, and $\lambda$ be the arrival rate of users/flows to the system. Then, if the system is stationary,*

$$E[P] = \lambda E[J]. \tag{3.3}$$

*Proof.* This result is intuitive and can be shown via Brumelle's theorem [65], which is a generalized version of Little's law. □

Based on Theorem 3, we below focus on minimizing the average system power consumption which in turn minimizes the average energy consumed by a typical mobile terminal.

### 3.2.4 Transmission power model

A key element of our work is to have a proper transmit power model. The power consumption in a real transmission chain depends on various factors such as drain efficiency

Table 3.1: Notation Summary

| | |
|---|---|
| $t$ | time frame index (discrete) |
| $s$ | time variable (continuous) |
| $i$ | user index (can be a subscript) |
| $\mathcal{A}$ | a set of ongoing users (= flows) |
| $n :=$ | $\lvert\mathcal{A}\rvert$, the number of flows in $\mathcal{A}$ |
| $x$ | instantaneous transmission rate |
| $w$ | spectral bandwidth |
| $\eta$ | drain efficiency |
| $g$ | channel gain |
| $\sigma^2$ | noise power ($=N_0 w$). |
| $\gamma :=$ | $\frac{\eta g}{\sigma^2}$, SNR with unit transmit power |
| $p_{\text{dc}}$ | circuit power |
| $p_{\text{idle}}$ | idling power |
| $\xi :=$ | $p_{\text{dc}} - p_{\text{idle}}$ |
| $\lambda$ | arrival rate of files |
| $\mu^{-1}$ | mean file size |
| $\rho :=$ | $\frac{\lambda}{\mu}$, traffic load |
| $q$ | a desired or target throughput per user |
| $c_{\max}$ | maximum system capacity |
| $p_{\max}$ | maximum *output* power |

of the RF power amplifier and associated circuit blocks [34,57]. It also depends on classes of power amplifiers, modulation schemes and power-saving mechanisms [74]. To have a realistic but also analytically tractable power model, we assume that the power consumed by the power amplifier is linearly dependent on output power of power amplifier, i.e., constant drain efficiency [34]. Then, the power equation $f(x)$ at transmission rate $x$ can be derived from (3.1) to give

$$f(x) \;=\; \begin{cases} \left(\exp(\frac{x}{w}) - 1\right) \frac{\sigma^2}{\eta g} + p_{\text{dc}} & (\text{active},\ x > 0) \\ p_{\text{idle}} & (\text{idling},\ x = 0), \end{cases} \tag{3.4}$$

where $\eta$ is the drain efficiency, which is defined as the ratio of the output power and the power consumed in the power amplifier; $p_{\text{dc}}$ is the circuit power; and $p_{\text{idle}}$ is the idling power [34, 57]. To simplify our notation, we let $\gamma = \frac{\eta g}{\sigma^2}$, i.e., the signal-to-noise ratio (SNR)

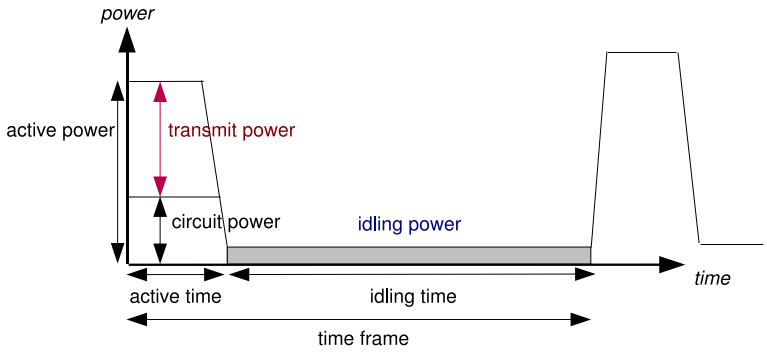



Figure 3.2: Transmission power model in TDMA systems.

when the transmit power, defined below, is 1. We summarize our terminologies as follows.

#### 3.2.4.1 Active power

When a user is transmitting, the active power is the collective power consumption in the transmission chain, i.e., the sum of the transmit power and circuit power as shown in Fig 3.2.

#### 3.2.4.2 Transmit power

We refer to $\frac{\exp(\frac{x}{w})-1}{\gamma}$ as the *transmit power* which captures the power consumed in the power amplifier. Transmit power is the main factor of power consumption in the transmission chain and equal to the output power divided by the drain efficiency.

#### 3.2.4.3 Circuit power $p_{\mathbf{dc}}$

The circuit power $p_{\mathrm{dc}}$ includes several circuit blocks in the transmission chain and remains almost constant irrespective of the transmission rate $x$. It is modeled in [34,57], by $p_{\mathrm{dc}} = p_{\mathrm{dac}} + p_{\mathrm{mix}} + p_{\mathrm{filt}} + p_{\mathrm{syn}}$, where $p_{\mathrm{dac}}$, $p_{\mathrm{mix}}$, $p_{\mathrm{filt}}$, $p_{\mathrm{syn}}$ stand for the power consumption from a digital-to-analog converter, a mixer, a filter, a frequency synthesizer, respectively.

Table 3.2: System Parameters

| | |
|---|---|
| $p_{\text{dac}} = 15.6$ mW | $\eta = 0.2$ |
| $p_{\text{mix}} = 30.3$ mW | $p_{\max} = 27.5$ dBm |
| $p_{\text{filt}} = 20.0$ mW | $w = 1$ MHz |
| $p_{\text{syn}} = 50.0$ mW | time frame = 5 ms |
| $p_{\text{dc}} = 115.9$ mW | $N_0 = -174$ dBm |
| $p_{\text{idle}} = 25$ mW | $\mu^{-1} = 60$ kbytes |

#### 3.2.4.4 Idling power $p_{\text{idle}}$

Recall that our focus herein is on TDMA systems; one user transmits at any time instance, and all other users wait to be scheduled. Users who do not transmit but wait are said to be *idling*, as opposed to *active*. As shown in Fig. 3.2, idling users turn off their transmission circuits and power amplifier to save energy, but they still consume idling power $p_{\text{idle}}$, ranging from a few to tens of mW, due to leakage currents [94]. Even though $p_{\text{idle}}$ could be negligible in a static system, it remains non-negligible in a dynamic system [57]. We will see the impact of idling power, particularly for the case of file transfers in Section 3.4. Notation are summarized in Table 3.1.

### 3.2.5 Discussion about practical issues

**Power-related parameters:** The power-related parameters in our model are summarized in Table 3.1. Note that these parameters can vary depending on the deployment scenario, the power amplifier design technologies, and specific implementation details, etc. For the case of cellular systems, which is the main focus of this dissertation, the above parameters are similar to those suggested by other researchers in the literature. For example, the Intel wireless standard group has studied improvement of client's energy efficiency for WiMAX terminals assuming the following: circuit power 100 mW and $p_{\max} = 33$ dBm [52, 82]. Note that wireless LAN have somewhat different parameters numbers, see

Table 3.3: System Parameters

| | |
|---|---|
| $p_{\text{dac}} = 15.6$ mW | $\eta = 0.2$ |
| $p_{\text{mix}} = 30.3$ mW | $p_{\text{max}} = 27.5$ dBm |
| $p_{\text{filt}} = 20.0$ mW | $w = 1$ MHz |
| $p_{\text{syn}} = 50.0$ mW | time frame $= 5$ ms |
| $p_{\text{dc}} = 115.9$ mW | $N_0 = -174$ dBm |
| $p_{\text{idle}} = 25$ mW | $\mu^{-1} = 60$ kbytes |

Table 3.4: WLAN transceiver parameters

| Mode | 802.11b | 802.11a | 802.11g |
|---|---|---|---|
| Sleep | 132 mW | 132 mW | 132 mW |
| Idle | 544 mW | 990 mW | 990 mW |
| Receive | 726 mW | 1320 mW | 1320 mW |
| Transmit | 1089 mW | 1815 mW | 1980 mW |

Table 3.4 quoted from [93]. The lack of a centralized controller makes the sleep and idle power consumption much higher versus that of cellular systems.

**Power function $f(x)$:** Our power function $f(x)$ is continuous. However, a step-wise continuous function of $f(x)$ that reflects adaptive modulation and coding (AMC) would better capture the characteristics of a system. The work in [32, 34] considers this for the case of a single transmitter/receiver pair. Since we address the problem in a dynamic system, we abstract for simplicity the rate-power tradeoff achievable at the physical layer as a continuous function based on Shannon's capacity formula as in [52, 81, 82]. Note that the spectral efficiency of AMC is a function of SNR and BER as given for example, by [48]:

$$\frac{R}{B} = \log_2\left(1 + \frac{-1.5}{\ln(5BER)}SNR\right)$$

where $R$ is the rate and $B$ is the spectral bandwidth. So, the power function can be well captured by an *exponential* function of the transmission rate, with a constant power gap $K = \frac{-1.5}{\ln(5BER)}$ between the ideal Tx power vs the real Tx power. Since practical systems and ideal systems are related by a scaling factor $K$, energy saving principle is not changed by such a scaling, see [34, 35, 80]. Hence, we use Shannon capacity for simplicity.

**Low SNR case:** Exponential relationship between the power and rate is a key characteristic enabling energy saving. It should be noted that energy savings can be achieved only if the SNR is relatively high. In our simulations we found that we can exploit the energy-delay tradeoff when SNR was at least 7 dB. This is because the power and rate relationship becomes roughly linear in low SNR regime (e.g., around 0 dB). When the relationship is exactly linear, then there is no energy saving to be had from reducing the transmission rate. In fact, faster transmission is beneficial as it minimizes the circuit energy consumption.

**Implications of energy savings:** To translate the energy savings in transmit chain into an extension of the battery lifetime – which is the ultimate objective of energy savings for mobile terminals – we need to know how much energy is consumed from transmission versus other energy consumption on the device including display, microprocessor, etc. While it has been reported that in a typical TDMA phone, approximately 60% of the battery consumption can be attributed to the transmission RF amplifier, it depends on the transmission technology as well. Generally, it has been shown that the radio interfaces, including bluetooth, Wi-Fi, and cellular communications, account for more than 50% of the overall system energy budget [3, 82], but of course this depends on the usage scenario. Hence, in describing the overall benefit of energy saving technique, we only consider the reduction in the transmit energy. We use a mean file size of 60 kbytes, which is suggested in the work of [21] modeling wireless systems. Another way of expressing the energy consumption/savings is using the notion of *energy-per-bit* rather than energy-per-flow, or *bit-per-Joule* as in [52, 82].

**Ideal vs real power amplifier:** For analytical tractability, we have assumed a constant drain efficiency $\eta$, which is commonly assumed in other work [32, 34, 52, 81, 82].

However, the efficiency of real power amplifiers is not constant. The efficiency and linearity of the power amplifier depend on the class of the power amplifier (PA), and generally speaking, efficiency is high when the output power is also high. Active research is still going on to improve the efficiency and linearity of the PA and also make the efficiency constant. For example, adaptive bias and supply of the RF PA provides good efficiency over the output power range of roughly $0 \sim 30$ dBm [74]. Without this technique, the range is roughly $10 \sim 30$ dBm. We expect the efficiency will be progressively improved, in which case the energy saving techniques developed in [32, 34, 52, 56, 58, 81, 82] including our work become more promising.

## 3.3 Energy Savings for Real-time Sessions

In this section, we consider realizing energy savings in systems supporting real-time, e.g., video/voice, sessions on the uplink. We show that the energy-optimal transmission policy is given by a dynamic policy determined by convex optimization problems associated with *fixed* user populations.

### 3.3.1 Problem formulation

We assume that the arrivals of real-time sessions follow a Poisson process with arrival rate $\lambda_s$ and have holding times which are identical, independent with mean $\mu_s^{-1}$. (Note that the distribution of the holding time is not necessarily exponential.) Let $r_i$ be the session rate requirement and $x_i$ be the instantaneous uplink transmission rate of user $i$. Then, in a TDMA system, the fraction of time user $i$ is active is $r_i/x_i$. Let $c_i$ be the maximum feasible transmission rate for user $i$, which depends on the maximum output power $p_{\max}$. Then, $c_i = w\log\left(1 + \frac{p_{\max} g_i}{\sigma^2}\right)$.

We assume that call admission control allows a new user into the system only if

there are resources to support the request, e.g.,

$$\sum_{i \in \mathcal{A} \cup \{k\}} \frac{r_i}{c_i} \leq 1 \tag{3.5}$$

where $\mathcal{A}$ denotes the set of ongoing users and $k$ is a new user (either new call or handoff). Let $n^*$ be the maximum number of users determined by a proper call admission control. From the insensitivity property, irrespective of the distribution of holding times, the stationary distribution is the same as that of an $M/M/m/m$- queue, i.e., the distribution $\pi_s(n)$ is simply given by [114]

$$\pi_s(n) = \pi_s(0) \frac{\rho_s^n}{n!}, \text{ for } n = \{0, \cdots, n^*\} \tag{3.6}$$

where $\rho_s := \frac{\lambda_s}{\mu_s}$ and $\pi_s(0) = \left[\sum_{n=0}^{n^*} \rho_s^n \frac{1}{n!}\right]^{-1}$. The blocking probability of real-time sessions is given by *Erlang-B formula* as $\pi_s(n^*)$ [14].

From Theorem 3, our objective is to minimize the average system power consumption $E[P]$ while satisfying $r_i$ for all $i \in \mathcal{A}$. Note that in this case backing off on transmit power will not change $\pi_s(n)$ since allocating more bandwidth does not imply real-time users would leave the system earlier. We refer to this as a *decoupling property.*[4] Thus, the problem reduces to one of optimizing power consumption for a static user population.

Now, we consider the convex optimization associated with minimizing power for a static user population. In every time frame $t$, we solve

$$\begin{aligned} \min_{\mathbf{x} \succ 0} \quad & \sum_{i \in \mathcal{A}} \frac{r_i}{x_i} \left[ \frac{\exp(\frac{x_i}{w}) - 1}{\gamma_i} + p_{\text{dc},i} + \sum_{j \in \mathcal{A} \setminus \{i\}} p_{\text{idle},j} \right] \\ & + \left(1 - \sum_{i \in \mathcal{A}} \frac{r_i}{x_i}\right) \sum_{i \in \mathcal{A}} p_{\text{idle},i} \\ s.t. \quad & \sum_{i \in \mathcal{A}} \frac{r_i}{x_i} \leq 1, \end{aligned} \tag{3.7}$$

[4] In Section 3.4 we will see that decoupling property does not hold for system dynamics of file transfers.

where $\mathbf{x}$ is a vector whose elements are $x_i$ and $\gamma_i = \frac{\eta g_i}{\sigma^2}$, $i \in \mathcal{A}$. We solve the optimization problem when the system is underutilized, i.e., when $r_i$ for all $i \in \mathcal{A}$ are achievable. We put the subscript $i$ for $p_{\text{dc}}$, $p_{\text{idle}}$, $g$ and $\gamma$ to accommodate the heterogeneous users.

Note that $\gamma_i$ and $\mathcal{A}$ may vary over different frames $t$ yet for simplicity we drop the time dependence. The optimization needs to be redone when $\gamma_i$ or $\mathcal{A}$ changes. As we will see in the sequel, the superlinear convergence speed and reuse of the previously determined optimal values make this optimization quickly computable on the fly.

The interpretation of the above optimization is as follows. When user $i$ transmits, the system power consumption is $\left(\exp(x_i/w) - 1\right)/\gamma_i + p_{\text{dc},i} + \sum_{j\in\mathcal{A}\setminus\{i\}} p_{\text{idle},j}$. This is weighted by $\frac{r_i}{x_i}$, the fraction of time user $i$ transmits. The sum over all users gives the average system power consumption. In addition, for a fraction of time $1 - \sum_{i\in\mathcal{A}} \frac{r_i}{x_i}$, all users consume idling power $\sum_{i\in\mathcal{A}} p_{\text{idle},i}$.

By manipulating the above we have an equivalent but simpler optimization problem given by,

*Problem* $\mathbb{O}1$:

$$\min_{\mathbf{x}\succ 0} \quad \sum_{i\in\mathcal{A}} \frac{r_i}{x_i}\left(\frac{\exp(\frac{x_i}{w}) - 1}{\gamma_i} + \xi_i\right) \tag{3.8}$$

$$s.t. \quad \sum_{i\in\mathcal{A}} \frac{r_i}{x_i} \leq 1, \tag{3.9}$$

where $\xi_i := p_{\text{dc},i} - p_{\text{idle},i}$. Note that Problem $\mathbb{O}1$ is a convex optimization with an inequality constraint because the objective function is a weighted sum of convex functions of $x_i$. Because the circuit power is higher than the idling power in practice, we assume $\xi_i \geq 0$. We note that it can be shown through a change of variables, $y_i = \frac{r_i}{x_i}$ and setting $p_{\text{idle},i} = p_{\text{dc},i}$, the original optimization problem also represents the optimal spectral bandwidth allocation

problem in a frequency-flat fading FDMA system where $y_i$ is the fractional bandwidth for user $i$ having the required rate $r_i$.

### 3.3.2 Solution: An Energy Optimal Transmission Policy

We propose an energy optimal transmission strategy for real-time sessions based on an iterative solution to Problem $\mathbb{O}1$. Given $\gamma_i$, $\xi_i$ and $r_i$, the base station solves the convex optimization problem using Lagrangian method. The optimal Lagrange multiplier is then computed by Newton's method, which guarantees superlinear convergence (faster than exponential). The base station then broadcasts the optimal Lagrange multiplier to mobile terminals, which, in turn, independently determine an associated transmission rate/power level. This makes for a scalable implementation.

Let $\kappa$ denote the Lagrange multiplier associated with the constraint in Problem $\mathbb{O}1$. The Lagrangian function is then given by

$$L(\mathbf{x}, \kappa) = \sum_{i \in \mathcal{A}} \frac{r_i}{x_i} \left( \frac{\exp(\frac{x_i}{w}) - 1}{\gamma_i} + \xi_i \right) + \kappa \left( \sum_{i \in \mathcal{A}} \frac{r_i}{x_i} - 1 \right).$$

This is a convex optimization so the necessary and sufficient conditions for optimality are given by Karush-Kuhn-Tucker (KKT) conditions [24], i.e., for all $i \in A$

$$\frac{\partial L}{\partial x_i^*} = 0 \text{ and } \kappa^* \left( \sum_{i \in \mathcal{A}} \frac{r_i}{x_i^*} - 1 \right) = 0 \tag{3.10}$$

where $\kappa^*$ denotes the optimal multiplier and $x_i^*$ is the optimal $x_i$. From $\frac{\partial L}{\partial x_i^*} = 0$, we have that

$$\kappa^* = \frac{1}{\gamma_i} \left( \exp\left(\frac{x_i^*}{w}\right) \left(\frac{x_i^*}{w} - 1\right) + 1 \right) - \xi_i, \, \forall i \in \mathcal{A}. \tag{3.11}$$

Suppose that $\kappa^*$ is known; the algorithm to compute $\kappa^*$ will be provided in Appendix I. Then the base station broadcasts $\kappa^*$, and mobile terminals solve (3.11). Unlike

the previous work which approximated the solution assuming high transmission rate [120] or used interior point method [33], we directly use the Lambert W function and obtain a closed form solution. Lambert W function also contributes to computing $\kappa^*$ in an efficient way combined with Newton's method, see Appendix I. Recall $W(z)$ is defined as [31]

$$W(z)e^{W(z)} = z, \tag{3.12}$$

and a concave, monotone increasing and differentiable function. We assume that mobile terminals have tabulated or can compute $W(z)$. The solution to (3.11) is then given by

$$x_i^* = \left( W\left( \frac{(\kappa^* + \xi_i)\gamma_i - 1}{e} \right) + 1 \right) w, \ i \in \mathcal{A} \tag{3.13}$$

and, the optimal output power level for $i \in \mathcal{A}$ is given by

$$p_i^* = \left( \exp\left( W\left( \frac{(\kappa^* + \xi_i)\gamma_i - 1}{e} \right) + 1 \right) - 1 \right) \frac{\sigma^2}{g_i}. \tag{3.14}$$

Let us consider two simple examples capturing the character of such uplink power control.

**Example 4** (Homogeneous Case 1)**.** *Suppose $\gamma_i = \gamma$, and $\xi_i = 0$, then we have that $x_i^* = \sum_{j \in \mathcal{A}} r_j$ for all $i \in \mathcal{A}$, i.e., the sum of all required rates. This yields the same power allocation across all users irrespective of their individual rate requirements, but a time allocation to each user is proportional to $r_i$.*

**Example 5** (Homogeneous Case 2)**.** *Suppose still that $\gamma_i = \gamma$, but now that $\xi_i = \xi > 0$. In this case (3.13) implies that $x_i^* = x^*$ for all $i \in \mathcal{A}$, but $x^*$ may be greater than $\sum_{i \in \mathcal{A}} r_i$. This will occur when the circuit power is large, so transmitting quickly and then idling is more beneficial than fully utilizing the time resource.*

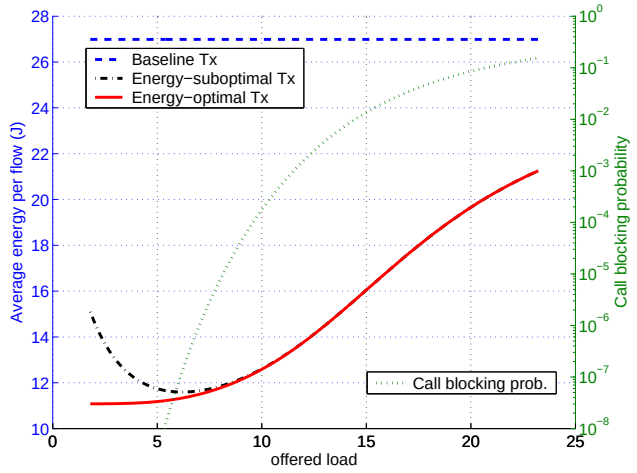



Figure 3.3: Energy saving for real-time sessions under various loads. $r_i = 150$ kbps for all users, $n^* = 23$, $c_{\max} = 3.49$ Mbps, $\mu_s^{-1} = 180$ sec, received SNR with full power transmit = 15 dB, other parameters are shown in Table 3.3.

### 3.3.3 Energy-savings under various loads

So far, we considered the optimization for a *fixed* number of users. Recall that our objective is to minimize the per-flow energy in a dynamic system, and it is of interest to see how energy saving benefits scale under various loads. To demonstrate this, we consider, for simplicity, homogeneous users with identical $\gamma$ and rate requirement $r$, so user index $i$ is dropped. We compare three transmission policies. The baseline policy is such that each terminal transmits at the maximum rate, i.e., the instantaneous transmission rate is $x = c_{\max}$. The second policy simply scales with the number of users, so $x = nr$, which fully utilizes the time resource. The third policy is our energy optimized one where $x^*$ is given by (3.13). Let $p(n)$ denote the system power consumption in state $n$; it is given by

$$p(n) = \frac{nr}{x}\left(f(x) + (n-1)p_{\text{idle}}\right) + \left(1 - \frac{nr}{x}\right) np_{\text{idle}}. \tag{3.15}$$

Then, the average system power consumption is $E[P] = \sum_{n=1}^{n^*} p(n)\pi_s(n)$ where $\pi_s(n)$ is given in (3.6). From Theorem 3 and considering the call blocking probability $\pi_s(n^*)$, the

average per-flow energy is given by

$$E[J] = \frac{E[P]}{\lambda_s(1 - \pi_s(n^*))}. \tag{3.16}$$

Representative results for the three policies are shown in Fig. 3.3. As can be seen, the optimal policy (solid line) significantly saves energy with respect to the baseline (dashed line). Per-flow energy is reduced by more than 50% when the call blocking probability is 0.1% or less. The energy saving benefits become more significant when the loads are low. Recall that energy savings come at no cost in terms of compromising user perceived performance.

*Remark* 3.3.1. The second policy $x = nr$ (dash-dot line) exhibits an interesting behavior in Fig. 3.3; this policy is asymptotically optimal as the loads grow, however, far from optimal when the loads are low. This is because of the impact of circuit power. When the loads are low, and $n$ is usually small, the circuit energy may dominate the transmit energy. Thus, transmitting faster than the required rate (i.e., $x^* > nr$) saves energy. Recall that Example 5 demonstrated this effect in a static system; here we see the analogous effect for the *dynamic* system.

### 3.3.4 Spatial power smoothing and fair energy savings

A further gain of our energy-optimal transmission policy is that *both* the output power levels *and* total power consumptions of mobile terminals are spatially smoothed. Let us consider an example. A base station is placed at $(0, 0)$ and 100 mobile terminals are placed every 30 m on a 10 by 10 square grid. We consider both of large and small scale fading; specifically path loss with exponent 3 and i.i.d. Rayleigh fading channels. Fig. 3.4 (a) exhibits the output power levels when all terminals are allocated an equal fraction of time. As can be seen, the output powers generally increase with the distance from the

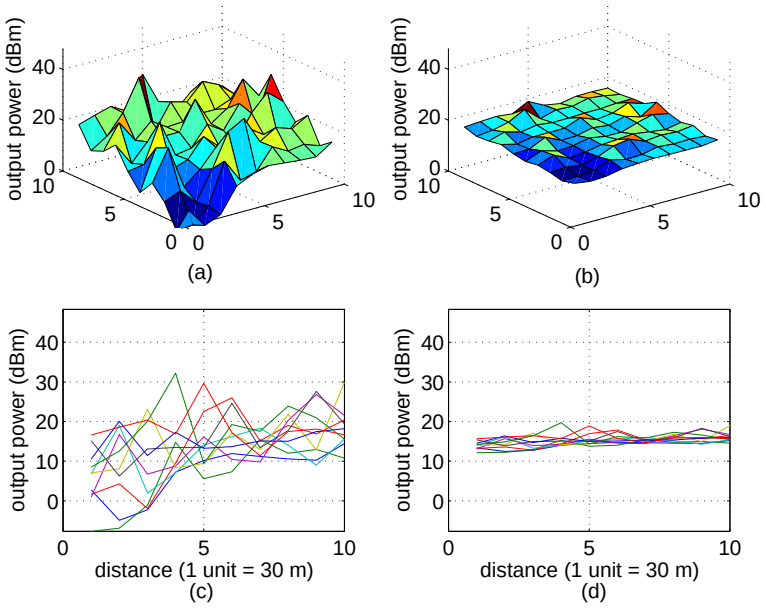



Figure 3.4: Spatial power smoothing, (a) Equal time fraction allocation (b) Optimized rate and time fraction (c) Side view of (a), (d) Side view of (b): $r_i = 50$ kbps, path loss exponent = 3, cell radius = 300 m, 100 users, carrier frequency = 1 GHz, other parameters are shown in Table 3.3.

base station. Fig. 3.4 (b) exhibits the output powers after applying our energy optimal transmission policy; the power levels are significantly smoothed and almost same across the cell. Fig. 3.4 (c) and (d) are the side views of (a) and (b), which reveal that the deviation of output powers are reduced significantly, i.e., from 40 dB to 5 dB. Furthermore, at the cell edge, the optimization reduces the output power levels by up to 15 dB. Even though we do not consider inter-cell interference here, reduced output power at the cell boundary suggests that our energy-saving mechanism could contribute to reducing inter-cell interference in multiple cell scenario.

## 3.4 Energy Savings for File Transfers

In this section, we consider energy savings in the context of uplink file transfers. Our focus is again on flow-level dynamics, and understanding how energy-savings can exploit times when the system is underloaded. A practical algorithm is proposed to achieve energy-efficiency and target throughput. The approach is then combined with opportunistic scheduling to exploit time-varying channels.

There are three key differences between achieving energy savings in system supporting real-time sessions versus file transfers. First, real-time sessions have strict rate requirements that must be achieved, otherwise, the sessions may be dropped. By contrast, file transfers are delay-tolerant, and users can specify a target throughput considering their preferences between energy savings and fast transmission. For example, a user with sufficient residual battery may prefer fast transmission, but another user with scarce battery may prefer slow transmission to benefit from the *energy-delay tradeoff*. Second, in the case of real-time sessions, the stationary distribution of the number of users is independent of the power control policy; we called this the decoupling property. In the case of file transfers, however, power control changes the stationary distribution, which makes the problem more challenging. Third, in determining energy-efficient transmission, circuit power was important for real-time sessions, but, as we will see, idling power plays a more crucial role in the case of file transfers.

### 3.4.1 Energy savings in an underutilized system

Recall our claim that energy can be saved without substantially impacting user perceived performance in an underutilized system. For purposes of developing some insight, consider two simple examples from the perspective of different time scales.

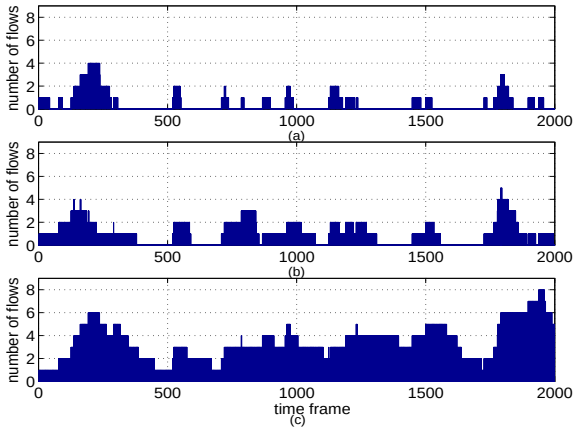



Figure 3.5: Time varying number of users in a dynamic system with offered load 30%. Individual target throughput is (a) 5.10 Mbps (b) 1,275 kbps (c) 318 kbps, and the arrival processes are identical. Simulation setup is given in Section 3.4.8.

**Example 6** (Long-term time scale)**.** *If an M/M/1 processor sharing system is stationary, the average file delay is given by* $d = \frac{1}{\mu c - \lambda}$ *where* $\mu^{-1}$ *is the average file size,* $\lambda$ *is the file arrival rate, and* $c$ *is the system capacity (or equivalently, system throughput.*[5]*) So, the system capacity to achieve an average delay* $d$ *is given by* $c = \frac{\lambda}{\mu} + \frac{1}{\mu d}$*. Suppose that the arrival rate over a long time scale is reduced to* $\lambda'$*. Then,* $c$ *could in principle be adapted to this and reduced to* $\frac{\lambda'}{\mu} + \frac{1}{\mu d}$ *and energy can be saved without impacting average file delay.*

**Example 7** (Short-term time scale)**.** *Fig. 3.5 exhibits* $n(t) = |\mathcal{A}(t)|$ *when the mean offered load is 30%. Unlike the previous case, let us consider short term dynamics. As can be seen in Fig. 3.5 (a), the base station frequently experiences periods when the system is idle, i.e., no users, corresponding to periods when the resources are essentially unused. These periods can be leveraged to save energy, by having users can backoff on their transmit power and rate as long as the resulting performance is acceptable. As shown in Figs. 3.5 (b) and (c) when such a strategy is used the system utilization increases, yet energy may be conserved.*

[5]System capacity in this chapter is not the same notion as the information theoretic capacity.

One might think that a backoff on transmit power decreases the SNR which may in turn increase the link error rate. This could lead to packet losses, which would be bad for our real-time sessions and bad for file transfers because TCP performance deteriorates over lossy wireless channels. However, adaptive modulation and coding (AMC) are used in real systems, and the transmission rate can be reduced accordingly as the SNR decreases while preserving the target bit error rate (BER) [4, 48]. In addition, Hybrid Automatic Repeat reQuest (HARQ) is used at the physical layer in cellular systems such as WiMAX, 3GPP-LTE, HSDPA, HSUPA, etc. to hide local link errors from TCP senders [4, 5]. As a consequence TCP performance is not likely to be degraded by using the energy saving techniques we present hereafter.

### 3.4.2 Problem formulation

Let us go back to the system model shown in Fig. 3.1 to formulate the problem in a dynamic system. Our objective is to minimize $E[J]$, see (3.2), while achieving a target throughput per user denoted by $q_i$; $q_i$ can be thought of as a tuning parameter controlling the tradeoff between fast transmission and energy savings.

In minimizing $E[J]$ in a stationary system, the two key elements are the system capacity, and how it is shared among ongoing flows. The system capacity not only determines the departure rate of flows but also controls the energy consumption of mobile terminals. We describe three models for the system capacity as a function of $n$, denoted by $c(n)$. We assume for simplicity that users have the same target throughput and experience homogeneous channels, so the user index $i$ is dropped.

***Baseline policy:*** Suppose all users are scheduled for an equal fraction of time and transmit at the full power to achieve the maximum achievable throughput. In this case the

system capacity is *not* state dependent, and given by

$$c(n) = c_{\max}, \tag{3.17}$$

where $c_{\max}$ is the maximum uplink capacity achievable by any individual user, and the scheduling discipline can be modeled as a processor sharing queue. Among the "fair" policies we consider, this one minimizes the file transfer delay, but expends the most power.

***State-dependent policy:*** Alternatively, consider a state-dependent transmission policy where the system capacity is given by

$$c(n) = \min(nq, c_{\max}). \tag{3.18}$$

The intuition underlying (3.18) is as follows. Assuming once again a processor sharing scheduling discipline, as long as the system is not overloaded, capacity is allocated so that each user sees its target throughput $q$, but no more than that. Thus the system capacity grows linearly in $n$, i.e., $c(n) = nq$ until it reaches the maximum system capacity $c_{\max}$. This policy represents a simple model for exploiting *dynamic spare capacity* to conserve energy by allowing the transmit power (and also the rate) to backoff.

***Opportunistic policy:*** If channels are time-varying, we may use opportunistic scheduling. In the simplest case where users are homogeneous, the system capacity using max-rate scheduling [63] would

$$c(n) = E[\max(R_1, \cdots, R_n)] \tag{3.19}$$

where $R_i$, $i \in \mathcal{A}$ is a random variable denoting the channel capacity of user $i$. Note that under max-rate scheduling for a homogeneous system each user would be served an equal fraction of time, thus processor sharing is again roughly a good approximation for how users are scheduled.

### 3.4.3 Flow-level dynamics

Given the above three simple models for system capacity we now obtain a Markov chain model for the number of ongoing flows in the system. We assume that the arrivals of file transfer requests follow an independent Poisson process with arrival rate $\lambda$ and have independent file sizes with mean $\mu^{-1}$. Note that we do not assume the file sizes are exponentially distributed. Let $\mathbf{N} = (N(s), s \geq 0)$ denote a random process representing the number of ongoing file transfers at time $s$. Then, if file sizes are exponentially distributed, $\mathbf{N}$ is a Markov process with state space $\mathbb{Z}^+$ and rate matrix $Q$ is given by

$$\begin{aligned} q(n, n+1) &= \lambda \\ q(n+1, n) &= \mu c(n+1) \quad \text{for } n \geq 0. \end{aligned}$$

The stationary distribution $\pi$, if it exists, is given by

$$\pi(n) = \pi(0) \frac{\rho^n}{\Pi_{m=1}^n c(m)}, \tag{3.20}$$

where $\rho := \frac{\lambda}{\mu}$ is the traffic load (bits per second) and $\pi(0) = \left(1 + \sum_{n=1}^{\infty} \frac{\rho^n}{\Pi_{m=1}^n c(m)}\right)^{-1}$. Note that the insensitivity property for Processor sharing queue ensures this distribution also holds for general file size distributions [18]. In the sequel we let $N$ be a random variable with distribution $\pi$. In steady state, the average system power consumption is given by $E[P] = \sum_{n=0}^{\infty} p(n)\pi(n)$ where $p(n)$ is a function which captures the overall system power expenditure in state $n$ and given by

$$p(n) = f(c(n)) + (n-1)p_{\text{idle}}, \tag{3.21}$$

because, at any time instance, one user is transmitting at the instantaneous rate $c(n)$ and $n-1$ users are idling. Finally, from Theorem 3, the average energy per flow is given by

$$E[J] = \frac{1}{\lambda} \sum_{n=1}^{\infty} \Big( f(c(n)) + (n-1)p_{\text{idle}} \Big) \pi(n). \tag{3.22}$$

Note that $\pi(n)$ depends on the system capacity $c(n)$, i.e., these are *coupled* together, see (3.20). Hence, the subtlety here is that, by backing off on transmit power one likely increases the number of flows in the system making the overall optimization of the dynamic system more challenging.

### 3.4.4 Energy-delay tradeoffs: Numerical results

Next, we investigate how changing the tuning parameter $q$ in (3.18) impacts the energy and delay performance; specifically, by reducing $q$ from $c_{\max}$, different performance pairs for delay and energy are obtained; these are shown in Fig. 3.6. When $q = c_{\max}$, the state-dependent policy is identical to the baseline; the delay is the smallest but the energy consumption is the highest. This baseline is exhibited by $\circ$ in Fig. 3.6. Then, as $q$ is reduced, energy is saved but average delay increases. We consider three power models, differing in whether they include the effect of circuit and/or idling power. As can be seen, Power Model 1 comprises both circuit and idling power and significant amount of energy, e.g., up-to 60% relative to the baseline, can be saved as $q$ is reduced (solid line). Interestingly, however, if $q$ is excessively reduced, the energy consumption grows again. This is because further reducing $q$ results in an increased number of idling users expending excessive idling energy. Thus there exists an *energy-optimal* target throughput where the most benefit is achieved.

**Example 8** (Energy-power equivalence)**.** *To better understand the relationship between energy and power in a dynamic system, here we provide an example based on Theorem 3,* $E[P] = \lambda E[J]$. *Fig. 3.6 shows the average energy consumption of a typical file transfer, which is measured during "file transfer time." The arrival rate* $\lambda$ *of file transfer requests 3.65/sec with average file size 60 kbytes gives traffic load of 1,752 Mbps* $(= 3.65 \times 60 \times 8)$. *Since* $c_{\max} = 5.84$ *Mbps, 1.752 Mbps divided by 5.84 Mbps gives 30% of offered load. Then,*

*here is a small calculation to provide intuition. The baseline with 240 mJ per file corresponds to the system power consumption 876 mW (= 3.65/sec × 240 mJ). The energy optimal point in Fig. 3.6 consumes 90 mJ per file, which corresponds to system power consumption 328 mW (=3.65/sec × 90 mJ). Then the reduction of system power consumption is 876-328 = 547.5 mW, which can be also computed by 3.65/sec × (240 - 90) mJ. So, we can compute the reduction of system power consumption by multiplying the reduction of energy per flow by the arrival rate. In this example, the total energy saving per file (or equivalently, the total system power saving) including circuit and idle power is 62%, but the file transfer delay is increased, which is called "energy-delay tradeoff".*

Before investigating energy optimal throughput, we first provide a lemma emphasizing the weak impact of circuit power on the energy consumption.

**Lemma 1** (**Bounded circuit energy**). *If a dynamic system is stationary, the impact of circuit energy per flow is monotonically increasing as the delay grows, but bounded by* $\frac{p_{dc}}{\lambda}$.

*Proof.* The average circuit power consumption in the system is $\sum_{n=1}^{\infty} p_{\mathrm{dc}}\pi(n) = p_{\mathrm{dc}}\left(1-\pi(0)\right)$. From Theorem 3, the average circuit energy per flow denoted by $\phi_c$ is given by $\phi_c = \frac{p_{\mathrm{dc}}(1-\pi(0))}{\lambda} \leq \frac{p_{\mathrm{dc}}}{\lambda}$. Since $\pi(0)$ is decreasing in delay, $\phi_c$ is monotonically increasing as delay grows, but bounded by $\frac{p_{\mathrm{dc}}}{\lambda}$. □

**Theorem 4** (**Asymptotically negligible circuit energy**). *If a dynamic system is stationary, the impact of circuit energy per flow becomes asymptotically negligible as the load grows.*

*Proof.* From Lemma 1, the bound $\frac{p_{\mathrm{dc}}}{\lambda}$ is decreasing as $\lambda$ grows, and thus the circuit energy becomes asymptotically negligible as the load grows. □

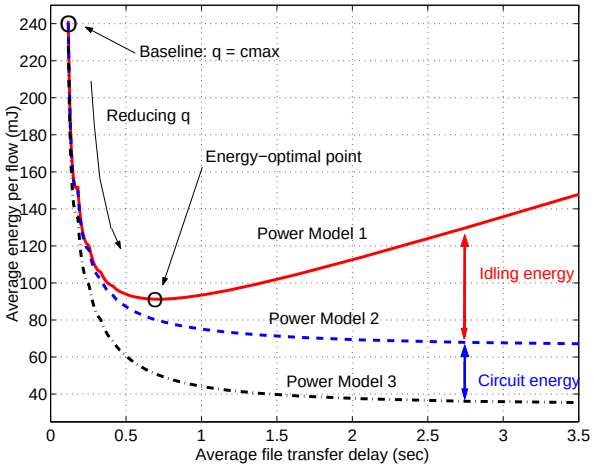



Figure 3.6: Energy-delay tradeoff for various throughput $q$. ($\lambda = 3.65$/sec, $c_{\max} = 5.84$ Mbps, offered load = 30%, received SNR with maximum rate transmission = 17.5 dB. Model 1: $p_{\rm dc} = 115.9$ mW, $p_{\rm idle} = 25$ mW, Model 2: $p_{\rm dc} = 115.9$ mW, and $p_{\rm idle} = 0$ mW, Model 3: $p_{\rm dc} = p_{\rm idle} = 0$ mW. Other parameters are given in Table 3.3.)

Although Lemma 1 and Theorem 4 are simple, they demonstrate a key difference between static and dynamic systems. Here are two supporting examples.

**Example 9.** *To focus on the circuit energy effect, we set the idling power as zero in this example. We compare Power Model 2 (with transmit and circuit power) with Model 3 (with transmit power only). Fig. 3.6 shows that Model 2 consumes more energy than Model 3 by the amount of circuit energy. As can be seen, the energy gap between Model 2 and 3 is monotonically increasing as the delay grows, but quickly saturates to* $\frac{p_{dc}}{\lambda}$*. As a result, the energy decreases monotonically in delay.*

This result is surprising because it is the *opposite* of what happens in static systems, i.e., long delay ultimately increased the energy consumption and thus there existed an energy-optimal throughput (or delay), see [34, 76, 93, 123].

**Example 10.** *To have an insight on diminishing impact of circuit energy, we plot the energy consumption for Model 2 for various offered loads. In Fig. 3.7, we exhibit the energy and*

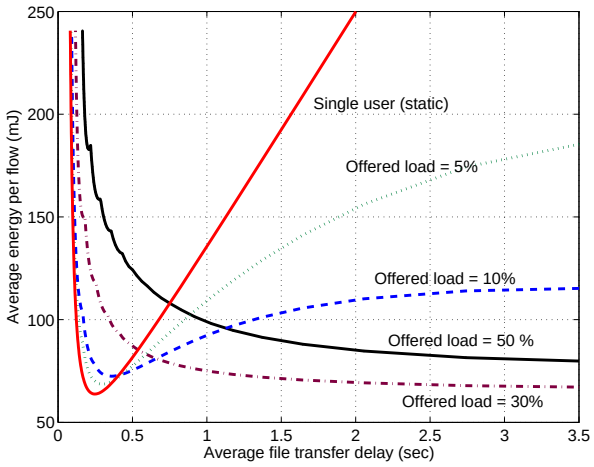



Figure 3.7: The weak impact of circuit power in energy-delay tradeoff: $p_{\rm dc} = 115.9$ mW, $p_{\rm idle} = 0$ mW, $c_{\max} = 5.84$ Mbps, received SNR with maximum rate transmission = 17.5 dB. Other parameters are given in Table 3.3.

*delay in the case of single user; the energy increases linearly when the delay is large (and the slope becomes identical to circuit power $p_{dc}$). However, for stationary systems, as the offered loads grow (5% to 50%), the impact of circuit energy is gradually diminishing, and finally, we see the monotonically decreasing energy consumption in delay. This confirms that for dynamic systems circuit energy is asymptotically negligible as the load grows.*

### 3.4.5 Stationary analysis

To enable a more quantitative analysis, we consider a regime where $c_{\max} \gg q$, i.e., the maximum system capacity far exceeds individual users' target throughput, and the system load is light. This captures the system dynamics as $q$ goes to zero (or delay goes to $\infty$). Then (3.22) can be simplified using the approximation $c(n) \approx nq$. The queue's stationary distribution $\pi(n)$ in (3.20) is then roughly Poisson with parameter $\frac{\lambda}{\mu q}$. Let $\phi(\lambda, q)$ denote the energy per flow at $(\lambda, q)$, i.e.,

$$\phi(\lambda, q) := \sum_{n=1}^{\infty} \Big( \frac{\exp(\frac{nq}{w}) - 1}{\gamma} + p_{\rm dc} + (n-1)p_{\rm idle} \Big) \frac{e^{-\frac{\lambda}{\mu q}} (\frac{\lambda}{\mu q})^n}{\lambda n!}.$$

Recognizing the first term as the moment generating function of a Poisson random variable, one obtains

$$\begin{aligned}\phi(\lambda,q) &= \frac{\exp\left(\frac{\lambda}{\mu q}(\exp(\frac{q}{w})-1)\right)-1}{\lambda\gamma}+ \\ &\quad p_{\mathrm{dc}}\frac{1-\exp(-\frac{\lambda}{\mu q})}{\lambda}+p_{\mathrm{idle}}\left(\frac{1}{\mu q}-\frac{1-\exp(-\frac{\lambda}{\mu q})}{\lambda}\right).\end{aligned} \tag{3.23}$$

Note that, as $\lambda \to 0$, (3.23) also captures the energy expenditure for a single user which sees no other flows than itself:

$$\lim_{\lambda\to 0}\phi(\lambda,q) = \frac{1}{\mu q}\left(\frac{\exp(\frac{q}{w})-1}{\gamma}+p_{\mathrm{dc}}\right). \tag{3.24}$$

The first term in (3.23) accounts for transmit energy, which increases exponentially in $\lambda$ given a fixed $q$. This implies if $\lambda$ is reduced (i.e., the system load is reduced), significant energy can be saved while maintaining the same $q$. The second term in (3.23) accounts for circuit energy. As mentioned in Lemma 1 and Theorem 4, as $q$ goes to zero, the circuit energy goes to $\frac{p_{\mathrm{dc}}}{\lambda}$. Furthermore as the load grows, it becomes asymptotically negligible.

The third term in (3.23) accounts for idling energy that plays a crucial role in determining the energy-efficiency. As can be seen, as $q$ is decreasing, the idling energy is increasing while the transmit energy (the first term) is decreasing. Hence, $\phi(\lambda,q)$ has an *energy-optimal* throughput for a given $\lambda$, which we denote by

$$e := \operatorname*{argmin}_{q>0}\phi(\lambda,q). \tag{3.25}$$

One can attempt to determine $e$ by solving $\frac{\partial}{\partial q}\phi(\lambda,q)=0$, yet this equation does not have a closed form solution. Instead, to get a sense of its characteristics, we will use a linear approximation around $q=0$, i.e., $\phi(\lambda,q)\approx s_1 q+s_2+\frac{p_{\mathrm{idle}}}{\mu q}$, where $s_1$ and $s_2$ are Taylor series

coefficients of $\phi(\lambda, q)$. Simple calculus gives the following approximation for the energy-optimal per-flow throughput:

$$e \approx w\sqrt{2\exp\left(-\frac{\rho}{w}\right)p_{\text{idle}}\gamma}. \tag{3.26}$$

*Remark* 3.4.1 (**Throughput region**). Eq. (3.26) suggests the *throughput region* $\{q|q \geq e\}$ where the throughput can be traded off with energy. Otherwise, both of the average delay and the energy performance are bad.

Interestingly, $e$ is an increasing function of SNR $\gamma$; so transmitting faster when channels are good indeed saves energy. In addition, fast transmissions are beneficial when idling power $p_{\text{idle}}$ is high; otherwise accumulated users will consume too much idling energy.

### 3.4.6 CUTE algorithm

Although we derived the energy-optimal throughput for a stationary system, it is not straightforward to apply this result in real system. Users experience heterogeneous and time-varying channels, the number of users will change, and the system may not be stationary; even if quasi stationary, it may not be easy to correctly estimate $\rho$ in (3.26). In this section we propose a simple practical algorithm that does not use the prior knowledge of the traffic load but simply relies on the current system state $n(t)$.

***Energy-efficient rate:*** The key idea is to replace the energy-optimal throughput (3.26), obtained in a stationary regime, with state-dependent one associated with each time frame $t$. Consider an uplink which is equally time shared by $n(t)$ users. The average energy per bit for user $i \in \mathcal{A}(t)$ to achieve throughput $x$ during one time frame is given by $\left(\frac{1}{n(t)}f_i(n(t)x) + \frac{n(t)-1}{n(t)}p_{\text{idle},i}\right)/x$ where $f_i(\cdot)$ is a user-indexed version of (3.4). Note that each user uses only a fraction $\frac{1}{n(t)}$ of time frame and so the instantaneous rate must be

$n(t)x$. The most energy-efficient individual throughput $e_i(t)$ can be determined based on

$$e_i(t) \quad := \quad \underset{x \geq 0}{\operatorname{argmin}} \left\{ \frac{f_i(n(t)x) + (n(t)-1)p_{\text{idle},i}}{x} \right\} \tag{3.27}$$

i.e., the throughput that minimizes the average energy per bit. Since (3.27) is differentiable and convex, $e_i(t)$ is given by simple calculus such as

$$e_i(t) = \frac{w}{n(t)} \left[ W\left( \frac{p_{\text{dc},i} + (n(t)-1)p_{\text{idle},i}}{e} \gamma_i(t) - \frac{1}{e} \right) + 1 \right]. \tag{3.28}$$

Using (3.28), each mobile can determine its own energy efficient rate $e_i(t)$ given $n(t)$.

*Remark* 3.4.2 (**Energy-opportunistic transmission**). Note that $e_i(t)$ is *energy-opportunistic* in the sense that $e_i(t)$ is an increasing function of $\gamma_i(t)$; if the channel is good, increasing the transmission rate saves energy (and vice versa). This is similar to the time-domain water filling, which is known to be the optimal transmission policy over a time-varying channel [48].

***Constraints:*** Two additional constraints play a role. First the maximum instantaneous transmission rate of a user is in practice bounded, say by $c_i(t)$. Thus when there are $n(t)$ users sharing the system, the highest achievable user throughput is $c_i(t)/n(t)$. Second users can specify their own target throughput $q_i$ considering their residual batteries and fast transmission. Thus, energy-efficient rate is upper and lower bounded, and the throughput for user $i$ is given by

$$r_i(t) \quad = \quad \min \left[ \max \left[ e_i(t), q_i \right], \frac{c_i(t)}{n(t)} \right], \; i \in \mathcal{A}(t). \tag{3.29}$$

***Relaxing target throughput:*** Since file transfers are delay tolerant, we do not need to achieve $q_i$ instantaneously. Instead, we might consider achieving it over a reasonable

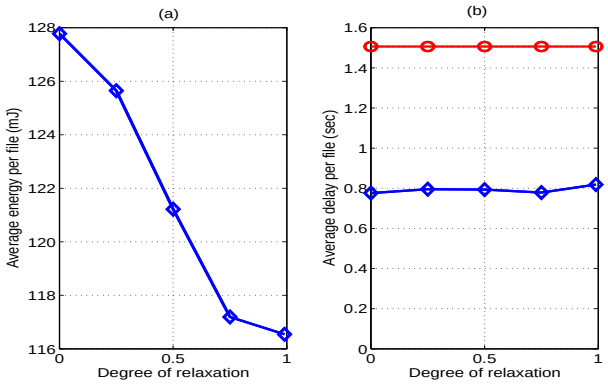



Figure 3.8: Additional energy saving by relaxed target rate in Rayleigh fading channels. $q_i$ = 320 kbps (1.5 second delay for 60 kbyte file), $c_{\max}$ = 5.1 Mbps, 30 % offered load. (a) average energy per file, (b) target delay ($\circ$), and the achieved delay ($\diamond$). Parameters are same to the simulations in Section 3.4.8

averaging window. We define the exponentially averaged throughput $\bar{r}_i(t)$ as

$$\bar{r}_i(t) = \nu \bar{r}_i(t-1) + (1-\nu) r_i(t),\ i \in \mathcal{A}(t) \tag{3.30}$$

where $\nu \in (0,1)$ corresponds to weight on the past. To meet $q_i$ on average, we choose $q_i(t)$ such that

$$q_i = \nu \bar{r}_i(t-1) + (1-\nu) q_i(t),\ i \in \mathcal{A}(t)$$

which yields

$$q_i(t) \quad = \quad \frac{q_i - \nu \bar{r}_i(t-1)}{1-\nu},\ i \in \mathcal{A}(t). \tag{3.31}$$

This relaxes the time scale over which the performance target should be met and contributes to further energy savings. Fig. 3.8 exhibits how such averaging time scales save energy while keeping the average file delays almost the same (solid $\diamond$). In summary the proposed algorithm realizes the following throughput

$$r_i(t) \quad = \quad \min\left[\max\left[e_i(t), q_i(t)\right], \frac{c_i(t)}{n(t)}\right],\ i \in \mathcal{A}(t). \tag{3.32}$$

We refer to this transmission policy as CUTE meaning Conserve User Terminals' Energy. In a run time CUTE alternates among three transmission modes– energy-efficient mode at $e_i(t)$, target mode at $q_i(t)$ and capacity-constrained mode at $c_i(t)/n(t)$ – in accordance with the system state, throughput history and channel fluctuations so that CUTE achieves (or exceeds) a target throughput while saves energy.

*Remark* 3.4.3 (**Energy-efficient mode**). The energy-efficient mode is the most 'desirable'; indeed when $e_i(t) \geq q_i(t)$ and feasible, user $i$ is served *faster* than its target and saves energy as well. If the system is underutilized, or channels are good, users are more likely to operate in this mode because $e_i(t)$ can be high, see (3.28).

Otherwise, if $q_i(t) > e_i(t)$, the user defers energy-saving and is served at $q_i(t)$ in order to meet the target throughput. Users with low SNR tend to operate in the target mode. If the system is congested or SNR is bad, that user may be in the capacity-constrained mode.

**The motivation for the study of convergence of CUTE:** We developed the CUTE algorithm to be practically applicable for real, dynamic systems where channels are time-varying and the number of users changes. Note that CUTE depends on attending to meet a relaxed target rate $q_i(t)$ which also varies as the number of users and channel gains change. However, when the channels and the number of users vary relatively slowly as compared to the length of the adaptation time frame, it is worthwhile to establish that the CUTE algorithm converges, i.e., in a quasi-static regime would quickly converge to a new stationary transmission rate. Such convergence is particularly meaningful when $n(t)$ and also the channel change slowly as compared to the length of the time associated with transmission frames, and also, when the file sizes are large enough. The following results are shown in the Appendix II.

**Theorem 5** (**Convergence of CUTE**). *Suppose that the number of users and channel*

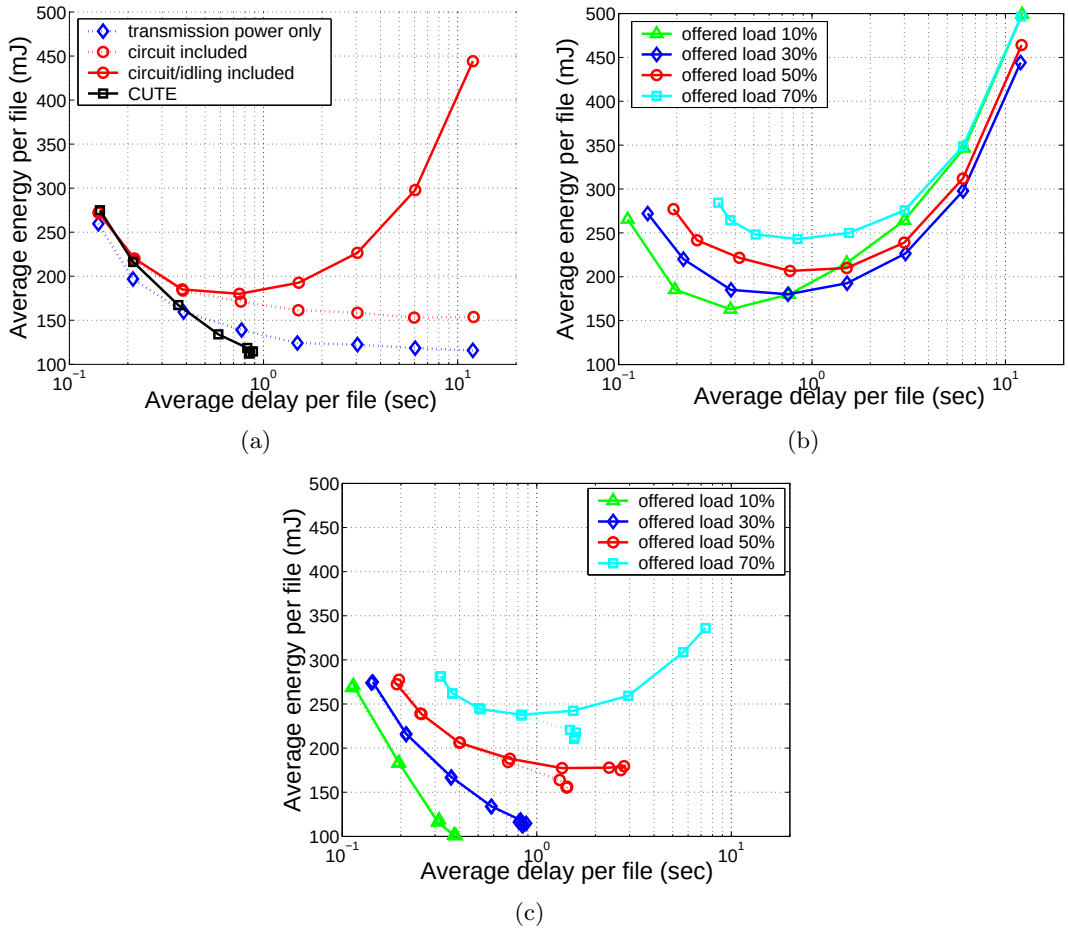



Figure 3.9: Energy-delay tradeoffs with round-robin scheduling. (a) CUTE algorithm to mitigate the impact of circuit/idling power on energy-delay tradeoff: $\lambda = 3.2$, offered load = 30 %. (b) Without energy-efficient rate. (c) With energy-efficient rate.

*gains are fixed, and consequently* $e_i(t) = e_i$ *and* $\frac{c_i(t)}{n(t)} = \frac{c_i}{n}$ *are fixed. Then, the average throughput* $\bar{r}_i(t)$ *and the transmission rate* $r_i(t)$ *both converge to* $\min(\max(q_i, e_i), \frac{c_i}{n})$*. Thus, if feasible, CUTE converges to the greater of* $q_i$ *and* $e_i$*, otherwise, to* $\frac{c_i}{n}$*.*

**Theorem 6 (Convergence speed).** *Both of* $\bar{r}_i(t)$ *and* $r_i(t)$ *converge to the equilibrium rate at least exponentially fast.*

### 3.4.7 CUTE with opportunistic scheduling

Opportunistic scheduling is desirable to enhance users' throughput when they see time-varying channels. Opportunistic scheduling for power control was first proposed in [68], but the authors exploited opportunism not to save energy but to enhance throughput. Clearly, opportunistic scheduling can serve both purposes. CUTE is compatible with various types of opportunistic scheduling such as [63, 71, 87, 89, 106]. The benefit of backing off the transmit power is more apparent when opportunistic scheduling is used versus round-robin scheduling because scheduled users are more likely to be experiencing high SNRs, and operating at energy-efficient mode, see Remark 3.4.3

To this end, we consider modifying our time sharing discipline. Consider the case where rather than serving all users in each frame, we schedule only a single user and assume the frame length is reduced to the channel coherence time. Let $s_\theta(t)$ denote the index of the scheduled user under an opportunistic policy $\theta$ on frame $t$. The proposed transmission policy under an opportunistic scheduling for user $i \in \mathcal{A}(t)$ is

$$r_i(t) = \min\left(\max\left(e_i(t), q_i(t)\right), c_i(t)\right) \mathbf{1}_{\{i=s_\theta(t)\}}, \tag{3.33}$$

where $\mathbf{1}_{\{\}}$ is the indicator function and $e_i(t)$ is redefined as

$$e_i(t) \quad = \quad \underset{x\geq 0}{\operatorname{argmin}} \left\{ \frac{f_i(x) + (n(t)-1)p_{\text{idle},i}}{x} \right\}$$

Note that we use $f_i(x)$ instead of $f_i(n(t)x)$ because only one user is scheduled per time frame. Also, note that $c_i(t)$ is used instead of $\frac{c_i(t)}{n(t)}$, and $q_i(t)$ is modified giving

$$q_i(t) \quad = \quad \frac{n(t)q_i - \bar{r}_i(t-1)\nu}{1-\nu}, \; i \in \mathcal{A}(t) \tag{3.34}$$

where $\bar{r}_i(t)$ is computed during the time frames where user $i$ has been served.

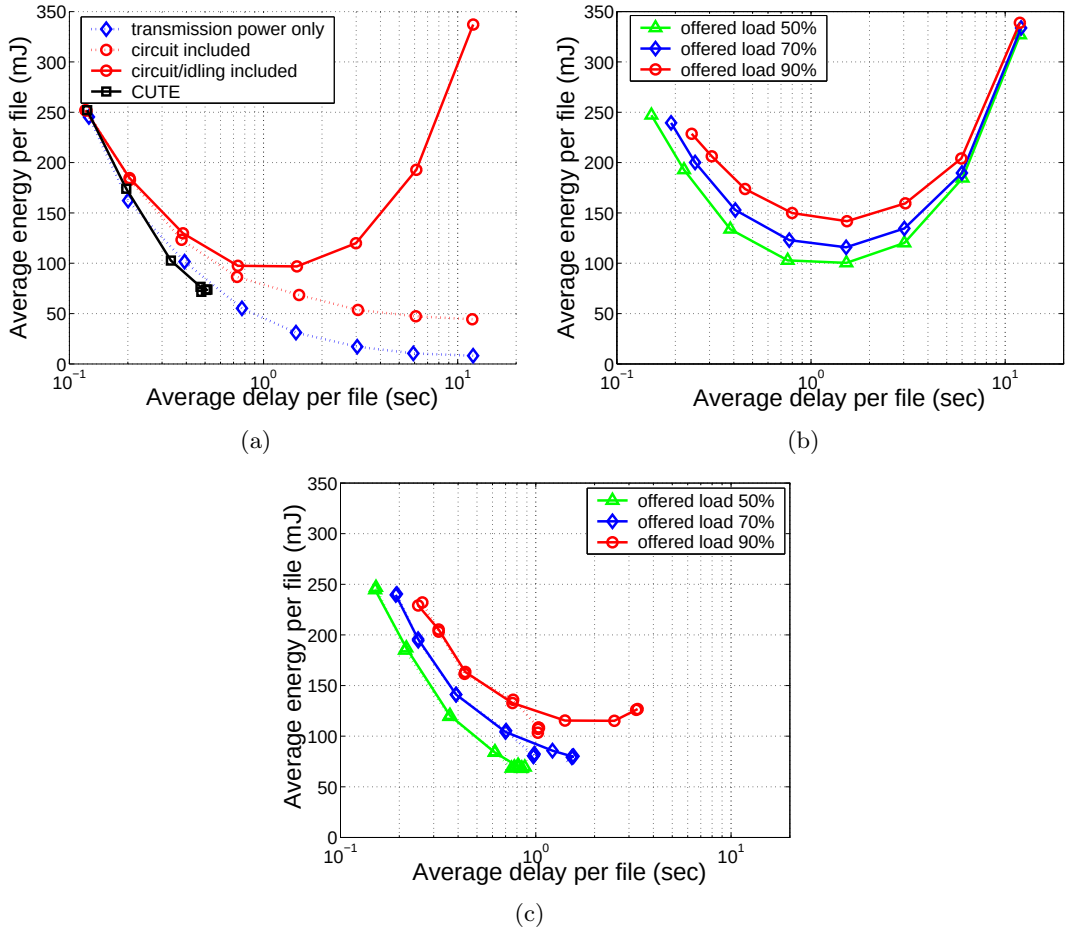



Figure 3.10: Energy-delay tradeoffs with opportunistic scheduling. (a) CUTE algorithm to mitigate the impact of circuit/idling power on energy-delay tradeoff: $\lambda = 3.2$, offered load = 30 %. (b) Without energy-efficient rate. (c) With energy-efficient rate.

### 3.4.8 Simulation results

To validate the effectiveness of the CUTE algorithm, we estimated the average energy consumption per file transfer versus the average delay using flow-level event-driven simulations. On each time frame, new user requests arrive according to a Poisson process with rate $\lambda$. Each user requests exactly one file that is log normally distributed with mean 60 kbytes [21]. Users are assumed to experience independent Rayleigh fading channels. Our

simulation parameters are $\nu = 0.95$, path loss $= -124$ dB, and an ergodic channel capacity is 5.1 Mbps. Other parameters are given in Table 3.3. The average received SNR at the base station when the mobile terminal transmits at its maximum output power is 17.5 dB. When mobile terminals reduce the target throughput, and power backoff is used, the average received SNR decreases. The number of time frames per simulation is 1,000,000. We plot the energy-delay tradeoff curves for $q_i = (1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}) \times 5.1$ Mbps to show how the user's preference on energy savings against fast transmission impacts the energy-delay tradeoff.

Fig. 3.9 demonstrates energy-delay tradeoffs under round-robin scheduling. Fig. 3.9 (a) exhibits four curves: transmit power only (dashed $\Diamond$), transmission and circuit power (dashed $\circ$), transmission, circuit and idling power (solid $\circ$), and CUTE algorithm (solid $\Box$). As expected, idling and circuit power increase the average energy. Furthermore, the impact of idling energy dominates when delay is large. This is because the accumulated users result in high idling energy consumption. By contrast, circuit energy becomes bounded by $\frac{p_{\mathrm{dc}}}{\lambda} = 36.2$ mW as stated in Lemma 1. Comparing solid $\circ$ line with solid $\Box$ line shows how the CUTE algorithm significantly improves the energy-delay performance in the presence of idling and circuit power. Perhaps surprisingly, CUTE dominates the case where the system energy expenditures involve only transmit power. This is because as mentioned in Remark 3.4.2 transmitting at rate $e_i(t)$ is energy-opportunistic.

Fig. 3.9 (b) shows the average energy and delay when

$$r_i(t) = \min\left[q_i(t), \frac{c_i(t)}{n(t)}\right], \; i \in \mathcal{A}(t) \tag{3.35}$$

i.e., without the energy efficient rate $e_i(t)$. The three curves correspond to offered loads of 10%, 30%, and 50% of the ergodic capacity. Without using $e_i(t)$, power backoff cannot fully realize energy-delay tradeoffs, moreover the adverse effect of idling power emerges when

delay is high. Interestingly, the curve for the offered load of 10% is different from the other two cases. This is because the circuit energy effect is relatively dominant when $\lambda$ is low, see Theorem 4 and Example 10.

Finally, Fig. 3.9 (c) shows the performance of CUTE when (3.35) is replaced by (3.32). Not only are undesirable energy-delay pairs removed but also energy savings can be seen to be significant– as much as 70%. We simulated various offered loads demonstrating that energy saving benefits are higher when the offered load is lower. Comparing subfigure (b) with (c) we see that CUTE significantly improves both energy and delay performance. For example, at an offered load 30%, the delay/energy pair at (3 sec, 225 mJ) in Fig. 3.9 (b) moves to (0.9 sec, 116 mJ) in Fig. 3.9 (c); the delay is reduced more than three times and energy consumption is cut by half. This is not surprising because the energy-efficient mode will serve a user faster than the target to save energy, see Remark 3.4.3.

Results for the case where opportunistic scheduling is used are shown in Fig. 3.10. The availability of perfect channel state information is assumed in simulating opportunistic scheduling. We reduce the time frame length to 1 msec. As with the case of the round-robin scheduling, energy consumption increases as the delay grows but CUTE successfully removes the undesirable energy and delay pairs. The energy consumption is, however, a lot less than the case of round-robin scheduling. For example, comparing Fig. 3.9 (a) and Fig. 3.10 (a) shows that, when the delay is 0.5 second, CUTE with round-robin consumes 140 mJ while CUTE with opportunistic scheduling expends 70 mJ. Comparing Fig. 3.9 (c) and Fig. 3.10 (c) with offered load 50% also shows that both of the energy and delay become less than half.

**Energy-saving of CUTE and interference:** We expect that our energy saving algorithm CUTE is also beneficial in reducing the inter-cell interference. Even though

we did not directly address this problem in the dissertation, reducing the transmission rate contributes to reducing the output power level and thus other cell interference is also reduced. We observed that this was the case when we considered the energy savings of real-time sessions in Section 3.3.

### 3.4.9 What happens in high loaded systems?

As can be seen in Fig. 3.9 (c), when the offered load is 70%, when $q_i \leq \frac{q_{\max}}{16}$ the energy/delay curve of the CUTE algorithm starts growing again. This highlights the fact that optimizing user-perceived energy consumption based on considering what is optimal for fixed numbers of users need not by optimal in the dynamic regime. Specifically the energy optimal rate $e_i(t)$, is specified given a *fixed* number of users, and is a monotonically decreasing function of $n(t)$, see (3.28). Thus when the offered load is high, and so is $n(t)$, users throughputs $\max[e_i(t), q_i(t)]$ may be too slow leading to increased numbers of users in the system and increased idling power consumption. As a consequence, operating at the static energy-optimal rate might not be good for a heavily loaded system. Note, in this respect, the work of [34, 35, 80, 93] may be also problematic.

To circumvent this potential problem one can restrict users from setting their target throughputs $q_i$ too small. Alternatively the energy efficient rate $e_i(t)$ can be modified so it has a lower bound; for example, $\frac{w}{n(t)}$ in (3.28) can be replaced by $\frac{w}{\min(n(t),\bar{n})}$ where $\bar{n}$ is given by $\bar{n} = \frac{\bar{u}}{1-\bar{u}}$. Here $\bar{u}$ is the utilization when the system is declared to be highly loaded. Note that $\bar{n}$ is the average number of users when the average utilization is $\bar{u}$ in $M/GI/1$ processor sharing system. Roughly speaking, when the load is less than $\bar{u}$, $e_i(t)$ operates as before, but when the load exceeds $\bar{u}$, $e_i(t)$ is increased by $\frac{n}{\bar{n}}$ for fast processing. Figs. 3.9 (c) and 3.10 (c) show the results when $\bar{n} = 8$ (dashed lines) and exhibit monotonically decreasing energy/delay curves (and also better performance) even when the system is overloaded.

## 3.5 Conclusion

This work is, to our knowledge, the first to study energy saving techniques for wireless systems subject to dynamic loads. The key idea is simple: to reduce uplink transmit power, but, to do so in a manner that neither leads to excessive idling/circuit power, nor degrades user perceived performance. We found that idling power, which was previously neglected in static systems, plays a crucial role in energy-efficiency when systems are dynamic, specifically for file transfers. By contrast, the impact of circuit power, which has been addressed in previous work, is limited and asymptotically negligible as the system load grows. Future broadband wireless systems promise to deliver much higher capacity, but in some cases at a much higher energy cost. As such, given the importance of battery lifetimes for mobile terminals, and potential savings in the uplink transmit energy on the order of more than 50% for real-time sessions and 35–75% for file transfers exhibited in this thesis, our approach appears to be quite promising.

## Appendix I
## Finding the optimal Lagrange multiplier

We determine the optimal Lagrange multiplier $\kappa^*$ based on an iterative method that exhibits superlinear convergence. Let $\delta$ denote the uplink utilization of the system, i.e.,

$$\delta = \sum_{i \in \mathcal{A}} \frac{r_i}{x_i}. \tag{3.36}$$

By substituting (3.13) into (3.36), we have

$$\delta(\kappa) = \frac{1}{w} \sum_{i \in \mathcal{A}} \frac{r_i}{W\left(\frac{(\kappa+\xi_i)\gamma_i - 1}{e}\right) + 1}. \tag{3.37}$$

Note that $\delta(\kappa)$ is a convex and monotone decreasing function of $\kappa$. From KKT conditions in (3.10), the optimal $\mathbf{x}$ satisfies $\delta < 1$ if and only if $\kappa^* = 0$. Otherwise $\delta = 1$. So consider setting the initial value as $\kappa_0 = 0$ and let us check two possible cases for $\delta(\kappa_0)$.

***Case 1)*** If $\delta(\kappa_0) \leq 1$, then $\kappa^* = 0$ and $x_i^*$ and $p_i^*$ are determined from (3.13) and (3.14). This is the case for Example 5.

***Case 2)*** If $\delta(\kappa_0) > 1$, the rate vector $\mathbf{x}$ is not feasible, and $\kappa$ should be increased until $\delta(\kappa)$ equals 1. Since $\delta(\kappa)$ is convex and monotonically decreasing in $\kappa$, Newton's method can be used to solve $\delta(\kappa) = 1$ iteratively, i.e.,

$$\kappa_{m+1} = \max\left[\kappa_m - \frac{\delta(\kappa_m) - 1}{\delta'(\kappa_m)}, \kappa^{\min}\right] \tag{3.38}$$

where

$$\delta'(\kappa) = -\frac{1}{w}\sum_{i\in\mathcal{A}} \frac{r_i W'\left(\frac{(\kappa+\xi_i)\gamma_i - 1}{e}\right)}{\left(W\left(\frac{(\kappa+\xi_i)\gamma_i - 1}{e}\right) + 1\right)^2}\frac{\gamma_i}{e} \tag{3.39}$$

and $W'(z) = \frac{W(z)}{z(1+W(z))}$ if $z \neq 0$, and $W'(0) = 1$ [31]. Although $\kappa_m$ converges to $\kappa^*$ superlinearly (because it is Newton's method [24]), a good initial value further reduces the number of iterations. In particular we start the iteration at $\kappa^{\min}$ where

$$\kappa^{\min} = \left[\min_i \left(\frac{\Big(\exp(v)(v-1)+1\Big)}{\gamma_i} - \xi_i\right)\right]^+ \tag{3.40}$$

and $v = \frac{\sum_{i\in\mathcal{A}} r_i}{w}$. Because $\delta(\kappa^{\min}) > 1$, $\lim_{\kappa\to\infty}\delta(\kappa) < 1$, and $\delta(\kappa)$ decreases monotonically, $\delta(\kappa)$ finally hits 1. The iteration ends when $\delta(\kappa_m)$ enters the interval $(1-\epsilon, 1)$ where we set $\epsilon = 10^{-6}$. The number of iterations to convergence is mostly less than 10. If starting with an optimal multiplier obtained in the previous time frame, the iterative optimization was found to converge after 3 – 5 iterations in a system with time-correlated Rayleigh fading channels.

## Appendix II

*Proof of Theorem 5.* If $\bar{r}_i(t)$ converges, then, from (3.30), it is obvious that $r_i(t)$ also converges to the same value. So, we only show that $\bar{r}_i(t)$ converges to $\min(\max(q_i, e_i), \frac{c_i}{n})$. By

substituting (3.32) into (3.30),

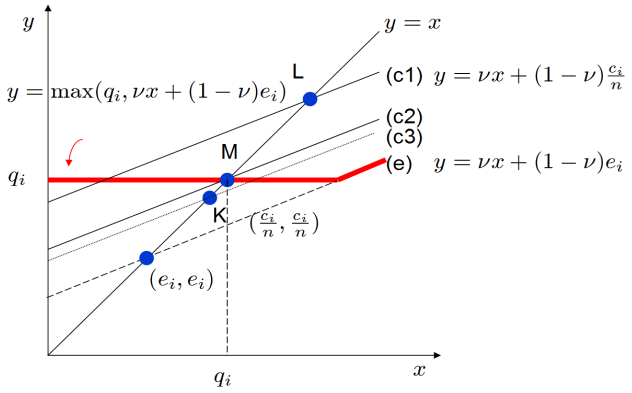
$$\begin{aligned}\bar{r}_i(t) &= \nu\bar{r}_i(t-1) + (1-\nu)\min\Big(\max\Big(\frac{q_i - \nu\bar{r}_i(t-1)\nu}{1-\nu}, e_i\Big), \frac{c_i}{n}\Big) \\ &= \min\Big(\max\big(q_i, \nu\bar{r}_i(t-1) + (1-\nu)e_i\big), \nu\bar{r}_i(t-1) + (1-\nu)\frac{c_i}{n}\Big).\end{aligned}$$

Let $f(x) = \min\big(\max\big(q_i, \nu x + (1-\nu)e_i\big), \nu x + (1-\nu)\frac{c_i}{n}\big)$. Then, $\bar{r}_i(t) = f(\bar{r}_i(t-1))$, and we show that $\bar{r}_i(t)$ converges to $\min(\max(q_i, e_i), \frac{c_i}{n})$ by considering the fixed point equation $f(x) = x$ and the geometry of the iteration. In Fig. 3.11 (a) where $q_i \geq e_i$, if $q_i$ is feasible, i.e., $\frac{c_i}{n} \geq q_i$ it is obvious from the figure that the convergence point is $M = (q_i, q_i)$, i.e., the intersection of $y = x$ and line (e) $y = \max(q_i, \nu x + (1-\nu)e_i)$. If $\frac{c_i}{n} < q_i$ as plotted by line (c3), the convergence point is $K = (\frac{c_i}{n}, \frac{c_i}{n})$. So, $\bar{r}_i(t)$ converges to $\min(q_i, \frac{c_i}{n})$. Similarly, in Fig. 3.11 (b) where $q_i < e_i$, if $e_i$ is feasible, i.e., $\frac{c_i}{n} \geq e_i$ it is obvious from the figure that the convergence point is $M = (e_i, e_i)$, i.e., the intersection of $y = x$ and line (e) $y = \max(q_i, \nu x + (1-\nu)e_i)$. If $\frac{c_i}{n} < e_i$ as plotted by line (c2), the convergence point is $K = (\frac{c_i}{n}, \frac{c_i}{n})$. So, $\bar{r}_i(t)$ converges to $\min(e_i, \frac{c_i}{n})$. Combining these two results completes the proof. □

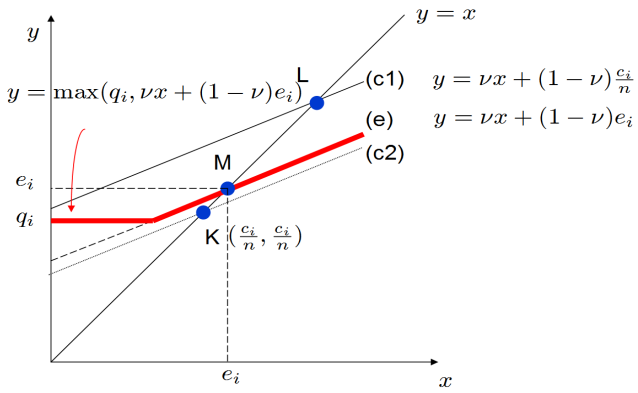
*Proof of Theorem 6.* If $y = x$ intersects $y = q_i$, $\bar{r}_i(t)$ converges in one iteration. If $y = x$ intersects $y = \nu x + (1-\nu)z_i$ where $z_i$ is either $\frac{c_i}{n}$ or $e_i$, $\bar{r}_i(t)$ converges to $z_i$ exponentially fast because

$$\begin{aligned}\bar{r}_i(t+1) &= f(\bar{r}_i(t)) = \nu\bar{r}_i(t) + (1-\nu)z_i \\ \Big|\frac{\bar{r}_i(t+1) - z_i}{\bar{r}_i(t) - z_i}\Big| &= \nu,\end{aligned}$$

and $0 < \nu < 1$. Thus, large $\nu$ means slow convergence. □




(a) In the case of $q_i \geq e_i$, $\bar{r}_i(t)$ converges to $\min(q_i, \frac{c_i}{n})$.




(b) In the case of $q_i < e_i$, $\bar{r}_i(t)$ converges to $\min(e_i, \frac{c_i}{n})$.

Figure 3.11: Geometric proof of convergence theorem.

# Chapter 4

# Energy-Efficient Adaptive MIMO Systems Leveraging Spare Capacity

## 4.1 Introduction

In Chapter 3 we showed that by exploiting dynamic spare capacity one can achieve significant energy savings at mobile terminals. In this chapter we further extend previous study for the case where mobile terminals have multiple antennas. As we have seen in Chapter 3, circuit power is one of the important factors to consider in achieving energy-efficient transmission. The impact of circuit power on energy efficiency is more critical when the mobile terminal (MT) has multiple transmit antennas because of the multiplicity of associated circuits such as mixers, synthesizers, digital-to-analog converters, filters, etc. The circuit power for a multiple-input multiple-output (MIMO) system is thus higher than that of a single-input multiple-output (SIMO) system by approximately $N_t$ times where $N_t$ is the number of transmit antennas. It is generally accepted that MIMO achieves better energy-efficiency than SIMO thanks to spatial multiplexing gain [46]. Altogether when circuit power is considered, however, MIMO may consume more power than SIMO at low spectral efficiency, thus circuit power hinders the use of MIMO on the uplink. This is one of the reasons why several emerging standards do not use MIMO for the uplink [40]. Mitigating the adverse impact of circuit power remains crucial to enabling the use of MIMO for the uplink.

Towards addressing the circuit power problem of MIMO systems, we identify a *crossover point* on the transmission rate (or spectral efficiency) below which SIMO is more

energy efficient than MIMO. We will focus on the case where $N_t = 2$ at the MT, which is perhaps the most practical assumption at this point given the antenna configurations of the IEEE802.16m standard [1]. We propose an *adaptive* switching mechanism between MIMO and SIMO. The key idea is simple. When the system is underutilized, the MT operates with SIMO at low spectral efficiency to save energy, but when congested, the MT operates with MIMO at high spectral efficiency to increase throughput. This is done in an adaptive way considering two aspects – dynamic network traffic and channel variations. In determining the crossover point, the circuit power is the main factor, but we will see that two other factors, the number of receive antennas and channel correlation also increase the crossover point and make mode switching more beneficial.

Prior work on adaptive MIMO techniques [26, 43, 51] has not specifically addressed energy conservation. The authors in [51] proposed mode selection criteria to improve link level bit error rate (BER) performance for a fixed rate. To increase throughput, several adaptive MIMO and link adaptation techniques have been proposed [27, 28, 43], but the authors mainly focused on the physical layer. Power-efficient MIMO systems were studied in terms of transmit diversity [83] and input covariance [117]. By contrast, our work is a *cross-layer* energy saving approach considering the role of *circuit power* at the circuit level, *multiple antennas* at the physical layer, and *dynamic user load* at the medium access control (MAC) layer and above.

One of the challenges in saving energy lies in the tradeoff between transmit energy and circuit energy; slowing down the transmission rate reduces transmit energy [95] but in turn increases circuit energy [23, 32, 34, 80, 93]. Thus, the total energy consumption becomes a convex function of the transmission rate, and an energy-optimal transmission rate exists. Previous work towards achieving energy-optimal transmission, e.g., [34], is

limited to physical layer modulation techniques with a single sender and receiver pair for sensor networks. A link-level multiantenna approach was proposed in [23], which adapts the transmission mode packet-per-packet among space-division multiplexing, space-time coding, and single antenna transmission in a wireless local area networks. The work in [35, 93] addresses multiple users including the MAC layer, but only for a *fixed* number of users in a wireless local area network. Unlike previous work [23, 32, 34, 35, 80, 93], our focus is on *dynamic* multi-users in a *cellular* system where new file transfers are initiated at random and leave the system after being served, i.e., "flow-level dynamics" [21].

**Our contributions.** This chapter makes three main contributions. First, we propose a mechanism for adaptively switching between MIMO and SIMO to conserve mobile terminals' energy. In a practical MIMO system with two transmit antennas at the MT and many receive antennas at the BS, we demonstrate that adaptive switching can save uplink transmission energy by more than 50% as compared to MIMO without substantially changing user-perceived performance. In addition our asymptotic analysis shows that the crossover point scales as $O(\log_2 N_r)$ where $N_r$ is the number of receive antennas at the BS, and thus increasing $N_r$ may improve mode switching benefits.

Our second contribution is to show that mode switching benefits are more significant when channels are correlated. Based on the exponential correlation model [75, 91], a closed form expression for the crossover point is provided as a function of a correlation coefficient. If MIMO uses a zero forcing receiver, the benefit of mode switching further increases because SIMO is more robust in ill-conditioned channels.

Our third contribution lies in that we are the first to consider exploiting *dynamic spare capacity* to realize energy savings in MIMO systems. Dynamic spare capacity is available when the system is underutilized, occasionally, due to changes in user population

and/or bursty traffic loads. Energy is saved by slowing down transmission rates when the system is underutilized. Circuit and idling power, however, deteriorate the energy saving benefit, and the total energy consumption may increase if the user's target throughput is too low. The proposed algorithm effectively avoids this problem by exploiting an energy-optimal transmission rate. We also propose an energy-opportunistic scheduler exploiting both multi-user and multi-mode diversity to further enhance the energy efficiency.

This chapter is organized as follows. In Section 4.2 we describe our system model and assumptions. Section 4.3 analyzes the impact of channel correlation and $N_r$ on the crossover point mainly from static single user scenario. We address the dynamic multi-user scenario and develop a practical energy-efficient adaptive MIMO algorithm in Section 4.4. Section 4.5 provides simulation results followed by conclusion in Section 4.6.

## 4.2 System model

### 4.2.1 Assumptions

We consider a centralized wireless communication system with one BS serving multiple MTs. Target systems could be, but are not limited to, WiMAX or 3GPP-LTE. We assume that the system is based on MIMO and shared via time division multiple access (TDMA). Since energy savings are more important at the MTs than at the BS, we focus on uplink transmissions. Our work is, however, also applicable to saving downlink energy at the BS. We assume that the channels experience flat fading[1] and the dimension of channel matrix $\mathbf{H}$ is $N_r \times N_t$ where $N_r$ is the number of receive antennas at the BS and $N_t$ is the number of transmit antennas at the MT. We focus on the case where $N_r \geq 2$ (at the BS) and $N_t = 2$ (at the MTs). The assumption of two transmit antennas at the MTs

[1]Flat fading can be obtained in practice using multiple input multiple output orthogonal frequency division multiplexing (MIMO-OFDM).

is in accordance with the antenna configurations of the IEEE802.16m standard [1].[2] One might question that current practical systems only use single transmit antenna on the uplink because of the implementation issues such as antenna spacing and circuit power [40]. In that case, our study becomes more meaningful because adaptive mode switching resolves the adverse impact of circuit power and justifies the use of two transmit antennas on the uplink.

We consider MIMO systems where the transmitter does not have channel state information (CSI), i.e., no instantaneous channel feedback. Thus, the transmission mode decision, either MIMO or SIMO, is made at the BS and fed back to the MS, which requires 1 bit of feedback. In addition, in the case of SIMO, the BS informs the MT of the index for the antenna with the highest channel gain, which requires an additional 1 bit of feedback. Our focus is on delivering delay-tolerant (best effort) traffic.

Concerning the MIMO channel model used in this chapter, we will derive the transmission power equation in Section 4.2.3 based on a general channel matrix. Then, we assume Rayleigh fading channel in Section 4.3 for the analysis of the crossover point and also for our simulations in Section 4.5. Note that Rayleigh fading is the most popular analytically tractable stochastic channel model where the components of the channel matrix coefficients are i.i.d. complex Gaussian random variable with zero mean and unit variance. We further consider the correlated channel based on a Kronecker correlation structure with transmit correlation [75, 91]. Even though Rayleigh fading gives analytical tractability, intensive research is still ongoing to better understand fading channels considering, for example, antenna depolarization, co- and cross-polarization isolations and mutual coupling, etc [15].

[2]Even though the 3GPP LTE considers two antennas at the MT but only one antenna is used for uplink.

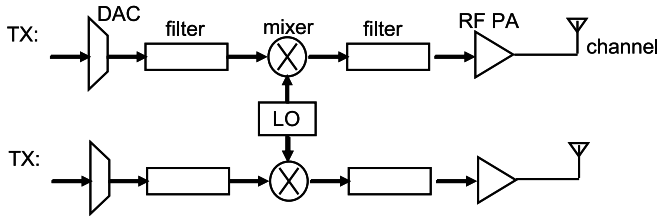



Figure 4.1: Transmission chain for a MIMO system with two antennas.

### 4.2.2 Problem definition

The key questions addressed in this chapter are 1) how to change transmission mode between MIMO and SIMO to save energy in a system supporting dynamic user populations (*mode switching*), and 2) how to determine the appropriate transmission rate considering circuit and idling power consumption[3] as well as the average target throughput of each user (*rate selection*).

### 4.2.3 Transmission power models

Fig. 1 (redrawn from [32]) illustrates the transmission chain for MIMO. A key element of our work is to have a reasonably accurate transmission power model, which we discuss next.

#### 4.2.3.1 MIMO power model

Assuming the MTs do not have access to CSI, we consider equal power allocation to each antenna, and thus do not consider precoding. Nevertheless for the actual transmission, the MS requires another kinds of feedback, for example, to implement the adaptive modulation and coding, and to adjust the tranmist power level in accordance with the pathloss.

[3]The definition and the impact of idling power, which plays a crucial role in dynamic systems, will be addressed in Section 4.4 in detail.

Our work is also applicable to closed loop MIMO, but it is harder to derive a closed form expression for the transmission power as a function of the rate. Let $\phi_1$ and $\phi_2$ be the eigenvalues of $\mathbf{H}^*\mathbf{H}$ where $\mathbf{H}^*$ is a complex conjugate of the channel matrix $\mathbf{H}$. Then, the achievable spectral efficiency of MIMO using equal power allocation into each antenna given $\mathbf{H}$ is expressed as [46]

$$C = \sum_{k=1}^{N_t} \log_2\Big(1 + \frac{P_o}{N_0 N_t}\phi_k\Big) \tag{4.1}$$

where $N_t = 2$ and $P_o$ is the output power (from the power amplifier) that is dissipated into the air, and $N_0$ is the noise power. We assume that $\mathbf{H}$ has a rank of $\min(N_t, N_r) = N_t$. Note that the channel matrix $\mathbf{H}$ is general. For example, if $\mathbf{H} = \mathbf{R}_r^{\frac{1}{2}}\mathbf{H}_w$ where $\mathbf{R}_r$ is a receive correlation matrix, then $\mathbf{H}$ captures the correlated channels at the receiver [91]. If $\mathbf{H} = \mathbf{H}_w$ where the elements of $\mathbf{H}_w$ are independent complex Gaussian random variables with zero mean, $\mathbf{H}$ models uncorrelated Rayleigh fading channels.

The MIMO transmission power model is derived from (4.1) as follows. With $N_t = 2$, we have

$$\begin{aligned} C &= \sum_{k=1}^{2} \log_2\Big(1 + \frac{P_o}{N_0 N_t}\phi_k\Big) && (4.2)\\ &= \log_2\left[\left(\frac{P_o}{2N_0}\right)^2 \phi_1\phi_2 + \frac{\phi_1 + \phi_2}{2N_0}P_o + 1\right], && (4.3) \end{aligned}$$

which leads to the following quadratic equation of $P_o$,

$$\phi_1\phi_2\left(\frac{P_o}{2N_0}\right)^2 + (\phi_1 + \phi_2)\frac{P_o}{2N_0} + (1 - 2^C) = 0. \tag{4.4}$$

The feasible solution to the above quadratic equation is given by

$$P_o = \frac{2N_0}{\phi_1\phi_2}\left[\sqrt{\left[\frac{\phi_1 + \phi_2}{2}\right]^2 + \phi_1\phi_2(2^C - 1)} - \frac{\phi_1 + \phi_2}{2}\right]. \tag{4.5}$$

Assuming that the power consumed by the power amplifier is linearly dependent on the output power [32], the power consumed in the power amplifier can be modeled as $\frac{P_o}{\eta}$ where $\eta$ is the drain efficiency of the power amplifier. Drain efficiency is the ratio of the output power, i.e., the electromagnetically radiated power into the air versus the power consumed in the power amplifier. The transmission power equation $f_m(r)$ for MIMO at transmission rate $r$ with spectral bandwidth $w$ and circuit power consumption $p_{\mathrm{dc},m}$ is then given by

$$f_m(r) = \frac{2N_0}{\eta\phi_1\phi_2}\left(\sqrt{\left(\frac{\phi_1+\phi_2}{2}\right)^2+\phi_1\phi_2(2^{r/w}-1)}-\frac{\phi_1+\phi_2}{2}\right)+p_{\mathrm{dc},m} \tag{4.6}$$

where the subscript $m$ stands for MIMO. For simplicity, we assume that possible transmission rates are continuous.[4]

Note that (4.1) and (4.6) are based on an ideal MIMO receiver. As an example of a practical linear receiver, we here consider a zero forcing receiver which gives us analytical tractability. Then as shown in [91], under an independent coding and detection assumption, (4.1) can be rewritten as,

$$C=\sum_{k=1}^{N_t}\log_2\left(1+\frac{P_o}{N_0N_t}\frac{1}{[(\mathbf{H}^*\mathbf{H})^{-1}]_{k,k}}\right) \tag{4.7}$$

where $[(\mathbf{H}^*\mathbf{H})^{-1}]_{k,k}$ denotes $k$th diagonal element of $(\mathbf{H}^*\mathbf{H})^{-1}$. Thus, if $\phi_k$ is replaced by $1/[(\mathbf{H}^*\mathbf{H})^{-1}]_{k,k}$ in (4.6), we obtain a transmission power model of MIMO with a zero forcing receiver. In the case of MIMO with a minimum mean square error (MMSE) receiver, we can obtain a transmission power model in a similar way, so for simplicity we only consider a zero forcing receiver.

In computing the circuit power, we assume that MIMO requires $N_t$ transmission blocks, but that the frequency synthesizer, i.e., local oscillator (LO), is shared by multiple

[4]For the discrete transmission rate, i.e., finite modulation order with BER constraint, see [32].

antennas [4, 32] as can be seen in Fig. 4.1. Then, the total circuit power consumption of MIMO is given by

$$p_{\mathrm{dc},m} = N_t(p_{\mathrm{dac}} + p_{\mathrm{mix}} + p_{\mathrm{filt}}) + p_{\mathrm{syn}} \tag{4.8}$$

where $p_{\mathrm{dac}}$, $p_{\mathrm{mix}}$, $p_{\mathrm{filt}}$, $p_{\mathrm{syn}}$ stand for the power consumption from a digital-to-analog converter, a mixer, a filter, and a frequency synthesizer, respectively. Our notation is summarized in Table 4.1.

#### 4.2.3.2 SIMO power model

The SIMO power equation can be derived from (1) with $N_t = 1$ and $\phi_k = \mathbf{h}_k^* \mathbf{h}_k = \sum_{\ell=1}^{N_r} |h_{\ell,k}|^2$ where $\mathbf{h}_k$ is the $k$-th column vector of $\mathbf{H}$, and $k = 1, 2$. Then, in selecting the transmit antenna out of the two, $\mathbf{h}_k$ with higher $\phi_k$ is chosen for the transmission. This decision is made at the BS, which implies 1 bit feedback from the BS to the MS is required. Then, the capacity can be expressed as

$$C = \log_2 \left[ 1 + \frac{P_o}{N_0} \sum_{\ell=1}^{N_r} |h_{\ell,\hat{k}}|^2 \right] \tag{4.9}$$

where $\hat{k}$ denotes the index for the selected antenna. From (4.9), we have

$$P_o = \frac{2^C - 1}{\sum_{\ell=1}^{N_r} |h_{\ell,\hat{k}}|^2} N_0, \tag{4.10}$$

and substituting $C$ with $\frac{r}{w}$ and introducing the drain efficiency $\eta$ and the circuit power of SIMO, $p_{\mathrm{dc},s}$, the transmission power $f_s(r)$ where the subscript $s$ stands for SIMO, is given by

$$f_s(r) = \frac{1}{\eta} \frac{2^{r/w} - 1}{\sum_{\ell=1}^{N_r} |h_{\ell,\hat{k}}|^2} N_0 + p_{\mathrm{dc},s}. \tag{4.11}$$

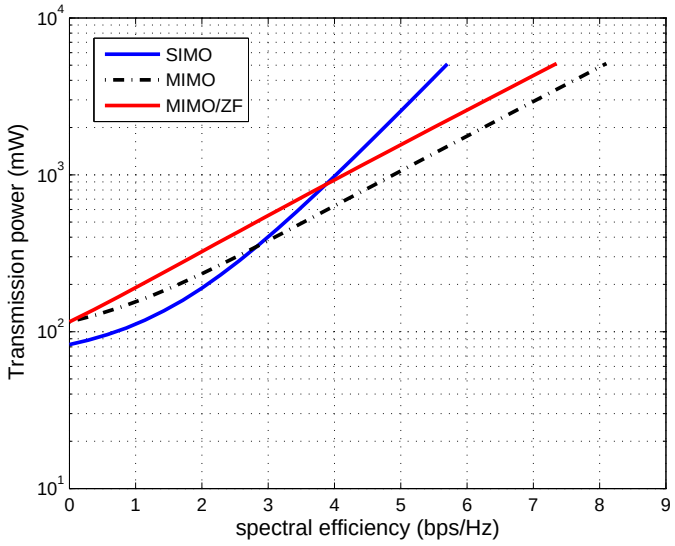



Figure 4.2: Transmission power consumption of mobile terminals including circuit power.

## 4.3 Analysis of the crossover point

### 4.3.1 Motivation for mode switching

Fig. 4.2 exhibits the transmission powers for both MIMO and SIMO in the case of Rayleigh fading channels. Note the crossover around $r/w = 3$ bps/Hz below which SIMO is more energy-efficient than MIMO. (This figure is an example of one realization for an uncorrelated Rayleigh fading channel – different realizations will give different results.) In addition, as an example of MIMO with a linear receiver, we plot the transmission power of MIMO with a zero forcing receiver. It is noticeable that the crossover point is higher than that for the ideal receiver, i.e., $r/w = 3.9$ bps/Hz. The crossover points exhibit the need for a smart switching policy between MIMO and SIMO considering the transmission rate, user-perceived throughput and energy efficiency.

### 4.3.2 The impact of channel correlation on the crossover point

Besides the circuit power, another important factor determining MIMO energy efficiency is spatial correlation among antennas. Since channel correlation degrades the capacity of MIMO systems [91], it further motivates mode switching. The rate regime where SIMO is more energy-efficient than MIMO expands further and thus mode switching becomes more plausible. To capture correlated channels, consider a channel model

$$\mathbf{H} = \mathbf{R}_r^{\frac{1}{2}} \mathbf{H}_w \tag{4.12}$$

where $\mathbf{R}_r$ is an $N_r \times N_r$ receive correlation matrix whose elements are $[\mathbf{R}_r]_{i,j} = \xi^{|i-j|}$, $0 \leq \xi < 1$. This model, called exponential correlation model, is extensively used in the literature [75, 91]. Since receive antennas at the BS are not surrounded by many scatters, it is reasonable to assume spatial correlation at receive antennas. By contrast, we assume that transmit antennas at the MT are not correlated for simplicity; otherwise, MIMO capacity is further degraded, and we have more chance to switch into SIMO. Even though we consider generally the case where $N_t = 2$ and $N_r \geq 2$, here we shall focus on the case where $N_t = N_r = 2$ to allow simple analysis. In this case, we can explicitly compute the crossover point in a high SNR regime. This result is given in the following proposition.

**Proposition 4.3.1.** *Assuming high received SNR, when a correlated channel matrix* $\mathbf{H}$ *is given by (4.12) with* $N_t = N_r = 2$*, the crossover point is explicitly given by*

$$r^* \approx 2log_2 \left\{ \frac{\varphi}{\sqrt{\phi_{w1}\phi_{w2}(1-\xi^2)}} + \sqrt{\frac{\varphi^2}{\phi_{w1}\phi_{w2}(1-\xi^2)} + \Delta p_{dc} \frac{\eta \varphi g}{N_0}} \right\} w \tag{4.13}$$

*where* $\xi$ *is the correlation coefficient,* $\phi_{w1}$ *and* $\phi_{w2}$ *are the eigenvalues of* $\mathbf{H}_w \mathbf{H}_w^*$*,* $\varphi := \mathbf{h}^* \mathbf{h}$*,* $\mathbf{h} = \mathbf{R}^{\frac{1}{2}} \mathbf{h}_w$ *and* $\mathbf{h}_w$ *is* $2 \times 1$ *vector whose elements are independent complex Gaussian random variables.*

*Proof.* The achievable spectral efficiency of MIMO with $\mathbf{H}$ in (4.12) and equal power allocation is given by

$$C = \log_2 \left| \mathbf{I} + \frac{P_o g}{N_0 N_t} \mathbf{R}_r^{\frac{1}{2}} \mathbf{H}_w \mathbf{H}_w^* \mathbf{R}_r^{*\frac{1}{2}} \right|. \tag{4.14}$$

Since the correlation matrix $\mathbf{R}_r$ has full rank, (4.14) is approximated under the high SNR regime by [91]

$$C \simeq N_t \log_2 \frac{P_o g}{N_0 N_t} + \log_2 |\mathbf{H}_w \mathbf{H}_w^*| + \log_2 |\mathbf{R}_r|. \tag{4.15}$$

Note $\log_2|\mathbf{R}_r| = (N_r - 1)\log_2\left(1 - \xi^2\right)$. Thus, the more correlated (i.e., $\xi$ is close to 1), the more the capacity is degraded. Since $\phi_{w1}$ and $\phi_{w2}$ are the eigenvalues of $\mathbf{H}_w\mathbf{H}_w^*$, $|\mathbf{H}_w\mathbf{H}_w^*|$ is given by $\phi_{w1}\phi_{w2}$. Then, in the case of $N_t = N_r = 2$, the power equation for MIMO is given by

$$f_m(r) = \frac{N_0 2}{\eta g} \frac{2^{\frac{r}{2w}}}{\sqrt{\phi_{w1}\phi_{w2}(1-\xi^2)}} + p_{\mathrm{dc},m}. \tag{4.16}$$

Similarly, the power equation for SIMO is given by

$$f_s(r) = \frac{N_0}{\eta g} \frac{2^{\frac{r}{w}}}{\varphi} + p_{\mathrm{dc},s}. \tag{4.17}$$

Then, the crossover point shown in (4.13) is obtained by equating $f_m(r)$ and $f_s(r)$. □

Thus, $r^*$ is an increasing function of $\xi$, and channel correlation makes the crossover point higher. Note that even if the circuit power is not factored, and thus $\Delta p_{\mathrm{dc}} = 0$, the crossover point still exists such as $r^* = 2\log_2\left(\frac{2\varphi}{\sqrt{\phi_{w1}\phi_{w2}(1-\xi^2)}}\right) w$, which further emphasizes the importance of mode switching in correlated channels.

### 4.3.3 The impact of the number of receive antennas on the crossover point

The number of receiver antennas at the BS also has an impact on the crossover point. To obtain an analytically tractable result we assume uncorrelated Rayleigh fading channels with many receive antennas at the BS. We first provide the following Lemma.

**Lemma 2.** *Suppose that* $\mathbf{H}$ *is an* $N_r \times N_t$ *matrix whose elements are independent complex Gaussian random variables with zero mean and variance* $g$*, i.e.,* $\mathbf{H}$ *represents uncorrelated Rayleigh fading channels. Then, as* $N_r$ *goes to infinity,* $\frac{1}{N_r g}\mathbf{H}^*\mathbf{H}$ *converges to an identity matrix by the law of large numbers.*

*Proof.* Let $\mathbf{A} = \frac{1}{N_r g}\mathbf{H}^*\mathbf{H}$. Let $[\mathbf{A}]_{i,j}$ denote the $(i,j)$ element of $\mathbf{A}$. Then, $[\mathbf{A}]_{i,j} = \sum_{\ell=1}^{N_r} \frac{1}{N_r g}[\mathbf{H}^*]_{i,\ell}[\mathbf{H}]_{\ell,j} = \sum_{\ell=1}^{N_r} \frac{1}{N_r g}\left([\mathbf{H}]_{\ell,i}\right)^* [\mathbf{H}]_{\ell,j}$. When $i = j$, $[\mathbf{A}]_{i,i} = \frac{1}{N_r}\sum_{\ell=1}^{N_r} |\frac{[\mathbf{H}]_{\ell,i}}{\sqrt{g}}|^2$, which is the mean of the square of $N_r$ independent normal complex Gaussian random variables. Then, as $N_r$ goes to infinity $[\mathbf{A}]_{i,i}$ converges to 1 by the law of large numbers. When $i \neq j$, $[\mathbf{A}]_{i,j}$ is the average of $N_r$ independent zero mean complex random variables. By the law of large numbers, $[\mathbf{A}]_{i,j}$ converges to zero as $N_r$ goes to infinity. Thus $\mathbf{A}$ converges to an identity matrix as $N_r$ goes to infinity. □

Then, two eigenvalues of $\mathbf{H}^*\mathbf{H}$ are asymptotically given by $N_r g$, and we have the following proposition.

**Proposition 4.3.2.** *Assuming uncorrelated Rayleigh fading channels, when* $N_r$ *is large enough (and* $N_t$ *is 2), the energy efficiency crossover point between SIMO and MIMO scales with* $O(log_2 N_r)$ *as* $N_r$ *grows. When* $N_r$ *goes to infinity, the ratio between the crossover point and the maximum achievable rate of MIMO converges to* $\frac{1}{2}$*.*

*Proof.* Assuming large $N_r$ and thus substituting $\phi_1 = \phi_2 = N_r g$ in (4.6), the power equation is given by

$$f_m(r) = \frac{N_0}{\eta g}\frac{2}{N_r}\left(2^{\frac{r}{2w}} - 1\right) + p_{\mathrm{dc},m}. \tag{4.18}$$

Similarly, (4.11) can be rewritten as

$$f_s(r) = \frac{N_0}{\eta g}\frac{1}{N_r}\left(2^{\frac{r}{w}} - 1\right) + p_{\mathrm{dc},s}. \tag{4.19}$$

The crossover point $r^*$ satisfies $f_s(r^*) = f_m(r^*)$. Solving this equation gives

$$r^* = 2\log_2\left(1 + \sqrt{\Delta p_{\mathrm{dc}} \frac{\eta N_r g}{N_0}}\right) w \tag{4.20}$$

where $\Delta p_{\mathrm{dc}} = p_{\mathrm{dc},m} - p_{\mathrm{dc},s}$. Thus, $r^*$ scales with $O(\log_2 N_r)$. The maximum transmission rate for MIMO is given by

$$r_{\max} = 2\log_2\left(1 + \frac{P_o}{2}\frac{N_r g}{N_0}\right) w. \tag{4.21}$$

Thus, the ratio of $r^*$ and $r_{\max}$ is

$$\frac{r^*}{r_{\max}} = \frac{\log_2\left(1 + \sqrt{\Delta p_{\mathrm{dc}} \frac{\eta N_r g}{N_0}}\right)}{\log_2\left(1 + \frac{P_o}{2}\frac{N_r g}{N_0}\right)}. \tag{4.22}$$

Since $\Delta p_{\mathrm{dc}} > 0$, as $N_r$ goes to infinity, $\frac{r^*}{r_{\max}}$ converges to $\frac{1}{2}$. □

Proposition 4.3.2 implies that as the number of receive antennas at the BS grows, the rate regime where SIMO is more energy-efficient than MIMO expands. This is because, as can be seen in (4.18) and (4.19), increasing $N_r$ makes $f_m(r)$ and $f_s(r)$ grow more slowly in $r$, and thus the impact of circuit power becomes dominant, which makes SIMO more energy-efficient. Finally, if $N_r$ is sufficiently large, the system operates at SIMO for the lower half of the feasible rates and then switched to MIMO for the higher half of the feasible rates.

### 4.3.4 Asymptotic analysis for many receive antennas using flow-level dynamics

So far we have focused on the link level performance, i.e., single user scenario. To better understand the impact of large $N_r$ on energy efficiency, we now proceed to a multi-user scenario and assume the number of ongoing users varies with time, i.e., a *dynamic* system. For large $N_r$, we have shown that the eigenvalues of $\mathbf{H}^*\mathbf{H}$ are approximately $N_r g$. Then, we perform stationary analysis as follows.

Users randomly arrive to the system according to a Poisson process and leave the system after finishing the file transfer. We are interested in the average energy consumption per file. To capture this, we use a *flow-level queuing model* [21], see Fig. 4.3. Flow-level analysis tracks the arrival and departure process of users. We will assume that each user arrives with exactly one file and thus corresponds to a single flow. We refer to the number of flows in the system $n$ as the system's state in the sequel.

Our objective is to minimize the average energy per file by switching between MIMO and SIMO transmission modes. For analytical simplicity, we assume that users have the same target throughput $q$ and are served via temporally fair TDMA scheduling. Then, the system capacity[5] in state $n$ is given by

$$c(n) = \min(nq, r_{\max}). \tag{4.23}$$

The system capacity increases linearly to satisfy the individual targets until the system is overloaded, i.e., $c(n) = r_{\max}$. Assuming a processor sharing scheduling discipline, if the system is not overloaded each user should see his target throughput $q$. This policy represents a simple approach towards exploiting *dynamic spare capacity* to conserve energy; when the system is congested, it operates at the maximum rate $r_{\max}$, however, when underutilized, the overall transmit power and the system capacity are reduced with $n$.

Given the above simple model for system capacity, we now obtain a Markov chain model for the number of ongoing flows in the system. We assume that the arrivals of file transfer requests follow an independent Poisson process with arrival rate $\lambda$ and have independent file sizes with mean $\mu^{-1}$. Let $\mathbf{N} = (N(u), u \geq 0)$ denote a random process representing the number of ongoing file transfers at time $u$. Then, if file sizes are exponentially

[5]System capacity $c(n)$ is not the same as the information theoretic capacity but implies the system throughput.

distributed, $\mathbf{N}$ is a Markov process with state space $\mathbb{Z}^+$ and the rate matrix $Q$ is given by

$$\begin{aligned} q(n, n+1) &= \lambda \\ q(n+1, n) &= \mu c(n+1) \quad \text{for } n \geq 0. \end{aligned}$$

The stationary distribution $\pi$, if it exists, is given by

$$\pi(n) = \pi(0) \frac{\rho^n}{\Pi_{m=1}^n c(m)}, \tag{4.24}$$

where $\rho := \frac{\lambda}{\mu}$ is the traffic load (bits per second) and $\pi(0) = \left(1 + \sum_{n=1}^{\infty} \frac{\rho^n}{\Pi_{m=1}^n c(m)}\right)^{-1}$. Note that the insensitivity property for processor sharing queue ensures this distribution also holds for general file size distributions. In the sequel we let $N$ be a random variable with distribution $\pi$. Let $P$ be random variable denoting the stationary system power consumption. In steady state, the average system power consumption is given by $E[P] = \sum_{n=0}^{\infty} p(n)\pi(n)$ where $p(n)$ is a function which captures the overall system power expenditure in state $n$ and is given by

$$p(n) = \min\left[f_m(c(n)), f_s(c(n))\right]. \tag{4.25}$$

Note that, from the crossover point $r^*$ in (4.20), if $nq \leq r^*$ then SIMO is more energy-efficient, and vice versa. Thus,

$$p(n) = \begin{cases} f_s(c(n)) & \text{if } n \leq n^* \\ f_m(c(n)) & \text{if } n > n^* \end{cases}, \tag{4.26}$$

where $n^* = \lfloor \frac{r^*}{q} \rfloor$. Thus if the number of users is small, e.g., less than or equal to $n^*$, then SIMO is selected, otherwise, MIMO.

Let $J$ be a random variable denoting the energy consumed to serve a typical user's flow. Then, *energy-power equivalence* in a stationary system given in Theorem 3 in Chapter 3 gives that

$$E[J] = \frac{1}{\lambda} E[P]. \tag{4.27}$$

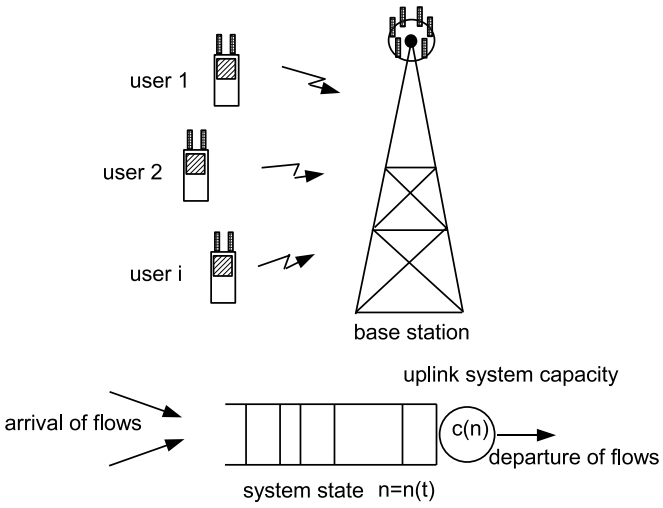



Figure 4.3: Flow-level model for uplink transmission in a dynamic system. One user corresponds to one flow.

Fig. 4.4 shows that $E[J]$ for a system with mode switching decreases faster than that of MIMO as $N_r$ grows. This result can be also anticipated from (4.20), i.e., $r^*$ scales with $O(\log_2 N_r)$. Although practical systems would not be able to employ a large number of antennas at the BS, we expect that our results provide an insight on the impact of receive antennas at the BS in designing practical systems, i.e., increasing $N_r$ improves the energy-saving benefit of mode switching.

## 4.4 Energy-efficient adaptive MIMO in dynamic user populations

In this section we investigate the mode switching combined with rate selection considering *multi-user* scenario in *dynamic* systems and propose a practical solution to realize energy-efficient adaptive MIMO systems. Energy-opportunistic scheduling exploiting multi-mode and multi-user diversity is also proposed to further enhance the energy efficiency.

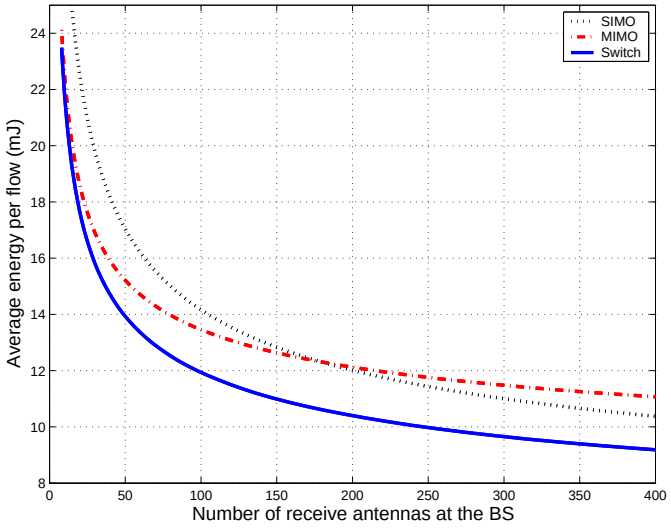



Figure 4.4: Flow-level analysis result for average energy per file vs the number of receive antennas: offered load 30% and $q = 0.01 \times r_{\max}$, i.e., 100 users can share the system without congestion.

### 4.4.1 Simple mode switching

If SIMO and MIMO use the same transmission rate $r$, it is straightforward to choose the best transmission mode; we pick the transmission mode that consumes least power at rate $r$, and the selected mode $\hat{z}$ at rate $r$ is

$$\hat{z}(r) = \underset{z \in \{m,s\}}{\operatorname{argmin}} f_z(r). \tag{4.28}$$

Let us call this *simple mode switching*, where $m$ and $s$ denote MIMO and SIMO respectively.

### 4.4.2 Challenges in mode switching and rate selection

In each mode $z$, however, we need to be careful to choose the transmission rate $r$ considering the tradeoff between transmit and circuit power consumption. As can be seen in (4.6) and (4.11), MIMO and SIMO have different transmit and circuit powers, and thus different energy-optimal transmission rates.

A dynamic user population makes realizing such energy-delay tradeoffs more challenging. To better understand the challenges involved, consider a TDMA system supporting a stationary dynamic load of file transfer requests. If one slows down the uplink transmission rate to save energy then the number of users in the system may grow, resulting in excess power consumption associated with users that *idle* while awaiting transmission. Indeed although ideally idling users turn off their transmission chains, in practice they still consume power due to leakage current[6] [57, 94]. Hence, in a dynamic system, if the transmission rates are excessively reduced, the number of users that are idling may accumulate resulting in excessive overall *idling power* consumption. Consequently, we need to judiciously select the transmission rate to avoid excessive idling power consumption. This makes tradeoffs between energy conservation and delay somewhat complex.

### 4.4.3 Proposed algorithm: CUTE

Next, we describe our proposed rate selection and mode switching algorithm for multiple users with time varying MIMO channels. This algorithm is named CUTE[7], which stands for *Conserving User Terminals' Energy.* The CUTE algorithm resolves two objectives: saving energy and achieving (or exceeding) a target user-perceived throughput. The underlying principle is to switch between SIMO and MIMO adaptively in accordance to the number of users, throughput history and channel fluctuations. In a TDMA system, we assume that time is divided into equal-sized *frames.* A frame is defined as the time period during which all users are scheduled once.

***Rate selection:*** Let $t$ denote the frame index. Let $v_i(t)$ denote the fraction of

[6] Idling power consumption depends on the specific power amplifier design. For example, power amplifier for WiMAX from Analog Devices consumes 2.5 to 25 mW during idling period [94].

[7] This is an extended version of CUTE introduced in Chapter 3, which was for single-input single-output (SISO) system, but we use the same name here.

time allocated to user $i$, and $\sum_i v_i(t) = 1$. Suppose that $n(t)$ users are sharing the uplink channel based on round-robin scheduling with weight $v_i(t)$. If temporally fair scheduling is used, $v_i(t)$ is simply $\frac{1}{n(t)}$. Let $r_{i,z}(t)$ be the transmission rate of user $i$ using transmission mode $z \in \{m, s\}$. We specify the maximum possible transmission rate as $c_{i,z}(t)$, which is determined by the time varying MIMO channel matrix $\mathbf{H}$ and the maximum output power. Since each user is only allocated a fraction $v_i(t)$ of the time frame, the maximum achievable rate of user $i$ is $v_i(t)c_{i,z}(t)$. Let $q_i(t)$ denote the target rate of user $i$. Since file transfers are delay-tolerant, users can specify their own target rate considering their preferences between energy savings and fast transmission. For example, a user with sufficient residual battery may prefer fast transmission, but another user with scarce battery may prefer slow transmission to benefit from the *energy-delay tradeoffs*. Note that the target rate should be independent of $z$, so we do not have a subscript $z$ in $q_i(t)$. Finally, we define an energy-optimal transmission rate as $e_{i,z}(t)$, which captures the circuit and idling power consumption. Then, the transmission rate $r_{i,z}(t)$ is given by

$$r_{i,z}(t) \quad = \quad \min\left[\max\left[e_{i,z}(t), q_i(t)\right], v_i(t)c_{i,z}(t)\right], \tag{4.29}$$

which means that we pick up the maximum of the energy-optimal rate and the target rate (if feasible). The specification of $e_{i,z}(t)$ and $q_i(t)$ are given later.

***Mode switching:*** To achieve $r_{i,z}(t)$ on average during one time frame, the instantaneous rate should be $r_{i,z}(t)/v_i(t)$ because user $i$ only uses a $v_i(t)$ fraction of a time frame, and the corresponding transmission power is $f_{i,z}\big(r_{i,z}(t)/v_i(t)\big)$. So, the energy per bit is given by $\frac{f_{i,z}\left(r_{i,z}(t)/v_i(t)\right)}{r_{i,z}(t)/v_i(t)}$, and the transmission mode of user $i$ is selected as that with the least energy per bit, i.e.,

$$\hat{z}_i \quad = \quad \underset{z\in\{m,s\}}{\operatorname{argmin}} \frac{f_{i,z}\big(r_{i,z}(t)/v_i(t)\big)}{r_{i,z}(t)/v_i(t)}. \tag{4.30}$$

Note that $r_{i,z}(t)$ might be different for MIMO and SIMO because of different $e_{i,z}(t)$, $v_i(t)$ or $c_{i,z}(t)$. If $r_{i,m}(t) = r_{i,s}(t)$, then the rule in (4.30) is identical to simple mode switching in (4.28). After we determine the mode, the service rate of user $i$ is

$$r_i(t) = r_{i,\hat{z}_i}(t). \tag{4.31}$$

Fig. 4.5 shows the overall operation of the proposed algorithm.

***Target rate*** $q_i(t)$*:* Suppose that user $i$ wants to achieve a throughput $q_i$. Since we focus on best effort traffic, which is assumed to be tolerant to transmission rate variation, we do not need to achieve $q_i$ instantaneously. Instead, we consider achieving $q_i$ on average. Based on an exponential averaging of $r_i(t)$, let us define the *average* rate $\bar{r}_i(t)$ seen by user $i$ up to time frame $t$ as $\bar{r}_i(t) = \bar{r}_i(t-1)\nu + r_i(t)(1-\nu)$ where $0 < \nu < 1$ corresponds to averaging weight on the past. We define a *relaxed target rate* $q_i(t)$ to satisfy $q_i = \bar{r}_i(t-1)\nu + q_i(t)(1-\nu)$ so $q_i(t)$ is given by

$$q_i(t) \;=\; \frac{q_i - \bar{r}_i(t-1)\nu}{1-\nu}, \tag{4.32}$$

which relaxes the time scale over which the performance target should be met. It is shown in Chapter 3 that the use of a relaxed target rate enables additional energy savings.

***Energy-optimal rate*** $e_{i,z}(t)$*:* Given $f_z(r)$ and idling power consumption $p_{\text{idle}}$, we define the energy-optimal transmission rate $e_{i,z}(t)$ as that which minimizes the energy per bit during a time frame such as

$$e_{i,z}(t) = \underset{r}{\operatorname{argmin}} \left[ v_i(t) f_{i,z}\big(r/v_i(t)\big) + \big(1 - v_i(t)\big)\, p_{\text{idle}} \right] \frac{1}{r}, \tag{4.33}$$

which means that user $i$ consumes $f_{i,z}\big(r/v_i(t)\big)$ power for $v_i(t)$ fraction of time and $p_{\text{idle}}$ for $(1 - v_i(t))$ fraction of time. Note that $p_{\text{idle}}$ is independent of $z$. The proposed algorithm converges *exponentially fast* to an equilibrium rate given a fixed $v_i(t) = 1/n(t)$ (i.e., temporally fair scheduling) and channel gains – the proof is given in Chapter 3.

### 4.4.4 Extension to energy-opportunistic scheduling

In order to further reduce transmit energy, CUTE algorithm can be easily extended to exploit multi-user diversity. In addition to the conventional opportunistic scheduling that selects the user who experiences the best channel condition [63], we need to consider the spatial transmission mode and the energy efficiency. The proposed *energy-opportunistic* scheduler selects the pair (the user and its transmission mode) that consumes the minimum energy per bit. To this end we slightly change the definition of the time frame. While in round-robin scheduling all users are scheduled a fraction $v_i(t)$ of each time frame, in energy-opportunistic scheduling only one user is selected and that user takes whole time frame. Hence, the length of time frame needs to shrink to a channel coherence time. Assuming users experience the same average channel gain (small-scale fading) and temporally fair scheduling, we use the following energy-opportunistic scheduling,

$$(i^*, \hat{z}_{i^*}) \quad = \quad \underset{(i \in \mathcal{A}(t),\, z \in \{m,s\})}{\operatorname{argmin}} \frac{f_{i,z}\big(u_{i,z}(t)\big)}{u_{i,z}(t)} \tag{4.34}$$

where $\mathcal{A}(t)$ is a set of all active users and $u_{i,z}(t)$ is defined by

$$u_{i,z}(t) \quad = \quad \min\left[\max\left[e_{i,z}(t), q_i\right], c_{i,z}(t)\right]. \tag{4.35}$$

Note that $v_i(t)c_{i,z}(t)$ in (4.29) is replaced by $c_{i,z}(t)$ because the selected user takes the whole time frame. In addition, (4.32) is modified to

$$q_i(t) = \frac{n(t)q_i - \bar{r}_i(t-1)\nu}{1-\nu} \tag{4.36}$$

where $\bar{r}_i(t)$ is computed during the time frames where user $i$ has been served, and (4.33) is modified to

$$e_{i,z}(t) = \underset{r}{\operatorname{argmin}} \frac{f_{i,z}\big(r\big) + \big(n(t)-1\big)\, p_{\text{idle}}}{r}. \tag{4.37}$$

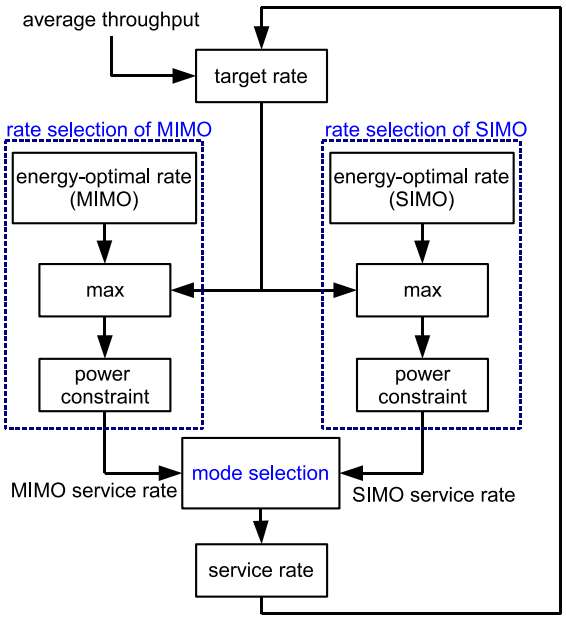



Figure 4.5: Flow chart of the proposed algorithm.

Finally, the service rate of user $i$ is given by

$$r_{i,z}(t) \quad = \quad u_{i,z}(t)\mathbf{1}_{\{(i,z)=s_\theta(t)\}}, \tag{4.38}$$

where $s_\theta(t)$ denotes the pair of the scheduled user and the transmission mode; i.e., $(i^*, \hat{z}_{i^*})$.

## 4.5 Simulation results

To validate the proposed algorithm, we estimate the average energy consumption per file transfer versus the average delay using flow-level event-driven simulations [21]. On each time frame, new user requests arrive according to a Poisson process. Each user requests exactly one file that is log normally distributed with mean 60 kbyte [21]. Users are assumed to experience $N_r \times 2$ spatially correlated Rayleigh fading channels. Our simulation parameters are $\eta = 0.2$, $\nu = 0.95$ for round-robin scheduling and $\nu = 0$ for energy-opportunistic scheduling, $\xi = 0.7$, $w = 1$ MHz, $N_0 = -114$ dBm, $g = -124$ dB, $p_{\text{mix}} = 30.3$ mW,

$p_{\text{syn}}$ = 50.0 mW, $p_{\text{filt}}$ = 20.0 mW, $p_{\text{idle}}$ = 25 mW, $p_{\text{dac}}$ = 15.6 mW, and the maximum output power of power amplifier is 27.5 dBm [4, 32, 94].[8] The duration of a time frame is 5 ms [4], and the number of time frames for the simulation is 1,000,000.

We choose $v_i(t) = \frac{1}{n(t)}$, i.e., temporally fair scheduling for our flow-level simulations. Interestingly in a dynamic system, temporally fair scheduling eventually gives more time resource to users with large files. This is because users with large files remain in the system for a long while users with small files quickly finish their uploads and leave the system. For example, suppose that user 1 with 1 Mbyte file and user 2 with 100 kbyte file share the uplink. As soon as user 2 finishes uploading, user 1 takes the whole time resource.

We plot the energy-delay tradeoff curves for $q_i = (1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32})$ of maximum achievable rate to show how user's preference on energy savings versus fast transmission impacts energy-delay tradeoff. The offered load is 30% of the maximum system capacity.

Fig. 4.6 to Fig. 4.9 show simulation results for MIMO with zero forcing receivers. Fig. 4.6 plots the pair of average delay and average energy per file transfer when circuit or idling power are not present. Three curves correspond to SIMO with antenna selection, MIMO with zero forcing receivers, and simple mode switching (SMS). Interestingly, we see significant energy savings with SMS even though circuit or idling power are not factored. This is because spatial correlation at the receive antennas makes the channel ill-conditioned and degrades the energy efficiency of MIMO. To compare the impact of spatial correlation we also plot the uncorrelated case (dotted lines). We see that energy efficiency of MIMO is greatly affected by the channel correlation while SIMO and SMS are not. Fig. 4.7 shows the result when $N_r$ is changed from 2 to 4. Comparing Fig. 4.7 with Fig. 4.6 exhibits that

[8] Since the value of $p_{\text{filt}}$ in [34] is too low for cellular systems, we adjust it from 2.5 to 20 mW, but the simulation results are almost the same.

increasing $N_r$ alleviates the impact of channel correlation on MIMO. We still see that SMS reduces the energy per file significantly against MIMO, e.g., more than 60% when the delay is 0.5 sec or larger. A dotted line represents SMS with random antenna selection for SIMO, which demonstrates additional energy savings by 1 bit antenna selection indicator.

Fig. 4.8 shows the energy-delay curves when circuit and idling power are factored. As can be seen, SMS saves energy significantly against MIMO. An interesting observation is that three energy curves of SIMO, MIMO and SMS go up again as the delays grow. This is because the effect of idling energy emerges when the file transfer delay is long. Hence, we cannot fully exploit energy-delay tradeoff. This problem is effectively solved by the proposed algorithm CUTE; CUTE removes the undesirable points, (i.e., long delay and large energy consumption) by incorporating the energy-optimal transmission rate; even if the user specifies an excessively low target throughput (and large delay), the proposed algorithm automatically sends *faster* than the user's requirement to save energy. We see that energy savings of CUTE versus MIMO are significant, e.g., more than 50% at 0.5 second delay. Fig. 4.9 shows the energy-delay curves when $N_r$ is changed from 4 to 8. We see that MIMO performance is improved because increasing $N_r$ alleviates the impact of channel correlation. However, still the SMS and CUTE algorithm substantially improve the energy-delay performance.

Fig. 4.10 illustrates the results for MIMO with ideal receivers when $N_r = 2$. Comparing Fig. 4.6 and Fig. 4.10 shows that the energy-delay performance of MIMO is better than that of MIMO with zero forcing receivers. Nevertheless, the performance of SMS and CUTE are almost the same as before, which implies that SIMO plays a major role in energy saving. In Fig. 4.10, we see that SMS performs better than MIMO even without circuit or idling power. This gain comes from SIMO antenna selection and MIMO channel correlation.

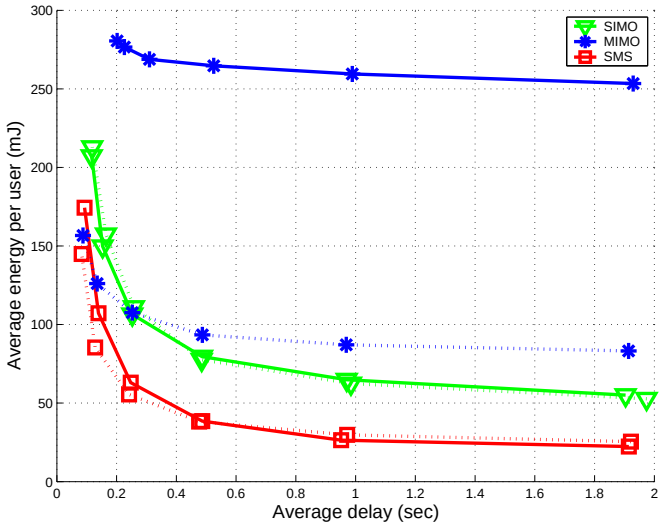



Figure 4.6: Energy-delay tradeoff curves without circuit and idling power: zero forcing receiver for MIMO, $N_r = 2$, $N_t = 2$, traffic load $\rho = 2.51$Mbps, $r_{\max} = 8.35$Mbps, correlation coefficient $\xi = 0.7$ (solid line), $\xi = 0$ (dotted line).

In Fig. 4.11, we also see that CUTE removes the undesirable delay and energy pairs and further improves the energy efficiency.

Fig. 4.12 shows the results under the same environment of Fig. 4.11 except that energy-opportunistic scheduling is adopted instead of round-robin scheduling. Comparing two figures shows substantial performance improvement. For example, the delay/energy pair of (0.54 sec, 46.2 mJ) in Fig. 4.11 shifts to (0.37 sec, 35.3 mJ) in Fig. 4.12, i.e., shorter delay and smaller energy consumption. Furthermore, considering the lower bound of the energy consumption is determined by the circuit energy consumption 22.2 mJ in the case of SIMO, the energy-opportunistic scheduling achieves additional 45% reduction of transmit energy at least.

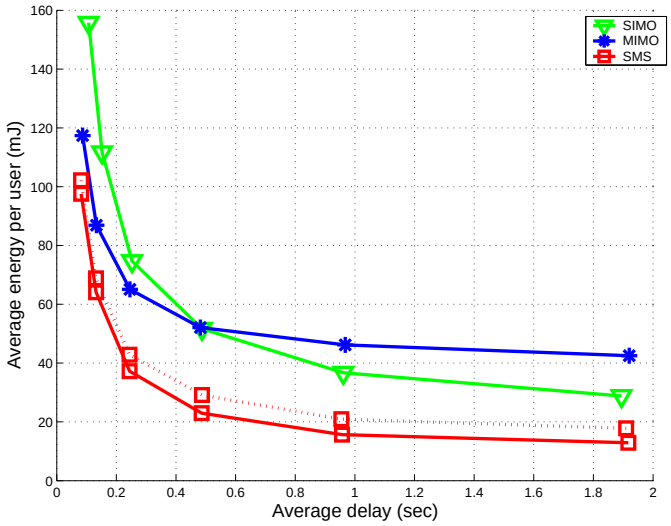



Figure 4.7: Energy-delay tradeoff curves without circuit and idling power: zero forcing receiver for MIMO, $N_r = 4$, $N_t = 2$, traffic load $\rho = 3.70$Mbps, $r_{\max} = 12.34$Mbps. Dotted line shows SMS with random antenna selection, i.e, without using 1 bit antenna selection indicator for SIMO.

## 4.6 Conclusion

In this chapter, we showed that significant energy-saving is achieved by transmission mode switching between MIMO and SIMO under dynamic loads. Even though MIMO is more energy-efficient than SIMO thanks to multiplexing gains, this may not be true when circuit power is factored. This is because circuit power can be dominant at low transmission rates, and MIMO consumes more circuit power than SIMO. Mode switching saves more energy for the case of MIMO with a zero forcing receiver, which occasionally suffers from ill-conditioned channels. In addition, spatial correlation among receive antennas further requires the mode switching because the energy efficiency of MIMO is degraded due to channel correlation. For large $N_r$ we showed that crossover point scales as $O(\log_2 N_r)$ and thus the benefit of mode switching increases with $N_r$. To capture the dynamic user population, we performed flow-level simulations under Rayleigh fading channels. In doing

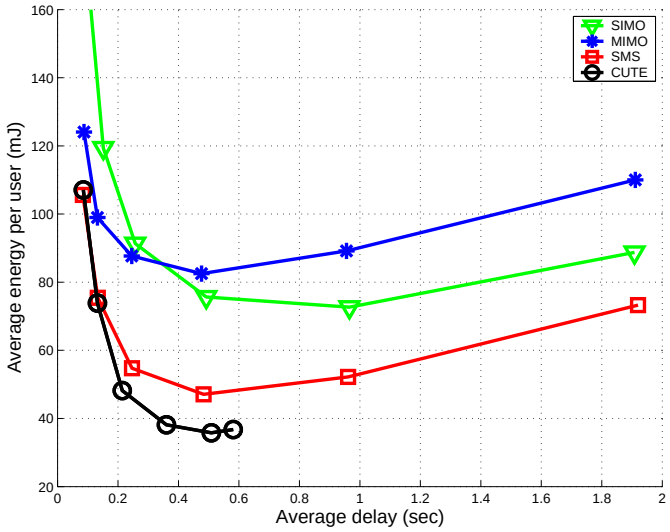



Figure 4.8: Energy-delay tradeoff curves *with* circuit and idling power: zero forcing receiver for MIMO, $N_r = 4$, $N_t = 2$, traffic load $\rho = 3.70$Mbps, $r_{\max} = 12.34$Mbps.

this, we considered the effect of idling power consumption, which led us to investigate the energy-optimal transmission rates, and solved the mode switching problem combined with rate selection. The proposed algorithm CUTE not only exhibited significant energy savings but also eliminated the undesirable operating points with excessive delay and/or energy consumption. Finally, we proposed the energy-opportunistic scheduling to further enhance the energy efficiency. Investigating the optimality of energy-delay tradeoff in a dynamic system remains for future research.

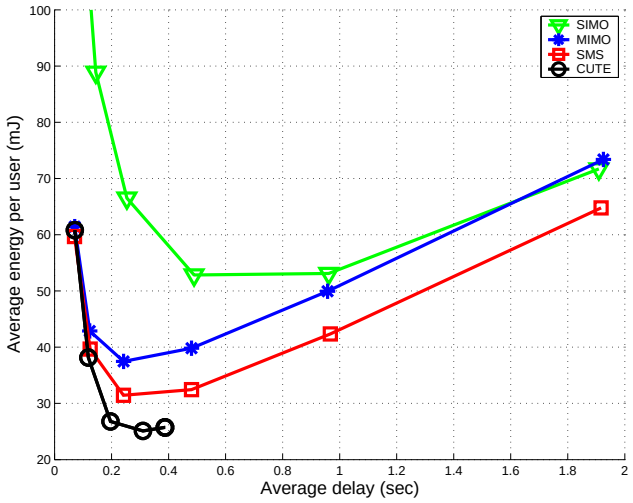



Figure 4.9: Energy-delay tradeoff curves *with* circuit and idling power: zero forcing receiver for MIMO, $N_r = 8$, $N_t = 2$, traffic load $\rho = 4.51$Mbps, $r_{\max} = 15.04$Mbps.

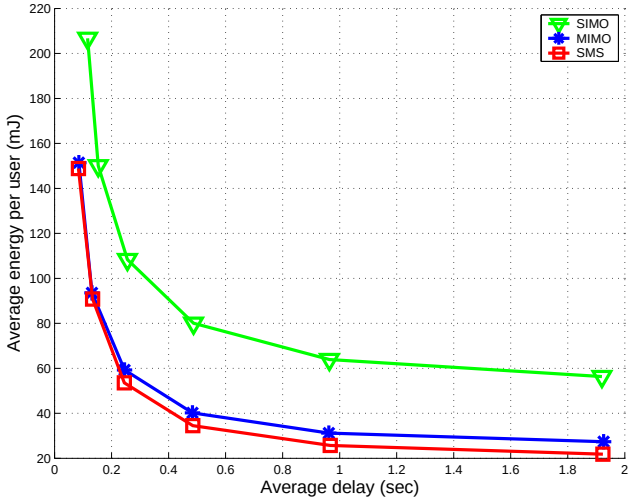



Figure 4.10: Energy-delay tradeoff curves without circuit and idling power: ideal receiver for MIMO, $N_r = 2$, $N_t = 2$, traffic load $\rho = 2.51$Mbps, $r_{\max} = 8.36$Mbps.

Table 4.1: Notation Summary.

| | |
|---|---|
| $\mathbf{H}$ | $N_r \times N_t$ uplink channel matrix |
| $N_r$ | the number of receive antennas at the BS ($\geq 2$) |
| $N_t$ | the number of transmit antennas at the MT ($= 2$) |
| $\phi_k$ | the eigenvalue of $\mathbf{H}^*\mathbf{H}$, $k = 1, 2$ |
| $N_0$ | noise power |
| $P_o$ | output power dissipated into the air from power amplifier |
| $w$ | spectral bandwidth (Hz) |
| $p_{\mathrm{dc,m}}$ | circuit power consumption of MIMO |
| $p_{\mathrm{dc,s}}$ | circuit power consumption of SIMO |
| $\eta$ | drain efficiency of power amplifier |
| $\mathbf{H}_w$ | complex Gaussian random matrix with mean 0 and variance $g$ |
| $g$ | path loss |
| $r^*$ | crossover point below which SIMO is more energy-efficient than MIMO |
| $q$ | target throughput per user |
| $r_{\max}$ | maximum system throughput |
| $\lambda$ | file arrival rate |
| $\mu^{-1}$ | average file size |
| $\rho$ | $:=\frac{\lambda}{\mu}$ traffic load (bps) |
| $\mathbf{R}$ | receive correlation matrix |
| $\xi$ | correlation coefficient |
| $z$ | transmission mode index |
| $i$ | user index |
| $t$ | time frame index |
| $n(t)$ | number of flows (users) at time frame $t$ |
| $v_i(t)$ | the fraction of time allocated to user $i$ |
| $\nu$ | exponential average parameter |

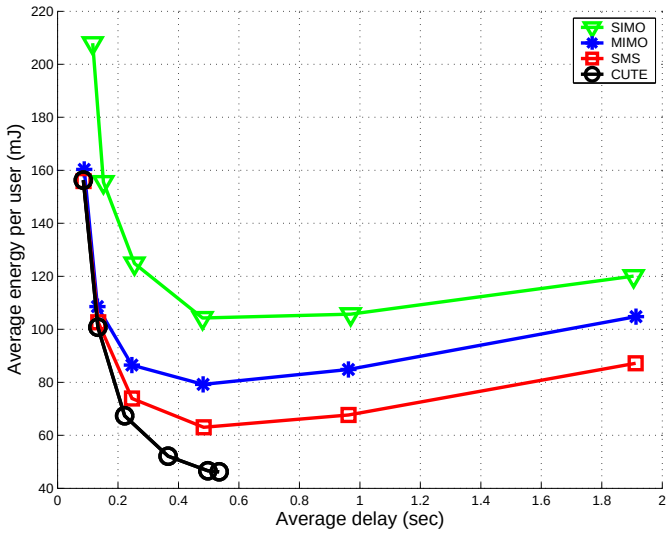



Figure 4.11: Energy-delay tradeoff curves *with* circuit and idling power: ideal receiver for MIMO, $N_r = 2$, $N_t = 2$, traffic load $\rho = 2.51$Mbps, $r_{\max} = 8.36$Mbps.

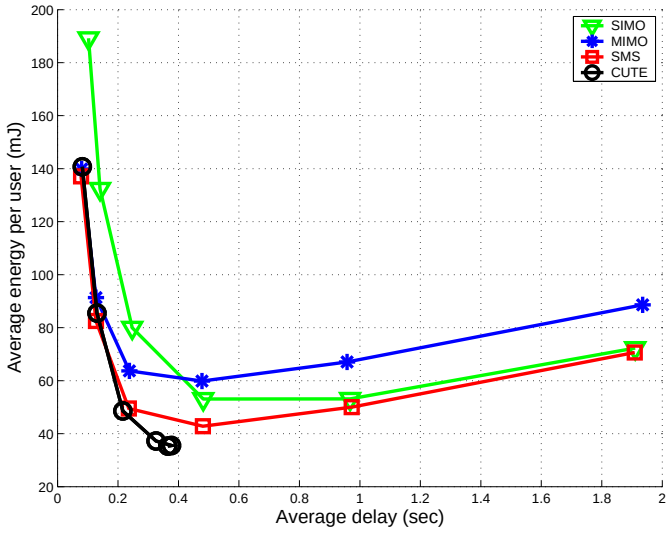



Figure 4.12: Energy-delay tradeoff curves based on energy-opportunistic scheduling (circuit and idling power included): ideal receiver for MIMO, $N_r = 2$, $N_t = 2$, traffic load $\rho = 2.51$Mbps, $r_{\max} = 8.36$Mbps, $\nu = 0$.

# Chapter 5

# $\alpha$-Optimal User Association and Cell Load Balancing in Wireless Networks

## 5.1 Introduction

While Chapters 2–4 focused on tradeoffs in single cell environments, in Chapter 5 we consider multiple cell scenario. One of the important problems in multi-cell data networks is properly associating mobile terminals (MTs) with serving base stations (BSs); this problem is usually called *user association*. In selecting the serving BS, two metrics - instantaneous achievable rate at the physical layer and cell load - should be considered. Since the achievable rate is computed from the received signal-to-interference-plus-noise ratio (SINR), the simplest (and thus widely accepted) rule is to choose the BS that gives the strongest downlink pilot signal. However, this rule is naive in the sense that it does not consider either inter-cell interference or cell load balancing.

In the literature, there have been many efforts towards developing user association rules considering interference avoidance and/or cell load balancing [17, 25, 36, 50, 69, 84, 99, 100, 102, 108, 127]. To avoid interference when frequency is universally reused and inter-cell interference is severe, *centralized* approaches have been considered [17, 69, 99, 108]. The basic idea is to schedule users across cells so that they do not severely interfere with each other. This is called inter-cell coordinated scheduling. Earlier work on load balancing also mostly assumed a centralized controller that governs the BSs and the MTs with access to all the necessary information [25, 36, 50, 84, 127]. However, centralized approaches, for either interference avoidance and/or load balancing, may require excessive computational

complexity and message overhead, which increase exponentially in the size of the network. Such centralized functionality is usually implemented in a server deep in the core network, which only allows slow adaptation at relatively long time scales. To avoid relying on a centralized controller, current systems are usually based on *fractional frequency reuse* or *interference randomization* [1, 40]. Distributed cell load balancing is also being considered as a basic requirement in upcoming standards. For example, IEEE 802.16m WiMAX2 recently included parameters such as cell load and cell type in the system information broadcast [1, 62].

In this chapter we investigate *distributed* user association policies. We will not consider interference avoidance that requires inter-cell coordinated scheduling. So, our approach is reasonable when fractional frequency reuse or interference randomization are being used. We focus on developing a theory and algorithms for user association that adapt to *spatially inhomogeneous* traffic. We consider stochastic traffic loads where new file transfers, or equivalently *flows*, are initiated at random and leave the system after being served – this is sometimes referred to as *flow-level dynamics* [17, 21].

Interestingly, even though user association in a dynamic setting can be viewed as a routing problem among queues, it is still not well understood; most work to date, is ad hoc in nature, and does not address dynamic systems [16, 36, 50, 84, 102, 108, 112]. The work in [17, 69, 127] explores flow-level dynamics for load balancing, but assumes a centralized controller. In particular none of these efforts fully explore the role of load balancing under spatially inhomogeneous traffic distributions.

One of the main challenges in developing a distributed user association policy is achieving global performance optimum without relying on a centralized controller, and doing so to track changes in traffic distributions; for example, day and night have quite

different spatial traffic distributions as may traffic on an hourly (or faster timescale) basis. Our proposed mechanism, denoted *$\alpha$-optimal user association*, effectively overcomes these challenges.

**Contributions.** We highlight the contributions of this chapter as follows. First, we provide a theoretical framework for user association, specifically focused on load balancing under spatially inhomogeneous traffic distributions in an infrastructure-based wireless network. We formulate the user association problem as a convex optimization problem. Then we show a fixed point optimality condition characterizing the spatial partitions (cell coverage areas) associated with minimizing a general system-level performance function. The optimal spatial partition is shown to be unique up to a set of traffic measure zero – this will be explained in the sequel. The optimality condition reveals many interesting facts, e.g. : cell loads are not interchangeable, and *balancing* loads to minimize delay does not imply *equalizing* loads at the BSs; Voronoi cells need not be delay optimal even if the traffic loads are spatially homogeneous; and cell coverage areas need not be contiguous, i.e., can be fragmented.

Second, we present a distributed algorithm and prove its convergence to a global optimum irrespective of the initial condition. Our algorithm could in principle track slowly varying traffic loads. It is also very simple and easily implementable; one need only implement a simple greedy behavior by MTs to achieve a global optimum. The proposed algorithm supports a family of load balancing objectives as $\alpha$ ranges from 0 to $\infty$: rate-optimal ($\alpha = 0$), throughput-optimal ($\alpha \geq 1$), delay-optimal ($\alpha = 2$), and equalizing BS loads ($\alpha = \infty$). Our work is general and applicable to various scenarios. For example, our model for achievable rate at the physical layer can capture shadowing. We do not assume the Tx power of BSs are the same, so our work is also applicable to heterogeneous BS

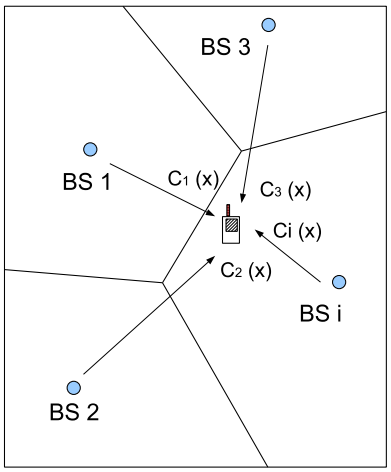



Figure 5.1: User association problem considering the capacity and the traffic loads.

deployments such as macro, micro, pico and even femto cells. Finally, our user association rule can easily address handoffs [62].

Third, we consider possible admission control policies when the system cannot be stabilized or is subject to excessively high loads. The work in [10,11] suggests that admission control is indeed required for best effort traffic in these circumstances. The optimal policy that minimizes our generalized system-level performance function plus blocking cost, results in blocking flows around the boundaries of BS coverage areas. In practice this may not be desirable, so we propose a policy that admits flows at the cell edge with a fixed probability, giving a minimum level of "connectivity" to all users.

**Organization.** This chapter is organized as follows. In Section 5.2, we describe our system model and assumptions. Section 5.3 is devoted to the distributed algorithm and fixed point optimality condition of user association under inhomogeneous traffic distribution. We prove the convergence of our algorithm in Section 5.4. We consider admission control in Section 5.5, and conclude the chapter in Section 5.6.

## 5.2 System Model

### 5.2.1 Assumptions

We consider an infrastructure based wireless communication system with multiple base stations. Target systems could be, but are not limited to, WiMAX2 or 3GPP-LTE. For simplicity, we focus on downlink communications but our method is also applicable to the uplink. We assume that other cell interference is static, and can be considered as noise. We consider a region $\mathcal{L} \in \mathbb{R}^2$ which is served by a set of base stations $\mathcal{B}$. Let $x \in \mathcal{L}$ denote a location and $i \in \mathcal{B}$ be a BS index. We assume that file transfer requests follow an inhomogeneous Poisson point process with arrival rate per unit area $\lambda(x)$ and file sizes which are independently distributed with mean $1/\mu(x)$ at location $x \in \mathcal{L}$, so the traffic load density is defined by $\gamma(x) := \frac{\lambda(x)}{\mu(x)}$; we assume $\gamma(x) < \infty$ for $x \in \mathcal{L}$. This captures spatial traffic variability. For example, a hot spot can be characterized by a high arrival rate and/or possibly large file sizes.

**Definition 1** (traffic load measure)**.** We define the traffic load *measure*, $m(\cdot)$, of a Borel set $\mathcal{G}$ as $m(\mathcal{G}) = \int_{\mathcal{G}} \gamma(x)dx$.

*Assumption* 5.2.1 (capacity function)*.* We assume the physical capacity each BS $i \in \mathcal{B}$ can deliver to location $x$, $c_i(x)$, is a Borel measurable function and for any $\eta > 0$ and $i, j \in \mathcal{B}$, the set

$$\mathcal{D}_{ij}(\eta) = \{x \in \mathcal{L} | c_i(x)/c_j(x) = \eta\} \tag{5.1}$$

has traffic load measure zero, i.e., $m(D_{ij}(\eta)) = 0$. Also to avoid unnecessary technicalities we assume $c_i(x) > 0$ for all $i \in \mathcal{B}$ and $x \in \mathcal{L}$.

As will be seen in the sequel this implies that cell 'boundaries' have zero traffic load measure. Note this model allows for a fairly general but *deterministic* capacity function.

*Remark* 5.2.1. When $c_i(x)$ is discrete valued, $\mathcal{D}_{ij}(\eta)$ may not have traffic load measure zero, so *non-trivial* tie breaking rules are necessary.

For simplicity, we use Shannon capacity function to model the transmission capacity that can be achieved from the BS $i$ to a user at location $x$, i.e.,

$$c_i(x) = \log_2(1 + SINR_i(x)) \tag{5.2}$$

where $SINR_i(x)$ is the received signal to interference plus noise ratio at location $x$ for the signal from BS $i$. Since we assumed that interference is randomized and/or fractional frequency reuse is used to mitigate interference, the sum of total interference power seen by the MT can be simply treated as another Gaussian-like noise [1, 40], and thus $SINR_i(x)$ is given by

$$SINR_i(x) = \frac{P_i g_i(x)}{\sigma^2 + I(x)} \tag{5.3}$$

where $P_i$ denotes the transmission power of the BS $i$, $g_i(x)$ denotes the total channel gain from the BS $i$ to the MT at location $x$, including path loss, shadowing, and others if any. Note that, however, fast fading is not considered here because the time scale for measuring $g_i(x)$ is much larger. $\sigma^2$ is noise power and $I(x)$ is the *average* interference seen by the MT at location $x$. It should be noted that $c_i(x)$ is *location-dependent* but not necessarily determined by the distance from the BS $i$. For example, $c_i(x)$ can be very small in a shadowed area where $g_i(x)$ is very small. Hence, $c_i(x)$ can capture shadowing as well.

The *system-load density* $\varrho_i(x)$ is then defined by $\varrho_i(x) := \frac{\gamma(x)}{c_i(x)}$, which denotes the fraction of time required to deliver traffic load $\gamma(x)$ from BS $i$ to location $x$. We assume that $\min_i \varrho_i(x)$ is finite, i.e., at least one BS has physical capacity to location $x \in \mathcal{L}$ that is not arbitrarily close to zero. Our notation is summarized in Table 5.1.

Table 5.1: Notation Summary

| | |
|---|---|
| $x$ | location in continuous space $\mathcal{L}$ |
| $i \in \mathcal{B}$ | BS index |
| $b$ | $:=\|\mathcal{B}\|$, the number of the BSs |
| $\lambda(x)$ | flow arrival rate per unit area |
| $1/\mu(x)$ | average file size at $x$ |
| $\gamma(x)$ | $:= \frac{\lambda(x)}{\mu(x)}$, inhomogeneous traffic load density |
| $c_i(x)$ | the physical capacity at $x$ from BS $i$ |
| $\varrho_i(x)$ | $:=\frac{\gamma(x)}{c_i(x)}$ system-load density (fractional time) |
| $p_i(x)$ | the routing probability to BS $i$ at $x$ |
| $p(x)$ | $:=(p_1(x), \cdots, p_b(x))$ |
| $\rho_i$ | $:=\int_{\mathcal{L}} \varrho_i(x) p_i(x) dx$ or $\int_{\mathcal{L}_i} \varrho_i(x) dx$ |
| $\rho$ | $:= (\rho_1, \cdots, \rho_b)$ |
| $\mathcal{L}_i$ | the coverage of BS $i$ |
| $\mathcal{P}$ | $:=\{\mathcal{L}_1, \cdots, \mathcal{L}_b\}$, the spatial partition |
| $q_i(x)$ | the binary-valued routing probability to $i$ at $x$ |
| $\mathcal{F}$ | a set of feasible $\rho$ |
| $\partial\mathcal{F}^o$ | a set of $T(\rho)$, $\rho \in \mathcal{F}$ |
| $(k)$ | iteration index in the superscript |

### 5.2.2 Problem formulation

Our problem is to find an optimal user association policy considering the physical capacity and cell load so as to minimize the system cost function given below. In doing this we introduce a routing function $p_i(x)$, which specifies the probability that a flow at location $x$ is associated with BS $i$. We will see that for our system model and Assumption 5.2.1 the optimal routing policy is deterministic, i.e., $p_i^*(x) \in \{0, 1\}$, which also uniquely determines spatial cell coverage areas $\{\mathcal{L}_i\}$.

**Definition 2** (Feasibility)**.** The set $\mathcal{F}$ of *feasible* BS loads $\rho = (\rho_1, \cdots, \rho_b)$, is given by

$$\mathcal{F} = \Big\{\rho \quad | \quad \rho_i = \int_{\mathcal{L}} \varrho_i(x) p_i(x) dx, \tag{5.4}$$

$$0 \leq \rho_i \leq 1 - \epsilon, \tag{5.5}$$

$$\sum_i p_i(x) = 1, \tag{5.6}$$

$$0 \leq p_i(x) \leq 1, \forall i \in \mathcal{B} \text{ and } \forall x \in \mathcal{L}\Big\}, \tag{5.7}$$

where $\epsilon$ is an arbitrarily small positive constant.

**Lemma 3.** *The feasible set $\mathcal{F}$ is convex.*

*Proof.* Consider two load vectors $\rho^1 \in \mathcal{F}$ and $\rho^2 \in \mathcal{F}$, $\rho^1 \neq \rho^2$. Then, there exist associated $p^1(x) = (p_1^1(x), \cdots, p_b^1(x))$ and $p^2(x) = (p_1^2(x), \cdots, p_b^2(x))$ such that $\rho_i^1 = \int \varrho_i(x) p_i^1(x) dx$ and $\rho_i^2 = \int \varrho_i(x) p_i^2(x) dx$ for all $i \in \mathcal{B}$. Now we make $\rho$ as a convex combination of $\rho_1$ and $\rho_2$, i.e., for $\theta \in [0, 1]$, $\rho_i = \theta\rho_i^1 + (1-\theta)\rho_i^2 = \int \varrho_i(x)[\theta p_i^1(x) + (1-\theta)p_i^2(x)]dx$ for all $i \in \mathcal{B}$. Let $p(x)$ be the routing probability associated with $\rho$. Then, $p_i(x) = \theta p_i^1(x) + (1-\theta)p_i^2(x)$, and it satisfies (5.4) to (5.7). Hence $\rho$ is feasible, and so $\mathcal{F}$ is a convex set. □

We formulate our problem as a convex optimization as follows.

**Problem 1**:

$$\min_{\rho} \left\{ \phi_\alpha(\rho) = \sum_i \frac{(1-\rho_i)^{1-\alpha}}{\alpha - 1} \,\Big|\, \rho \in \mathcal{F} \right\} \tag{5.8}$$

where $\alpha \geq 0$ is a parameter specifying the desired degree of load balancing. When $\alpha = 1$ the objective function is defined as $\sum_i \log(\frac{1}{1-\rho_i})$. Problem 1 is said to be feasible if $\mathcal{F}$ is non-empty. Otherwise we shall consider admission control, which will be discussed in Section 5.5.

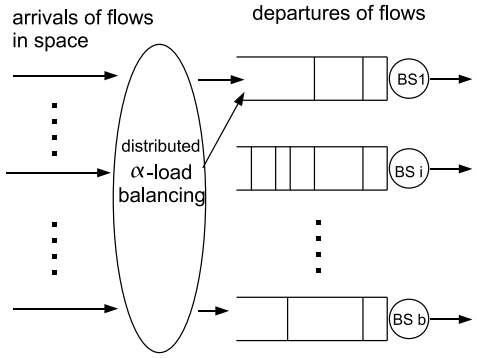



Figure 5.2: Flow-level queueing model for user association problem.

### 5.2.3 Motivation for the objective function

Optimizing $\phi_\alpha(\rho)$ for the case $\alpha = 2$ corresponds to minimizing the overall average flow delay in the system if MTs that are associated with a BS are served by a temporally fair scheduler. Consider a dynamic system where new flows (or file transfer requests) arrive randomly (Poisson) into the system and leave after being served. The dynamics of this system are captured by a *flow-level queuing model* as shown in Fig. 5.2 which tracks the arrival and departure processes of users (or flows, file requests), see e.g., [20, 45, 59].

Let $\mathbf{N_i} = (N_i(t), t \geq 0)$ denote a random process representing the number of ongoing file transfers served by BS $i$ at time $t$. Then, if the system is stationary, the stationary distribution $\pi_i$ of $N_i$ is identical to that of an $M/GI/1$ multi-class processor sharing system [114], and given by $\pi_i(n_i) = (1-\rho_i)\rho_i^{n_i}$. Multi-class reflects the fact that users see different service rates and file sizes based on their locations. We consider infinitely many classes because we address this problem in a continuous space $\mathcal{L}$. The average number of flows at BS $i$ is then simply given by $E[N_i] = \frac{\rho_i}{1-\rho_i}$ and total number of flows in $\mathcal{L}$ is $E[N] = \sum_i E[N_i] = \sum_i \frac{\rho_i}{1-\rho_i}$. From Little's formula, minimizing the average number of flows is equivalent to minimizing the average delay experienced by a *typical* flow. Minimizing $\sum_i \frac{\rho_i}{1-\rho_i}$ is equivalent to (5.8) when $\alpha = 2$ because $\sum_i \left(\frac{\rho_i}{1-\rho_i} + 1\right) = \sum_i \frac{1}{1-\rho_i}$, which

does not change the optimization problem.

### 5.2.4 $\alpha$-optimal user association

Before discussing the optimal user association and how to achieve it, we first discuss the implications of this framework. The solution to Problem 1 gives a unified approach that allows the mobile terminals to select the BS considering signal strength (a user point of view) and the degree of load balancing (the network point of view). Throughout this chapter we will see that if Problem 1 is feasible, the optimal decision made by the mobile terminal located at $x$ is to join BS $i(x)$ given by

$$i(x) = \underset{j \in \mathcal{B}}{\operatorname{argmax}}\, c_j(x)(1-\rho_j^*)^\alpha, \;\; \forall x \in \mathcal{L} \tag{5.9}$$

where $\rho^* = (\rho_1^*, \cdots, \rho_b^*)$ denotes an optimal load vector, i.e., solution to Problem 1.

*Remark* 5.2.2 (Tie-breaking). A location $x \in \mathcal{L}$ is called a cell boundary if a tie of argmax operation in (5.9) happens at $x$. Based on Assumption 5.2.1, cell boundaries have traffic load measure zero; nevertheless, for completeness *if a tie happens, we shall hereafter assume that the MT at such a location chooses the lower indexed BS.*

From (5.9) the mobile terminal chooses a BS that provides the highest physical capacity weighted by a power of *BS's idle time*. By a BS's idle time we refer to the fraction of time it is inactive, i.e., $1-\rho_i$. Depending on the value of $\alpha$ we categorize $\alpha$-optimal user association policies into four cases.

#### 5.2.4.1 Rate-optimal policy

When $\alpha = 0$, the decision is purely based on user's perspective, i.e., based on the physical capacity only (or SINR), and oblivious of network traffic condition. In this case one can show that $\alpha$-optimal user association maximizes the *arithmetic* mean of the BSs'

idle times.

#### 5.2.4.2 Throughput-optimal policy

As $\alpha$ increases, the BS selection criteria gradually shifts from user's perspective to network perspective, and $\alpha = 1$ is a critical point. This is because $\phi_\alpha(\rho)$ goes to infinity with loads close to 1 only if $\alpha \geq 1$ and ensures a stable behavior as long as Problem 1 is feasible. When $\alpha = 1$, it can be shown that the *geometric* mean of the BSs' idle time is maximized.

#### 5.2.4.3 Delay-optimal policy

When $\alpha = 2$, average file transfer delay is minimized as we have seen. In addition, one can show that the *harmonic* mean of the BSs' idle time is maximized.

#### 5.2.4.4 Equalizing-load policy

As $\alpha$ further increases, the rule is such that more emphasis is placed on the traffic loads rather than the physical capacity. One can show that as $\alpha \to \infty$, $\alpha$-optimal user association minimizes the maximum utilization, i.e., min-max utilization, and furthermore it equalizes the utilization of all the BSs.

## 5.3 Distributed Iteration Achieving Optimality

In this section we propose a distributed adaptive user association algorithm that achieves the global optimum of Problem 1 in an iterative manner. The algorithm is simple; BSs periodically share their time average loads with MTs, and MTs use this information to make decisions over these periods. We will show that if spatial loads are temporally stationary, the load vector eventually converges to the unique solution of Problem 1, which in turn

determines spatial coverage areas associated with each BS. However to show convergence we shall assume the following simplifying assumption.

*Assumption* 5.3.1 (Separation of time scales). We shall assume the flow arrival and departure process is very fast relative to the period on which BSs advertise their loads. In particular, once the BSs advertise their load vector, prior to the next update the BSs are able to measure the new steady state loads associated with MT decisions under the advertised vector.

### 5.3.1 Distributed-decision algorithm

The algorithm involves two parts.

***Mobile terminal***: At the start of the $k$-th period MTs receive the load vector $\rho^{(k)}$, e.g., through broadcast control messages from BSs.[1] Then, a new flow request for a MT located at $x$ simply selects the BS $i(x)$ using the deterministic rule given by (5.9) where $\rho^*$ is replaced by $\rho^{(k)}$. Let $\mathcal{L}_i^{(k)}$ denote the coverage area of BS $i$ at $k$-th period. Then, a new spatial partition $\mathcal{P}^{(k)} = \{\mathcal{L}_1^{(k)}, \cdots, \mathcal{L}_b^{(k)}\}$ is determined by $\rho^{(k)}$ and given by

$$\mathcal{L}_i^{(k)} = \left\{ x \in \mathcal{L} | i = \underset{j}{\operatorname{argmax}}\, c_j(x) \left[1 - \rho_j^{(k)}\right]^{\alpha} \right\}, \ \forall i \in \mathcal{B}. \tag{5.10}$$

***Base station:*** During the $k$-th period BSs measure their average utilizations. Due to Assumption 5.3.1, the measured utilization converges to $T_i(\rho^{(k)})$ given by

$$T_i(\rho^{(k)}) = \min \left[ \int_{\mathcal{L}_i^{(k)}} \varrho_i(x) dx, 1 - \epsilon \right], \quad \forall i \in \mathcal{B}. \tag{5.11}$$

Note that the measured utilization, i.e., average busy fractional time of the BS $i$ cannot exceed 1. To avoid unnecessary technicalities we introduce an arbitrarily small positive

[1] IEEE 802.16m facilitates this type of message structure [1, 62].

constant $\epsilon$. It can be shown that $T(\rho) = \{T_i(\rho)\}$ is a continuous mapping defined on $[0, 1-\epsilon]^b$ to itself.

After $T(\rho^{(k)})$ is measured, BSs compute and advertise their next broadcast message $\rho^{(k+1)}$ given by

$$\rho^{(k+1)} = \beta^{(k)}\rho^{(k)} + (1-\beta^{(k)})T(\rho^{(k)}) := S(\rho^{(k)}) \tag{5.12}$$

where $\beta^{(k)} \in [0, 1)$ is an exponential-averaging parameter. It should be noted that $T(\rho^{(k)})$ corresponds to the average loads seen during the $k$-th period while $\rho^{(k)}$ is an exponential average of $T(\rho^{(\ell)})$ across periods, i.e., $\ell = 0, \cdots, k-1$ with some initial loads $\rho^{(0)} \in \mathcal{F}$.

### 5.3.2 Fixed point achieves optimality

Note that if $\rho^{(k)}$ converges it must converge to a fixed point of (5.12), i.e., a solution to

$$\rho^* = T(\rho^*). \tag{5.13}$$

The convergence of (5.12) will be shown in Section 5.4. Below we will show that $T(\cdot)$ has a unique fixed point $\rho^*$ corresponding to the optimal load vector associated with Problem 1.

**Theorem 7.** *Suppose that Problem 1 is feasible. Then, $T$ has a unique fixed point which is the optimal solution to Problem 1. In addition, under Assumption 5.2.1 this fixed point determines a unique optimal spatial partition $\mathcal{P}^*$ up to a set of traffic measure zero.*

*Proof.* Since $T$ is a continuous mapping defined on compact set $[0, 1-\epsilon]^b$ to itself, by Brouwer's fixed point theorem, a solution of $T(\rho^*) = \rho^*$ must exist. Now we prove that $\rho^*$ is the optimal solution of Problem 1. Since $\phi_\alpha(\rho)$ is a convex function over a convex set, if $\rho^*$ satisfies the following condition

$$\langle \nabla\phi_\alpha(\rho^*), \Delta\rho^* \rangle \geq 0 \tag{5.14}$$

for all $\rho \in \mathcal{F}$ where $\Delta\rho^* = \rho - \rho^*$, then $\rho^*$ is the optimal solution of Problem 1.

Let $p(x)$ and $p^*(x)$ be the associated routing probabilities for $\rho$ and $\rho^*$, respectively. From (5.10), we have

$$p_i^*(x) \quad = \quad \mathbf{1}\left\{ i = \underset{j}{\operatorname{argmax}}\, c_j(x)\left(1-\rho_j^*\right)^{\alpha} \right\}, \tag{5.15}$$

and the inner product is computed such as

$$\begin{aligned} \langle \nabla\phi_\alpha(\rho^*), \Delta\rho^* \rangle &= \sum_i \frac{1}{(1-\rho_i^*)^\alpha}\left(\rho_i - \rho_i^*\right) \\ &= \sum_i \frac{\int_{\mathcal{L}} \varrho_i(x)\left(p_i(x) - p_i^*(x)\right) dx}{(1-\rho_i^*)^\alpha} \\ &= \int_{\mathcal{L}} \gamma(x) \left[ \sum_i \frac{p_i(x) - p_i^*(x)}{c_i(x)(1-\rho_i^*)^\alpha} \right] dx. \end{aligned} \tag{5.16}$$

Note that

$$\sum_i \frac{p_i(x)}{c_i(x)(1-\rho_i^*)^\alpha} \geq \sum_i \frac{p_i^*(x)}{c_i(x)(1-\rho_i^*)^\alpha}$$

holds because $p_i^*(x)$ in (5.15) is an indicator for the maximizer of $c_j(x)(1-\rho_j^*)^\alpha$, for all $j \in \mathcal{B}$. Hence, $\langle \nabla\phi_\alpha(\rho^*), \Delta\rho^* \rangle \geq 0$. When $\alpha > 0$ Problem 1 is strictly convex, and $\rho^*$ should be unique, and so is the fixed point. When $\alpha = 0$, the optimal policy selects the BS that gives the highest $c_i(x)$ without considering load. Hence $T(\rho)$ is independent of the load vector $\rho$ and a constant function, which ensures that $\rho^*$ is unique.

In addition we can show that $\rho^*$ has a corresponding spatial partition $\mathcal{P}^* = \{\mathcal{L}_i^*, i \in \mathcal{B}\}$ which is unique up to a set of traffic measure zero. Suppose that there are two such partitions $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$ associated with $\rho^*$, and there exists a set $\mathcal{M} \subset \mathcal{L}$ with non-zero traffic measure where $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$ differ, i.e., user associations are different. In particular, without loss of generality on $\mathcal{M}$, under $\mathcal{P}_1^*$, users at those locations associated with BS 1, while under $\mathcal{P}_2^*$ they associate with BS 2. It follows that on $\mathcal{M}$ there must be a tie, yet by

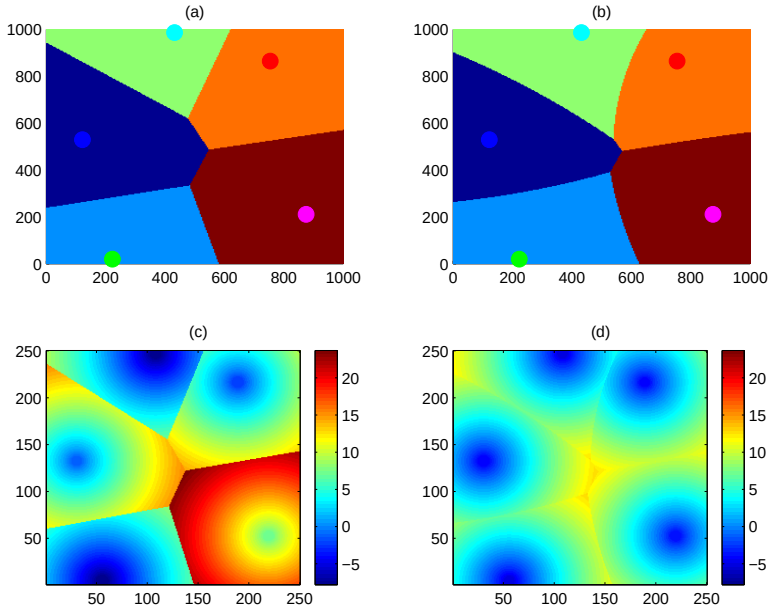



Figure 5.3: Voronoi cells vs Delay-optimal cells (a–b), and spatial distribution of conditional average delay (dB scale) in each case.

Assumption 5.2.1 such sets have traffic measure zero. This is then a contradiction. It follows that the induced partition $\mathcal{P}^*$ is unique except on sets which have zero traffic measure. □

### 5.3.3 Examples

We provide four examples to exhibit the properties of $\alpha$-optimal user association.

**Example 11** (Spatial delay smoothing)**.** *The first example shows the BS coverage areas and geographical distribution of average file transfer delays. Five BSs are randomly placed in 1000m* $\times$ *1000m region. As an example of inhomogeneous traffic loads, a linearly increasing load in the diagonal direction is considered. The Tx power of all the BSs was normalized to* 1*. We assume hereafter that the Tx power is 1, unless otherwise specified, throughout this chapter. In addition,* $c_i(x)$ *is computed using pathloss exponent 3. Fig. 5.3 (a) shows the partition when* $\alpha = 0$ *(Voronoi cells), and Fig. 5.3 (b) shows the partition when* $\alpha = 2$

*(delay-optimal cells). Fig 5.3 (c) and (d) show the conditional average file transfer delays (dB scale) at $x$, which is given by*

$$E[D_i|X=x] = \frac{1}{\lambda(x)}\frac{\varrho_i(x)}{\rho_i}\frac{\rho_i}{1-\rho_i} = \frac{1}{\mu(x)c_i(x)(1-\rho_i)}$$

*in the case of an $M/GI/1$ multi-class processor sharing system model. For simplicity, we set $1/\mu(x) = 1$ and show the average 1-bit transmit time. The benefit of delay-optimal load balancing is clearly shown in Fig. 5.3. A slight modification of the cell coverages significantly improves the delay performance, specifically, of the congested cell at the right lower corner.*

**Example 12** (Voronoi cells vs delay-optimal cells)**.** *One might think that Voronoi cells are delay-optimal for homogeneous traffic loads. However, that is not generally true. We consider a case where the traffic loads are homogeneous, i.e., $\lambda(x) = \lambda$. Then, from (5.15), the delay optimal cell boundary $\ell_{ij}$ for two adjacent cells, $\mathcal{L}_i$ and $\mathcal{L}_j$ is given by*

$$\ell_{ij} = \left\{x|c_i(x)(1-\rho_i^*)^2 = c_j(x)(1-\rho_j^*)^2\right\}. \tag{5.17}$$

*Since $c_i(x) = c_j(x)$ at the Voronoi cell boundaries, (5.17) is satisfied when $\rho_i^* = \rho_j^*$. However, two adjacent Voronoi cells do not necessarily have the same loads in their Voronoi cells so $\rho_i^* = \rho_j^*$ is not guaranteed. In fact Voronoi cells are delay-optimal only if in addition Voronoi cells have the same loads. Fig. 5.4 shows an example of delay-optimal cells that are far from Voronoi cells even though the traffic loads are homogeneous.*

**Example 13** (Fragmented cells)**.** *One might think that coverage areas associated with BSs should be contiguous. However, optimal BS coverage areas may be fragmented. Fig. 5.5 (a) shows an example of fragmented delay-optimal cells. For illustrative purposes $\lambda(x)$ here is given by a quadratic function as shown in Fig. 5.5 (b). Fragmented cells can exist because $(1-\rho_i^*)^\alpha$ and $(1-\rho_j^*)^\alpha$ in (5.17) play a role in determining the boundary. Fig. 5.6 illustrates $c_i(x)(1-\rho_i^*)^\alpha$ and $c_j(x)(1-\rho_j^*)^\alpha$ in 1-D and shows how non-contiguous coverage areas may*

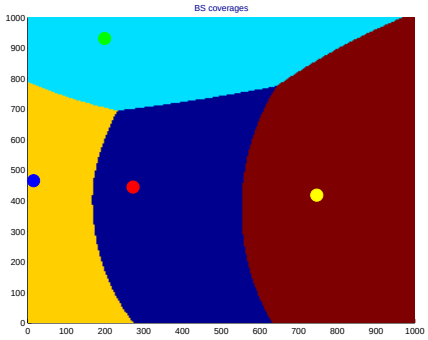



Figure 5.4: Voronoi cells are not delay-optimal even if the traffic loads are homogeneous.

*exist depending on* $\rho_i^*$ *and* $\rho_j^*$. *Fig. 5.5 (c) and (d) show the overall average delay and utilization converge to the delay-optimal ones when the proposed iteration is used.*

**Example 14** (Delay for various $\alpha$)**.** *Fig. 5.7 shows the average delay performance for different* $\alpha$. *For illustrative purposes* $\lambda(x)$ *is same as Fig. 5.5 (b) and four BSs are randomly placed on 1000m* $\times$ *1000m. We exclude the results when* $\alpha < 1$ *because they result in excessive delays. It can be clearly seen that* $\alpha = 2$ *minimizes the average delay.*

*Remark* 5.3.1. Utilizations can be estimated by measuring the average number of flows as well. In an $M/GI/1$ queue, the average number of flows is given by $E[N_i] = \frac{\rho_i}{1-\rho_i}$, which in turn yields $\rho_i = \frac{E[N_i]}{E[N_i]+1}$. Replacing $\rho_i = \frac{E[N_i]}{E[N_i]+1}$ into (5.9) when $\alpha = 1$ gives

$$i(x) = \operatorname*{argmax}_j \frac{c_j(x)}{E[N_j]+1}. \tag{5.18}$$

This rule is a special case of $\alpha$-optimal user association and was earlier proposed as a heuristic in [102, 108] and [50].

## 5.4 Convergence of Distributed Iteration

In this section we prove the convergence of the proposed distributed algorithm, i.e., the convergence of $\rho^{(k)}$. When $\alpha = 0$, $T(\rho)$ is constant and (5.12) with $\beta = 0$ converges in

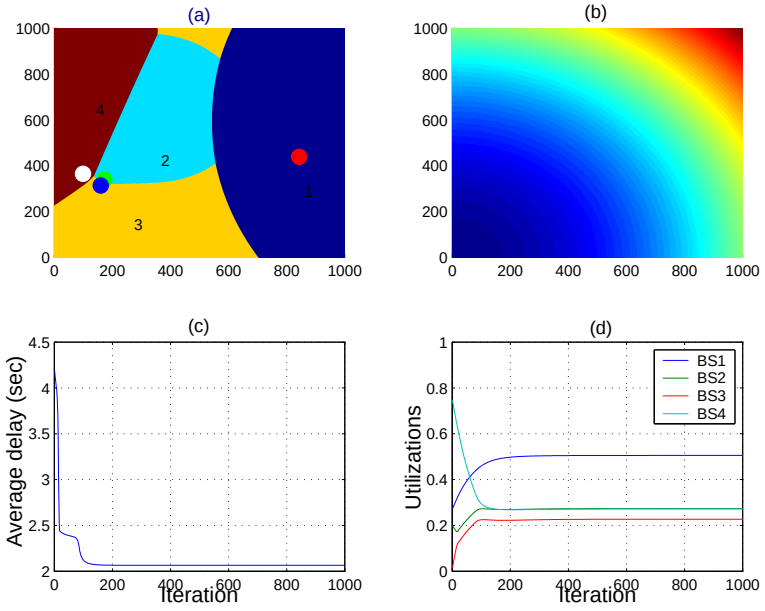



Figure 5.5: Cell coverage areas can be fragmented.

one iteration, so hereafter we focus on the case when $\alpha > 0$.

### 5.4.1 Proof of convergence

As seen earlier the proposed algorithm can be interpreted as iteratively applying a mapping $S$ to an initial load $\rho^{(0)}$. We shall prove the convergence of the loads by first considering the characteristics of $T$ mapping. If $T$ were a contraction mapping, then iterating $T$ would guarantee convergence to the unique fixed point associated with the global optimum. However, $T$ is not necessarily a contraction mapping, in particular when the system is highly loaded. So the proposed algorithm is a damped version of $T$, i.e., $S(\rho) = \beta\rho + (1-\beta)T(\rho)$. We first show the following two lemmas associated with the $T$ mapping, and then prove the convergence of the $S$ mapping.

**Lemma 4.** *If $\rho \in \mathcal{F}$, then $T(\rho)$ is on the boundary of $\mathcal{F}$ that faces the origin, see e.g., Fig 5.8.*

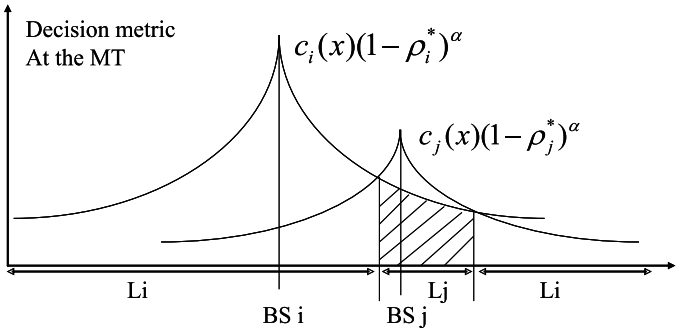



Figure 5.6: Illustration of fragmented cell coverage areas.

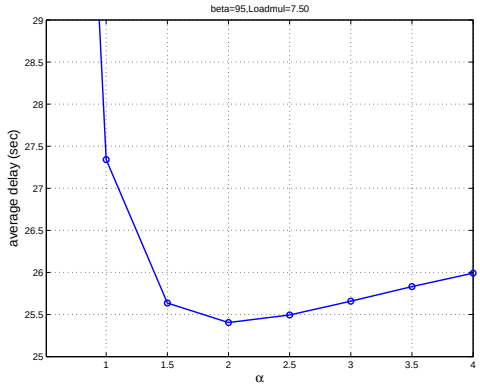



Figure 5.7: Average delay obtained for different values of $\alpha$.

The proof is included in that of Lemma 5.

In the case of two BSs, the fixed point of $T$ can be visualized as shown in Fig. 5.8. The dashed line denotes a set $\partial\mathcal{F}^o = \{T(\rho)|\rho \in \mathcal{F}\}$, i.e., the boundary of $\mathcal{F}$ facing the origin. Since the level sets of $\phi_\alpha(\rho)$ are concave functions (solid lines), $\rho^*$ is the point where the level set touches a convex set $\mathcal{F}$. Note that the shape of $\mathcal{F}$ and $\rho^*$ depend on the spatial traffic distribution.

*Remark* 5.4.1. From (5.11), $T(\rho)$ is associated with deterministic BS coverage areas, and the routing probability that specifies $T(\rho)$ is binary, i.e., either 1 or 0. Hence in describing the routing probability associated with $T(\rho)$ we will use the notation $q_i(x) \in \{0, 1\}$ instead

of $p_i(x)$.

Next we show two interesting properties of $T$ mapping. The first is that $T(\rho) - \rho$ is a descent direction of $\phi_\alpha(\rho)$. The second is that $T(\rho) - \rho$ is a vector that minimizes the inner product with $\nabla\phi_\alpha(\rho)$. This is formally stated in the following lemma.

**Lemma 5** (Descent direction)**.** *For $\rho \in \mathcal{F}$ and $\rho \neq \rho^*$, $T(\rho)$ gives a descent direction at $\rho$, i.e.,*

$$\langle \nabla\phi_\alpha(\rho), T(\rho) - \rho \rangle < 0.$$

*In addition, $T(\rho)$ is the feasible load vector that minimizes the inner product with the gradient at $\rho$, i.e.,*

$$T(\rho) = \underset{\hat{\rho}\in\mathcal{F}}{\operatorname{argmin}} \langle \nabla\phi_\alpha(\rho), \hat{\rho} - \rho \rangle. \tag{5.19}$$

*Proof.* Let $p(x)$ and $q(x)$ be the routing probability associated with $\rho$ and $T(\rho)$, respectively. From (5.11) $T_i$ is associated with deterministic cell coverage area $\mathcal{L}_i$, and thus its routing probability $q_i(x)$ is given by binary, i.e.,

$$q_i(x) = \mathbf{1}\left\{ i = \underset{j}{\operatorname{argmax}}\, c_j(x)(1-\rho_j)^\alpha \right\}, \forall i \in \mathcal{B}, \forall x \in \mathcal{L}, \tag{5.20}$$

with ties broken in favor of lowest index BS. Let $\Delta\rho = T(\rho) - \rho$. Then, $\langle \nabla\phi_\alpha(\rho), \Delta\rho \rangle$ can be computed as follows:

$$\begin{aligned}
\langle \nabla\phi_\alpha(\rho), \Delta\rho \rangle &= \sum_i \frac{T_i(\rho) - \rho_i}{(1-\rho_i)^\alpha} \\
&= \sum_i \frac{\int_{\mathcal{L}} \varrho_i(x)\,(q_i(x) - p_i(x))\,dx}{(1-\rho_i)^\alpha} \\
&= \int_{\mathcal{L}} \gamma(x) \left( \sum_i \frac{q_i(x) - p_i(x)}{c_i(x)(1-\rho_i)^\alpha} \right) dx.
\end{aligned}$$

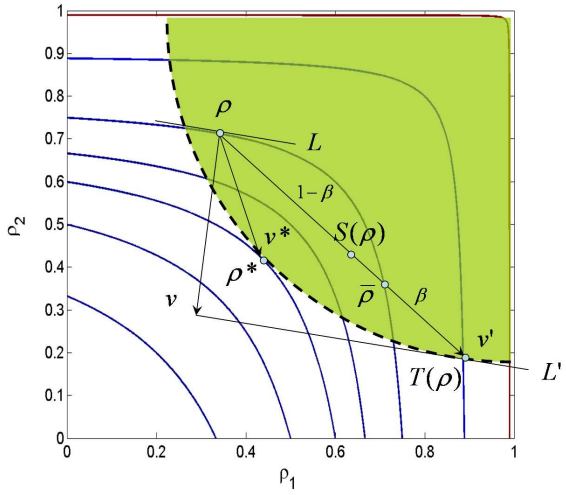



Figure 5.8: Convergence property of $S$ mapping.

By definition $q_i(x)$ satisfies

$$\sum_i \frac{q_i(x) - p_i(x)}{c_i(x)(1-\rho_i)^\alpha} \leq 0. \tag{5.21}$$

Since $\rho \neq \rho^*$ it must be that $p(x) \neq q(x)$ on a set which has non-zero traffic load measure. Then, multiplying (5.21) by $\gamma(x)$ and integrating over $\mathcal{L}$ gives $\langle \nabla \phi_\alpha(\rho), \Delta\rho \rangle < 0$.

Furthermore, we have the following property,

$$q_i(x) \quad = \quad \underset{\hat{p}_i(x)}{\operatorname{argmin}} \sum_i \frac{\hat{p}_i(x) - p_i(x)}{c_i(x)(1-\rho_j)^\alpha}, \tag{5.22}$$

because (5.21) holds for arbitrary $p_i(x) \neq q_i(x)$. Then, multiplying (5.22) with $\gamma(x)$ and integrating (5.22) over $\mathcal{L}$ proves (5.19). Finally, (5.19) implies that $T(\rho)$ is on the boundary of $\mathcal{F}$, and Lemma 4 is proved. □

Fig. 5.8 exhibits $T(\rho)$. Suppose that $v$ is the opposite direction of $\nabla\phi_\alpha(\rho)$ and $L$ is the tangent line of the level set at $\rho$. Then, the feasible vector that maximizes the inner product with $v$ can be found by drawing a line $L'$ which is parallel to $L$ and tangent to the boundary $\partial\mathcal{F}^o$; the tangent point is then $T(\rho)$. In Fig. 5.8 we see the case of

$\phi_\alpha(T(\rho)) > \phi_\alpha(\rho)$, which implies that $T(\rho)$ gives a descent direction, but it does not necessarily result in a monotonic decreasing sequence $\phi_\alpha(\rho^{(k)})$. Indeed $T$ mapping can overshoot along the descent direction, in particular when the loads are high. Introducing the weighting parameter $\beta$ in (5.12) alleviates such overshooting. Fig. 5.8 shows $\phi_\alpha(S(\rho)) < \phi_\alpha(\rho)$ if $S(\rho)$ is selected between $\rho$ and $\bar{\rho}$, where $\bar{\rho}$ is the intersection of $T(\rho) - \rho$ and the level set at $\rho$. Based on this we prove the convergence of $S$ iteration in Lemma 6 and Theorem 8.

**Lemma 6.** *For $\rho \in \mathcal{F}$ and $\rho \neq \rho^*$, there exists $\beta \in [0, 1)$ such that $\phi_\alpha(S(\rho)) < \phi_\alpha(\rho)$.*

*Proof.* Since $S(\rho) - \rho = \beta\rho + (1-\beta)T(\rho) - \rho = (1-\beta)(T(\rho) - \rho)$, $S(\rho) - \rho$ is also a descent direction. Since the level sets of $\phi_\alpha(\rho)$ are strictly concave functions when $\alpha > 0$ and $S(\rho)$ gives a descent direction at $\rho$, there exists a $\beta \in [0, 1)$ that makes $\phi_\alpha(S(\rho)) < \phi_\alpha(\rho)$. □

**Theorem 8** (Convergence)**.** *Suppose that Problem 1 is feasible. If $\rho^{(0)} \in \mathcal{F}$ and $\beta^{(k)}$ is chosen so that $\phi_\alpha(S(\rho^{(k)})) < \phi_\alpha(\rho^{(k)})$, then $\rho^{(k+1)} = S(\rho^{(k)})$ converges to $\rho^*$.*

*Proof.* $\phi_\alpha(\rho^{(k)})$ is a monotonically decreasing sequence in $k$ and also lower-bounded by 0, so $\phi_\alpha(\rho^{(k)})$ converges. Suppose that $\phi_\alpha(\rho^{(k)})$ converges to something other than $\phi_\alpha(\rho^*)$. Then $S$ produces a descent direction again, and by Lemma 6, $\phi_\alpha(\rho^{(k)})$ can further decrease in next iteration. This contradicts the convergence assumption and $\rho^{(k)}$ should converge to $\rho^*$. □

*Remark* 5.4.2. A fixed $\beta$ close to 1 generally works well for the convergence. However, the magnitude of $\beta$ guaranteeing convergence depends on the network load. When the system is not congested, even $\beta = 0$ can guarantee convergence as $T(\rho)$ may be a contraction mapping. However, when the system is congested, $\beta$ needs to be close to 1, e.g., 0.95 to 0.99. The convergence speed also depends on $\beta$. When the loads are low, $\beta$ can be small

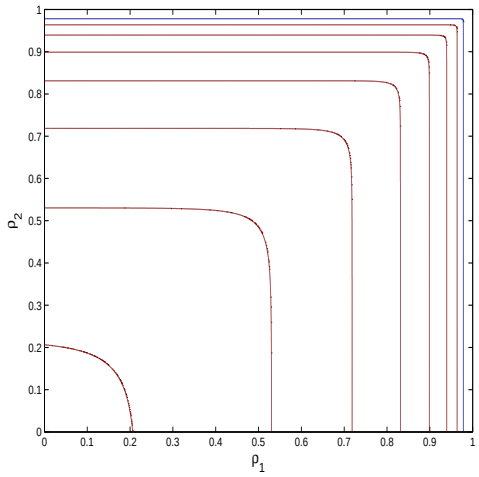



Figure 5.9: Level sets of $\phi_\alpha(\rho)$ when $\alpha = 10$ (two BSs case).

and exhibits fast convergence. In practice $\beta$ is a design parameter that should be selected to balance speed of convergence vs. stable system behavior.

*Remark* 5.4.3. As stated earlier without proof, the optimal solution equalizes $\rho_i$ for all $i \in \mathcal{B}$ when $\alpha = \infty$. This can be easily proven when the level sets of $\phi_\alpha(\rho)$ are plotted. Fig. 5.9 shows the level sets when $\alpha = 10$. In fact the level sets become more and more sharp as $\alpha$ grows, and thus the optimal utilization where the level set touches $\mathcal{F}$ occurs when $\rho_i$ are all equal.

### 5.4.2 The convergence independent of initial condition

So far we assumed $\rho^{(0)} \in \mathcal{F}$, and then $\rho^{(k)}$ remains in the feasible set $\mathcal{F}$ during the iteration. One can however show that the iteration converges to the optimal point as long as $\rho^{(0)} \in [0, 1-\epsilon]^b$. This property is important in real implementation because it makes the algorithm *robust* to changes in the traffic spatial distribution. As an example, suppose that at time $t = t_0$, the stationary file arrival process with $\lambda^1(x)$ changes to another stationary process with $\lambda^2(x)$. However, the optimal solution for $\lambda^1(x)$ may not be in the feasible set

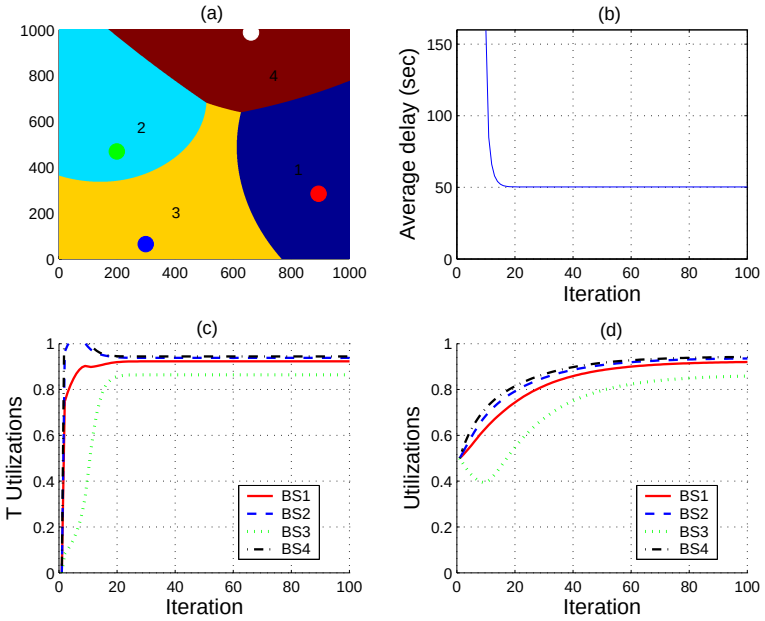



Figure 5.10: Example of convergence: (a) delay-optimal partition (b) average delay (c) $T(\rho^{(k)})$ (d) $\rho^{(k)}$.

associated with $\lambda_2(x)$. Nevertheless, our algorithm would converge to new optimal point. Using an *affine-invariant* property of the $T$ mapping given below, we prove the convergence of the iteration irrespective of the feasibility of the initial condition.

**Definition 3** (Affine set of $\rho$)**.** For $\rho \in [0, 1-\epsilon]^b$, we define an affine set $\mathcal{A}(\rho) = \{\tilde{\rho} \in [0, 1-\epsilon]^b | \tilde{\rho} = \theta\rho + (1-\theta)e, \theta \geq 0\}$ where $e = (1, \cdots, 1)$. Hence $\mathcal{A}$ is a set of points on the line connecting $\rho$ and $e$, see Fig. 5.11.

**Lemma 7** (Affine invariance of $T$)**.** *For $\rho \in [0, 1-\epsilon]^b$ and $\tilde{\rho} \in \mathcal{A}(\rho)$, we have $T(\rho) = T(\tilde{\rho})$, which implies that all the points in the affine set yield the same partition by $T$. In fact $T$ is not a one-to-one mapping but many-to-one, and thus non-invertible.*

*Proof.* From (5.10) we see that scaling of $(1-\rho)$ does not change the partition $\mathcal{P}$ because the decision metrics for all BSs are scaled in the same way. Hence $T(\rho) = T(\tilde{\rho})$. □

**Lemma 8.** *For $\rho \in [0, 1-\epsilon]^b \cap \mathcal{F}^c$, there exist $\tilde{\rho} \in \mathcal{A}(\rho)$ and $\beta, \tilde{\beta} \in [0, 1)$ such that $\phi_\alpha(S(\tilde{\rho})) < \phi_\alpha(\tilde{\rho})$ where $S(\tilde{\rho}) = \tilde{\beta}\tilde{\rho} + (1-\tilde{\beta})T(\tilde{\rho})$, see Fig. 5.11.*

*Proof.* We consider a mirror image of $\rho$, denoted by $\tilde{\rho}$, such that $\tilde{\rho} \in \mathcal{F} \cap \mathcal{A}(\rho)$. Then, $T(\rho) = T(\tilde{\rho})$ by Lemma 7. Note that $\tilde{\rho}$ is not unique. Let $\tilde{L}$ denote a line connecting $\tilde{\rho}$ and $T(\tilde{\rho})$. Similarly, let $L$ denote a line connecting $\rho$ and $T(\rho)$. We pick up $S(\rho)$ with some $\beta$ on line $L$ and determine $S(\tilde{\rho})$ as an intersection of $\tilde{L}$ and $\mathcal{A}(S(\rho))$. From Lemma 6, if $\beta$ is sufficiently close to 1, which in turn implies $\tilde{\beta}$ is also sufficiently close to 1, we have $\phi_\alpha(S(\tilde{\rho})) < \phi_\alpha(\tilde{\rho})$. □

**Theorem 9.** *Suppose that Problem 1 is feasible. If $\rho^{(0)} \in [0, 1-\epsilon]^b$ and $\beta^{(k)}$ is chosen so that $\phi_\alpha(S(\tilde{\rho}^{(k)})) < \phi_\alpha(\tilde{\rho}^{(k)})$, then $\rho^{(k+1)} = S(\rho^{(k)})$ converges to $\rho^*$.*

*Proof.* If $\beta^{(k)}$ is chosen so that $\phi_\alpha(S(\tilde{\rho}^{(k)})) < \phi_\alpha(\tilde{\rho}^{(k)})$, then $\tilde{\rho}^{(k)}$ converges to $\rho^*$ as like Theorem 8. Since $T(\rho^{(k)}) = T(\tilde{\rho}^{(k)})$, $T(\rho^{(k)})$ also converge to $\rho^*$. Since $T$ is a continuous mapping, $\rho^{(k)}$ also converges to $\rho^*$. □

**Example 15.** *Fig. 5.10 shows the convergence of two different utilizations: $T(\rho^{(k)})$ and $\rho^{(k)}$. The iteration can start at any $\rho^{(0)} \in [0, 1-\epsilon]^b$, so we simply pick up $\rho^{(0)} = (0.5, 0.5, 0.5, 0.5)$, which gives Voronoi cells at the first iteration. In this example, traffic loads are chosen so that Voronoi cells cannot stabilize the system. Hence the delays would be infinite for the first few iterations, see Fig. 5.10 (b) and (c). Nevertheless, our algorithm converges quickly to the optimal point. For illustrative purposes $\lambda(x)$ here is given by a quadratic function and $\beta = 0.95$ is used.*

*Remark* 5.4.4 (Throughput-optimality). The proposed algorithm is throughput-optimal when $\alpha \geq 1$. This is because $\phi_\alpha(\rho)$ goes to infinity when $\rho_i$ approaches 1. Then, if the system can be stabilized, there exists a partition $\mathcal{P}$ and corresponding $\rho$ such that $\phi_\alpha(\rho) < \infty$.

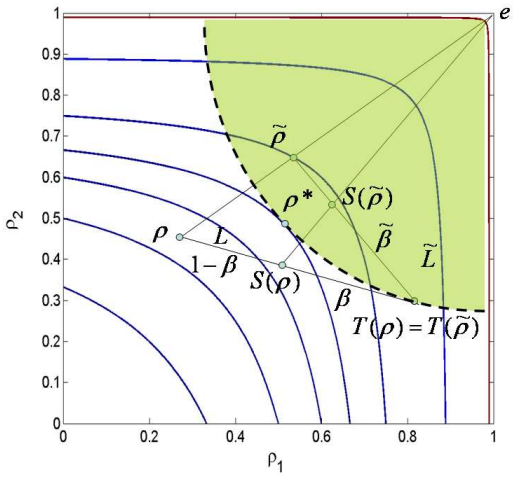



Figure 5.11: Convergence property starting from arbitrary $\rho$.

Then, $\phi_\alpha(\rho^*) \leq \phi_\alpha(\rho)$, i.e., $\phi_\alpha(\rho^*)$ is also finite. Since the algorithm converges to $\rho^*$ for any $\rho^{(0)} \in [0, 1-\epsilon]^b$, it stabilizes the system if the system can be stabilized.

## 5.5 Admission Control

So far we have assumed that Problem 1 is feasible, i.e., the system can be stabilized and Problem 1 has a solution. However, when the traffic loads are too high, the system may not be stabilizable, or may perform very poorly so admission control is required. In this section we consider admission policies for such regimes. Our objective is to minimize a system cost function which includes a cost associated with flow blocking. We assume that blocking cost is proportional to the volume of the blocked traffic. Since the flow blocking determines the user's satisfaction and unsatisfied users may switch operators, such admission control policies may reflect operators' business concerns.

### 5.5.1 Optimality condition

We assume that the flows that are blocked go to a sink, or *null* BS. Let $\mathcal{B}_0$ denote a set of all BSs including BS 0, i.e., the null BS and redefine $\rho$ as $\rho = (\rho_0, \rho_1, \cdots, \rho_b)$. It should be noted that $\rho_0$ is not a utilization; it is defined as $\rho_0 := \int_{\mathcal{L}} \gamma(x) p_0(x) dx$ where $p_0(x)$ is the flow blocking probability at location $x$. Hence, $\rho_0$ can be greater than 1. The total blocking cost is given by $\xi\rho_0$ where $\xi$ is a blocking cost parameter per bit. Parallelly Problem 1, we define a feasible set $\mathcal{F}_0$ including $\rho_0$ as

$$\begin{aligned}\mathcal{F}_0 = \Big\{\rho \quad | \quad & \rho_0 = \int_{\mathcal{L}} \gamma(x) p_0(x) dx, \\ & \rho_i = \int_{\mathcal{L}} \varrho_i(x) p_i(x) dx, \ \ \forall i \in \mathcal{B}, \\ & 0 \le \rho_i \le 1 - \epsilon, \quad \forall i \in \mathcal{B}, \\ & \sum_{i \in \mathcal{B}_0} p_i(x) = 1, \ \ \forall x \in \mathcal{L}, \\ & 0 \le p_i(x) \le 1, \ \ \forall i \in \mathcal{B}_0 \text{ and } \forall x \in \mathcal{L} \Big\}.\end{aligned}$$

It can be shown that $\mathcal{F}_0 \in \mathbb{R}^{b+1}$ is a convex set. Our objective function is then given by

**Problem 2**:

$$\min_{\rho} \left\{ \phi_\alpha^\xi(\rho) = \sum_{i \in \mathcal{B}} \frac{(1-\rho_i)^{(1-\alpha)}}{\alpha - 1} + \xi\rho_0 \Big| \rho \in \mathcal{F}_0 \right\}. \tag{5.23}$$

Note that Problem 2 is a simple convex generalization of Problem 1; if the system can be stabilized, Problem 2 is equivalent to Problem 1 as $\xi$ goes to infinity. An optimality condition for this Problem based on which we can develop adaptive admission control and user association policy is proposed next.

An iterative algorithm to Problem 2 and its behavior are similar to (5.9) to (5.12); at the start of the $k$-th period, MTs choose their serving BSs using (5.9) where $\rho^*$ is replaced

with $\rho^{(k)}$, and BSs update their load vector using (5.12). However, the BS may block a MT based on a threshold; so (5.10) is replaced by

$$\mathcal{L}_i^{(k)} = \left\{ x \in \mathcal{L} | i = \underset{j \in \mathcal{B}_0}{\operatorname{argmax}} \nu_j^{(k)}(x) \right\}, \ \forall i \in \mathcal{B}_0 \tag{5.24}$$

where

$$\nu_j^{(k)}(x) \quad = \quad \begin{cases} \frac{1}{\xi}, & \text{if } j = 0, \\ c_j(x)\left(1 - \rho_j^{(k)}\right)^{\alpha}, & \text{if } j \in \mathcal{B}. \end{cases} \tag{5.25}$$

Note that a BS blocks flows that do not see good performance as compared to $1/\xi$. For example if $\alpha = 1$, the expected throughput of a MT at $x$, i.e., $\max_j c_j(x)(1 - \rho_j^{(k)})$ needs to exceed $1/\xi$ in order not to be blocked. The meaning of a threshold $1/\xi$ depends on $\alpha$: when $\alpha = 0$, $1/\xi$ is simply the minimum achievable rate; when $\alpha = 2$, $\xi$ corresponds to a maximum marginal 1 bit transmit time. In addition to (5.11), we define $T_0(\rho^{(k)}) := \int_{\mathcal{L}_0^{(k)}} \gamma(x) dx$, and $T$ is redefined on $[0, M] \times [0, 1-\epsilon]^b$ to itself where $M < \infty$. One can show that $\rho^{(k)}$ converges to $\rho^*$, i.e., a fixed point of $T$.

**Theorem 10.** *$T$ has a unique fixed point which is the optimal solution to Problem 2. In addition, under Assumption 5.2.1 this fixed point determines a unique optimal spatial partition up to a set of traffic measure zero.*

*Proof.* Since $T$ is a continuous mapping defined on compact set to itself, by Brouwer's fixed point theorem, a solution of $T(\rho^*) = \rho^*$ must exist. Now we prove that $\rho^*$ is the optimal solution of Problem 2. Since $\phi_\alpha^\xi(\rho)$ is a convex function over a convex set, if $\rho^*$ satisfies the following condition

$$\langle \nabla \phi_\alpha^\xi(\rho^*), \Delta\rho^* \rangle \geq 0 \tag{5.26}$$

for all $\rho \in \mathcal{F}$ where $\Delta\rho^* = \rho - \rho^*$, then $\rho^*$ is the optimal solution of Problem 2.

Let $p(x)$ and $p^*(x)$ be the associated routing probabilities for $\rho$ and $\rho^*$, respectively. From (5.24), we have

$$p_i^*(x) \quad = \quad \mathbf{1}\left\{ i = \underset{j}{\operatorname{argmax}}\, \nu_j^*(x) \right\}, \tag{5.27}$$

where $\nu^*(x)$ is given by (5.25) with the optimal $\rho^*$. Then, the inner product is computed such as

$$\begin{aligned}
\langle \nabla \phi_\alpha^\xi(\rho^*), \Delta\rho^* \rangle \quad &= \sum_{i \in \mathcal{B}_0} \frac{\partial \phi_\alpha^\xi}{\partial \rho_i} (\rho_i - \rho_i^*) \\
&= \sum_{i \in \mathcal{B}} \frac{\int_{\mathcal{L}} \varrho_i(x) \left(p_i(x) - p_i^*(x)\right) dx}{(1-\rho_i^*)^\alpha} + \xi \int_{\mathcal{L}} \gamma(x)(p_0(x) - p_0^*(x))dx \\
&= \int_{\mathcal{L}} \gamma(x) \left[ \sum_{i \in \mathcal{B}} \frac{p_i(x) - p_i^*(x)}{c_i(x)(1-\rho_i^*)^\alpha} + \xi(p_0(x) - p_0^*(x)) \right] dx \\
&= \int_{\mathcal{L}} \gamma(x) \left[ \sum_{i \in \mathcal{B}_0} \frac{p_i(x) - p_i^*(x)}{\nu_i^*(x)} \right] dx \\
&\overset{(a)}{\geq} 0
\end{aligned} \tag{5.28}$$

where (a) follows from that $p_i^*(x)$ is the maximizer of $\nu_i^*(x)$ for $i \in \mathcal{B}_0$ (or the minimizer of their inverses). The uniqueness of the spatial partition can be proven similarly as that of Problem 1. □

It is also reported in [17] that admission control under heavily congested system blocks the flows around the cell edge. This is because users at the cell edge consume most of the system resources, i.e., time.

### 5.5.2 Minimal connectivity

Fig. 5.12 shows an example of admission control policy when the traffic load is heavy. For illustrative purposes, four BSs are placed at the four corners and one at the

center. $\lambda(x)$ is given by a quadratic function as shown in Fig. 5.5 (b) and is such that the system is not stabilizable. Fig. 5.12 (a) shows the coverage areas of five BSs. As can be seen, the flows around the cell edge are blocked (bright gray areas). In practice, however, allowing some level of minimal connectivity might be beneficial from a higher layer QoS perspective, i.e., the tradeoff between delay and service outage probability; providing such minimum connectivity will compromise overall delay performance and lead to additional blocking for customers closer to the BSs. To capture this tradeoff we add the following constraint

$$0 \leq p_0(x) \leq 1 - \delta \tag{5.29}$$

in $\mathcal{F}_0$ where $\delta$ specifies the minimum connectivity probability.

**Corollary 2.** *An optimal user association policy of Problem 2 with additional constraint (5.29) is still (5.9), but with probability* $1-\delta$ *the flow is blocked if* $\max_{i\in\mathcal{B}} c_i(x)(1-\rho_i^*)^\alpha < 1/\xi$.

*Proof.* Since $\mathcal{F}^0$ is still a convex set with additional constraint of (5.29), it is sufficent to show that (5.28) is satisfied when $\rho^*$ and its associated $p^*(x)$ are given as follows:

*if* $\max_{i\in\mathcal{B}} \nu_i^*(x) \geq \frac{1}{\xi}$,

$$p_i^*(x) = \mathbf{1}\left\{ i = \underset{j\in\mathcal{B}}{\operatorname{argmax}}\, \nu_j^*(x) \right\}, \ \forall i \in \mathcal{B}_0, \tag{5.30}$$

*otherwise*,

$$p_0^*(x) = 1 - \delta, \tag{5.31}$$

$$p_i^*(x) = \delta \times \mathbf{1}\left\{ i = \underset{j\in\mathcal{B}}{\operatorname{argmax}}\, \nu_j^*(x) \right\}, \ \forall i \in \mathcal{B}. \tag{5.32}$$

The proof is essentially same to the proof of Theorem 10, and condition (a) in (5.28) is satisfied when $p_i^*(x)$ is chosen as stated in (5.30) to (5.32) because $p_i^*(x)$ gives most of its weight on at most two $\nu_i^*(x)$ under the constraint of (5.29). □

Fig. 5.12 (b) shows the areas where the flows are partially admitted with a fixed probability (denoted by PA $i$, $i = 1, \cdots, 5$) and completely admitted (denoted by A $i$). In this example $\delta = 0.5$ is used. Comparing Fig. 5.12 (a) and (b) shows that areas where flows are admitted with probability 1 shrinks as $\delta$ increases from 0 to 0.5. In addition, increasing $\delta$ also degrades overall system performance. Fig. 5.12 (c) shows its trade off; as $\delta$ grows, $\phi_\alpha^\xi(\rho)$ increases. Hence, $\delta$ should be carefully chosen considering the tradeoff between minimum connectivity and performance degradation.

*Remark* 5.5.1. The addition of a minimal connectivity (or admission probability) $\delta$ irrespective of location, means once more that the network may not be stabilizable. That is, even if flows are blocked with probability $1 - \delta$ everywhere, there may not exist a user association policy that stabilizes the remaining load. We envisage the service provider having sufficient knowledge of the traffic loads on its network to balance the selection of the two parameters $\xi$ and $\delta$: balancing blocking (and stability) vs. flow-level performance.

## 5.6 Conclusion

In this chapter we proposed a theoretical (and also practical) framework for user association problem in wireless networks. We specifically focused on distributed load balancing under spatially inhomogeneous traffic distributions and showed the optimality condition of cell coverage areas that minimizes generalized system performance function. Interestingly, the optimal user association policy, i.e., routing of flows to appropriate BSs is deterministic even though probabilistic routing is allowed. This deterministic property enables us to

develop a simple distributed-decision algorithm at the MTs, which is easily implementable and compliant with upcoming standards, e.g., WiMAX2. Our distributed algorithm was shown to converge to the global optimum and also robust to changes of traffic distributions. Finally, our work was extended to the case where system cannot be stabilized due to excessive traffic loads. Under such heavy traffic regimes, we proposed optimal admission control policies considering tradeoffs between two QoS metrics: average delay vs. maintaining a minimum level of connectivity to users independent of their location.

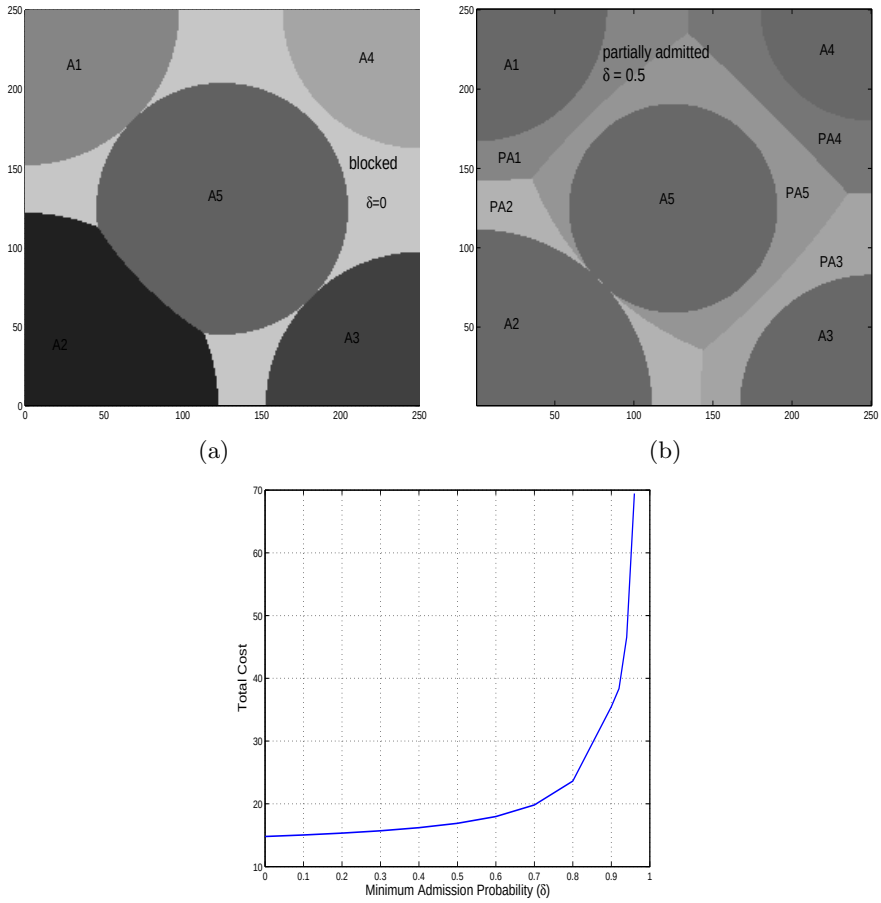



Figure 5.12: Cell coverage areas under admission control. (a) $\delta = 0$ (b) $\delta = 0.5$: each cell has completely admitted area A $i$ and partially admitted area PA $i$. (c) tradeoff between $\delta$ and performance.

# Chapter 6

# Conclusion and Future Work

## 6.1 Summary

In this dissertation we studied flow-level performance under stochastic traffic loads with specific focus on tradeoffs amongst QoS, capacity and energy-efficiency. The emphasis on flow-level dynamics is driven by the desire to better capture user-perceived performance in real systems and reveal properties that cannot be observed in static systems.

We first evaluated the tradeoff between QoS and system capacity when integrating QoS and best effort flows in an opportunistic wireless system. We showed that integration of QoS and best effort flows results in loss in opportunism, which in turn results in a reduction of the stability region, degradation in system throughput, and increased file transfer delay. This occurs because guaranteeing a minimum average bandwidth for a subset of users may compromise opportunism by requiring that users be scheduled that do not currently have good channels. These losses were shown to be proportional to opportunistic gains, the guaranteed bandwidth and the number of QoS flows, but inversely proportional to SNR under a Rayleigh fading channel model.

Second, we proposed and investigated an approach to exploit dynamic spare capacity in wireless system in order to conserve mobile terminals' energy. Mobile terminals adapt their data transmission rate in accordance to the current work utilization; when networks are underutilized, an energy saving mechanism slows down the transmission rates, but does so judiciously in order to prevent network congestion. We considered systems with flow-level

dynamics supporting either real-time or best effort sessions. The energy-optimal transmission strategy for real-time sessions was determined by solving a convex optimization and was shown to achieve a substantial energy savings, e.g., more than 50% when the session blocking probability is 0.1% or less. The case of file transfers proposed to be more subtle because power backoff changes the system dynamics. We studied energy-efficient transmission strategies that enable energy-delay tradeoffs. The proposed mechanism achieved a 35–75% energy savings depending on the load and file transfer target throughput. A key insight, relative to previous work focusing on static scenarios, is that idling power has a significant impact on energy-efficiency, while circuit power has limited impact as the load increases.

We then extended our energy saving technique to multiple antenna systems. We proposed a mechanism to switch between MIMO with two transmit antennas and SIMO in order to conserve mobile terminals energy. We showed that there exists a crossover point for the transmission rate, below which SIMO consumes less power than MIMO when circuit power is included. The crossover point was an increasing function of the circuit power, the number of receive antennas and channel correlation, all of which increased the potential energy savings resulting from mode switching. We proposed an adaptive mode switching algorithm combined with rate selection to maintain users' target throughput while achieving energy efficiency. Extensive flow-level simulations under dynamic loads confirmed that the proposed technique reduced the transmission energy by more than 50% and enabled an effective tradeoff between file transfer delay and energy conservation.

Finally, we developed a framework for user association in infrastructure-based wireless networks, specifically focused on flow-level cell load balancing under spatially inhomogeneous traffic distributions. We proved that the optimal load vector that minimizes a generalized system performance function is the fixed point of a certain mapping. An itera-

tive distributed adaptive user association policy was proposed and was shown to converge to a globally optimal policy in steady state. We further addressed admission control policies for the case where the system is overloaded. For an appropriate system level cost function the optimal admission control policy blocks all flows at cell's edges. However, providing a minimum level of connectivity to all spatial locations might be desirable, to this end a channel and load dependent random blocking and user association policy were proposed.

## 6.2 Future Work

**Energy-conservation and interference.** Our energy saving techniques were mainly based on TDMA for multiple access scenario. However, we expect our approach to be suitable for a broader set of multiple access technologies, e.g., beyond TDMA, FDMA to OFDMA, and extended to multiple cell scenarios. Another interesting observation is that such energy saving techniques effectively reduce the output power level of mobile terminals and this in turn might be beneficial to mitigating inter-cell interference. Thus one might expect to achieve even better energy savings, higher system capacity and/or, in the case of file transfers, to see an improved energy-delay tradeoffs.

**Coexistence of two types of users and energy-conservation.** In Chapter 3 we have deliberately partitioned users into two types: real-time sessions and file transfers. A question then arises as to whether it is possible to consider them joint optimization of energy and performance criteria. Considering the difficulty of optimizing the case of file transfers alone, this remains an open question, particularly in a dynamic system. One possible way is to use time-scale separation. Assuming the time scale of real-time sessions is much larger than that of file transfers, the state-dependent algorithm for file transfers and the optimization for real-time sessions can be done at the same time, but "separately". This requires dividing the time frame into two subframes, one for real-time and the other

for file transfers. The ratio between two subframes can be judiciously determined from the engineering point of view by service providers considering the reliability of real-time sessions and energy saving benefits of each type. One might try to find the optimal subframe ratio – we leave this as a problem for further research.

**Future work on energy saving for MIMO.** So far we have studied the energy saving technique that is applicable to the single-cell uplink scenario. One possible future research direction is devising energy efficient uplink MIMO transmissions for multi-cell scenarios. Specifically, the MT at the cell edge can transmit to only one of the BSs or transmit simultaneously to multiple BSs to exploit macro diversity. In either case, the MT also can choose one out of the three possible modes: spatial multiplexing, diversity or antenna selection as similar to what we proposed in Chapter 4. We will, therefore, design a new uplink MIMO transmit algorithm to save energy.

**Implementing load balancing.** There are several factors to consider in implementing our proposed adaptive load balancing algorithm in real systems. First, the load vector should be estimated, but estimation errors may in turn affect system performance. Though we have not addressed the impact of estimation errors on the algorithm, this is an interesting topic for future research. For example, while $\alpha = 2$ should give delay-optimal user association, it is not clear using $\alpha = 2$ would be robust against the estimation errors because $(1 - \rho_i)^2$ amplifies errors particularly when $\rho_i$ is close to 1.

Second, we considered the case where *synchronous* updates of $\rho^{(k)}$ were being made while assuming a separation of time scales between such updates and the flow-level dynamics; BSs update their utilizations periodically after they see stationary traffic distributions, and MTs then follow the new association rule. In real systems one need not enforce such synchronous behavior; the load vector, say $\rho(t)$, can be *continuously* updated in time. This

modified algorithm would work well in real systems because our original synchronous update algorithm achieves optimality when $\beta$ is close to 1, i.e., when there are no abrupt changes in $\rho^{(k)}$.

Third, we have not considered the case where users see discrete valued rates. In real systems the rates are quantized because of a finite number of modulation orders and code rates. As a consequence, there exists non-zero probability of tie-breaking, which may require appropriate tie-breaking rule.

Overall the work in this dissertation suggests that the design space of wireless systems offers a variety of practically interesting tradeoffs one can realize in designing such systems. The critical challenge however remains appropriately gauging the relative importance of various performance metrics to the application/user and to the provider. For example, a simple question such as quantifying the impact of blocking request vs offering poor performance to a population of users is critical in designing an appropriate call admission policy for best effort flows. Providers have a great concern with losing customers (churn) due to poor coverage or performance, yet the central question of how the system should be optimized (if possible) to minimize such losses cannot be resolved until the connection between user perceived performance and choice of (change of) provider behaviors are understood.

# Bibliography


[1] IEEE P802.16m-2007 draft standards for local and metropolitan area networks part 16: Air interface for fixed broadcast wireless access systems. *IEEE Standard 802.16m*, 2007.

[2] E. Altman, A. Orda, and N. Shimkin. Bandwidth allocation for guaranteed versus best effort service categories. *Queueing Systems*, 36:89–105, 2000.

[3] A. Anand, C. Manikopoulos, Q. Jones, and C. Borcea. A quantitative analysis of power consumption for location-aware applications on smart phones. In *Proc. IEEE ISIE*, June 2007.

[4] J. G. Andrews, A. Ghosh, and R. Muhamed. Fundamentals of WiMAX. *Prentice Hall*, 2007.

[5] M. Assaad and D. Zeghlache. Cross-layer design in HSDPA system to reduce the TCP effect. *IEEE Jour. Select. Areas in Comm.*, 2006.

[6] F. Baccelli and P. Bremaud. Elements of queueing theory. *Springer-Verlag*, 2004.

[7] H. Balakrishnan, V. Padmanabhan, S. Seshan, M. Stemm, and R. H. Katz. A comparison of mechanisms for improving TCP performance. *IEEE/ACM Trans. Networking.*, 5:756–769, 1997.

[8] N. Bambos. Towards power-sensitive network architectures in wireless communications. *IEEE Personal Communications*, pages 50–59, June 1998.

[9] L. A. Barroso and U. Holzle. The case for energy-proportional computing. *IEEE Computer*, 40(12):33–37, Dec. 2007.

[10] N. Benameur, S. Ben Fredj, F. Delcoigne, S. Oueslait-Boulahia, and J.W. Roberts. Integrated admission control for streaming and elastic traffic. *QofIS*, pages 69–81, 2001.

[11] N. Benameur, S. Ben Fredj, S. Oueslait-Boulahia, and J. W. Roberts. Quality of service and flow level admission control in the internet. *Computer Networks*, 40:57–71, 2002.

[12] P. Bender and A. Viterbi. CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users. *IEEE Comm. Mag.*, pages 70–77, July 2000.

[13] R. A. Berry and R. G. Gallager. Communication over fading channels with delay constraints. *IEEE Trans. Infomation Theory*, 48(5):1135–1149, May 2002.

[14] D. Bertsekas and R. Gallager. *Data Networks*. 1992.

[15] Ramya Bhagavatula, Claude Oestges, and Jr. Robert W. Heath. A new double directional channel model including antenna patterns, array orientation and depolarization. *in revision, IEEE Transactions on Vehicular Technology*, 2009.

[16] G. Bianchi and I. Tinnirello. Improving load balancing mechanisms in wireless packet networks. In *Proc. IEEE Int. Conf. on Comm. (ICC)*, pages 891–95, 2002.

[17] T. Bonald, S. Borst, and A. Proutiere. Inter-cell scheduling in wireless data networks. In *European Wireless Conference*, 2005.

[18] T. Bonald and A. Proutiere. Insensitive bandwidth sharing in data networks. *Queueing systems*, 44:69–100, 2003.

[19] T. Bonald and A. Proutiere. On performance bounds for the integration of elastic and adaptive streaming flows. *ACM SIGMETRICS*, pages 235–245, 2004.

[20] T. Bonald and J. W. Roberts. Congestion at flow level and the impact of user behaviour. *Computer Networks*, pages 521–536, 2003.

[21] S. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. *Proc. IEEE INFOCOM*, 1:321–331, Apr. 2003.

[22] S. Borst, I. Saniee, and A. Whiting. Distributed dynamic load balancing in wireless networks. In *ITC*, pages 1024–37, 2007.

[23] B. Bougard, G. Lenoir, A. Dejonghe, L. V. der Perre, F. Catthoor, and W. Dehaene. SmartMIMO: an energy-aware adaptive MIMO-OFDM radio link control for next-generation wireless local area networks. *EURASIP Journal on Wireless Communications and Networking*, 2007:1–15, 2007.

[24] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[25] T. Bu, L. Li, and R. Ramjee. Generalized proportional fair scheduling in third generation wireless data networks. In *Proc. IEEE INFOCOM*, 2006.

[26] S. Catreux, V. Erceg, D. Gesbert, and Jr R. W. Heath. Adaptive modulation and MIMO coding for broadband wireless datanetworks. *IEEE Comm. Mag.*, 40:108–115, Jun. 2002.

[27] C.-B. Chae, A. Forenza, R. W. Heath, Jr., M. R. McKay, and I. B. Collings. Adaptive MIMO transmission techniques for broadband wireless communication systems. *submitted to IEEE Comm. Mag.*, March 2007.

[28] C.-B. Chae, M. Katz, C. Suh, and H. Jeong. Adaptive spatial modulation for MIMO-OFDM. *Proc. IEEE WCNC*, pages 87–92, March 2004.

[29] N. M. Chaskar, T. V. Lakshman, and U. Madhow. TCP over wireless with link level error control: Analysis and design methodology. *IEEE/ACM Trans. Networking.*, 7:605–615, 1999.

[30] B. E. Collins and R. L. Cruz. Transmission policies for time varying channels with average delay constraints. In *Proc. Allerton Conf. on Comm. Control and Comp.*, 1999.

[31] R. M. Corless, G. H. Gonnet, D. E. G. Hare, and D. E. Knuth D. J. Jeffrey. On the Lambert W function. *Advances in Computational Mathematics*, 5:329–359, 1996.

[32] S. Cui, A. J. Goldsmith, and A. Bahai. Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks. *IEEE Jour. Select. Areas in Comm.*, 22:1089–1098, Aug. 2004.

[33] S. Cui, A. J. Goldsmith, and A. Bahai. Joint modulation and multiple access optimization under energy constraints. In *Proc. IEEE Glob. Telecom. Conf.*, pages 151–155, 2004.

[34] S. Cui, A. J. Goldsmith, and A. Bahai. Energy-constrained modulation optimization. *IEEE Trans. Wireless Commun.*, 4:2349–32360, Sep. 2005.

[35] S. Cui, R. Madan, A. J. Goldsmith, and S. Lall. Cross-layer energy and delay optimization in small-scale sensor networks. *IEEE Trans. Wireless Commun.*, 6:3688–3699, Oct. 2007.

[36] S Das, H Viswanathan, and G Rittenhouse. Dynamic load balancing through coordinated scheduling in packet data systems. In *Proc. IEEE INFOCOM*, pages 786–96, 2003.

[37] G. de Veciana, T.-J. Lee, and T. Konstantopoulos. Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Trans. Networking.*, 9:2–14, 2001.

[38] F. Delcoigne, A. Proutiere, and G. Regnie. Modeling integration of streaming and data traffic. *Performance Evaluation*, 55, 2004.

[39] A. El Gamal, C. Nair, B. Prabhakar, E. Uysal-Biyikoglu, and S. Zahedi. Energy-efficient scheduling of packet transmissions over wireless networks. In *Proc. IEEE INFOCOM*, volume 3, pages 1773–1782, 2002.

[40] 3GPP Long Term Evolution. Physical layer aspects of UTRA high speed downlink packet access. *Technical Report TR25.814*, 2006.

[41] G. Fayolle, V. A. Malyshev, and M. V. Menshikov. Topics in the constructive theory of countable markov chains. *Cambridge University Press*, 1985.

[42] M. Fiorenzi, D. Girella, N. Moller, A. Arvidssony, R. Skogy, J. Peterssony, P. Karlsson, C. Fischione, and K. H. Johansson. Enhancing TCP over HSDPA by cross-layer signalling. In *Proc. IEEE Glob. Telecom. Conf.*, 2007.

[43] A. Forenza, M. R. McKay, A. Pandharipande, R. W. Heath, Jr., and I. B. Collings. Adaptive MIMO transmission for exploiting the capacity of spatially correlated channels. *IEEE Trans. on Veh. Technol.*, 56:619–630, Mar. 2007.

[44] G. J. Foschini and Z. Miljanic. A simple distributed autonomous power control algorithm and its convergence. *IEEE Trans. on Veh. Technol.*, 42(4):641–646, Nov. 1993.

[45] S. B. Fredj, T. Bonald, A. Proutiere, G. Regnie, and J. W. Roberts. Statistical bandwidth sharing: a study of congestion at flow level. In *ACM SIGCOMM*, 2001.

[46] D. Gesbert, M. Shafi, D. Shiu, P. J. Smith, and A. Naguib. From theory to practice: an overview of MIMO space-time coded wireless systems. *IEEE Jour. Select. Areas in Comm.*, 21:281–302, Apr. 2003.

[47] A. J. Goldsmith. *Wireless Communications*. Cambridge, 2005.

[48] A. J. Goldsmith and P. P. Varaiya. Capacity of fading channels with channel side information. *IEEE Trans. Infomation Theory*, 43(6):1986–1992, Nov. 1997.

[49] Goliath. Financial assessment of citywide wi-fi/wimax deployment. *http: //goliath.ecnext.com/coms2/summary_0199-6709510_ITM*, 2006.

[50] P. Hande, S. Patil, and H. Myung. Distributed load-balancing in a multi-carrier system. In *Proc. IEEE Wireless Comm. and Net. Conf.*, 2009.

[51] R. W. Heath, Jr. and A. Paulraj. Switching between spatial multiplexing and transmit diversity based on constellation distance. *IEEE Trans. Communications*, 53:962–968, June 2005.

[52] N. Himayat, M. Venkatachalam, A. Koc, S. Talwar, H. Yin, S. Ahmadi, M. Ho (Intel Corporation), G. Miao, Ye (Geoffrey) Li, and H. Kim. Improving client energy consumption in 802.16m. *IEEE 802.16m Standard Contribution IEEE C802.16m-09/0107*, January 2009.

[53] A. Jalali, R. Padovani, and R. Pankaj. Data throughput of CDMA-HDR a high efficient-high data rate personal communication wireless system. In *Proc. IEEE Veh. Technol. Conf.*, pages 1854–1858, May 2000.

[54] P. Key, L. Massoulie, A. Bain, and F. Kelly. Fair internet traffic integration: network flow models and analysis. *Annals of Telecommunications*, 59, 2004.

[55] H. Kim. Experience from broadband rollout and future steps in KT. *Telektronikk*, pages 123–131, 2/3.2002. Invited paper.

[56] H. Kim, C-B. Chae, G. de Veciana, and R. W. Heath Jr. A cross-layer approach to energy efficiency for adaptive MIMO systems exploiting spare capacity. *IEEE Trans. Wireless Communications*.

[57] H. Kim, C-B. Chae, G. de Veciana, and R. W. Heath Jr. Energy-efficient adaptive MIMO systems leveraging dynamic spare capacity. In *Proc. Conference on Information Sciences and Systems (CISS)*, 2008.

[58] H. Kim and G. de Veciana. Leveraging dynamic spare capacity in wireless systems to conserve mobile terminals' energy. *to appear in IEEE/ACM Trans. Networking*.

[59] H. Kim and G. de Veciana. Losing opportunism: evaluating service integration in an opportunistic wireless system. In *Proc. IEEE INFOCOM*, 2007.

[60] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam. $\alpha$-optimal user association and cell load balancing in wireless networks. *submitted to IEEE INFOCOM*, 2009.

[61] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam. $\alpha$-optimal user association and cell load balancing in wireless networks. *Technical Report, UT-Austin, available at http: //users.ece.utexas.edu/ hkim4/index_files/alpha_optimal2009.pdf*, 2009.

[62] H. Kim, X. Yang, M. Venkatachalam, Y.-S. Chen, K. Chou, I.-K. Fu, and P. Cheng. Handover and load balancing rules for 16m. In *IEEE C802.16m-09/0136r1*, pages 1024–37, Jan. 2009.

[63] R. Knopp and P.A. Humblet. Information capacity and power control in single-cell multiuser communications. In *Proc. IEEE Int. Conf. on Comm. (ICC)*, pages 331–335, 1995.

[64] M. Kodialam, T. V. Lakshman, and S. Sengupta. Traffic-oblivious routing for guaranteed bandwidth performance. *IEEE Comm. Mag.*, pages 46–51, Apr. 2007.

[65] A. Kumar, D. Manjunath, and J. Kuri. *Communication Networking.* Elsevier, 2003.

[66] T-J. Lee and G. de Veciana. Model and performance evaluation for multiservice network link supporting ABR and CBR services. *IEEE Commun. Letters*, 4, 2000.

[67] Yong-Kyung Lee and Dongmyun Lee. Broadband access in korea: Experience and future perspective. *IEEE Comm. Mag.*, pages 30–36, Dec. 2003.

[68] K-K. Leung and C. W. Sung. An opportunistic power control algorithm for cellular network. *IEEE/ACM Trans. Networking.*, 14(3):470–478, Jun. 2006.

[69] S. Liu and J. Virtamo. Inter-cell coordination with inhomogeneous traffic distribution. In *Next Generation Internet Design and Engineering Conference*, 2006.

[70] X. Liu, E. K. P. Chong, and N. B. Shroff. Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. *IEEE Jour. Select. Areas in Comm.*, 19, Oct. 2001.

[71] X. Liu, E. K. P. Chong, and N. B. Shroff. A framework for opportunistic scheduling in wireless networks. *Computer Networks*, 41, 2003.

[72] R. Madan, S. Cui, S. Lall, and A. J. Goldsmith. Cross-layer design for lifetime maximization in interference-limited wireless sensor networks. *IEEE Trans. Wireless Commun.*, 5(11):3142–3152, Nov. 2006.

[73] R. Madan, S. Cui, S. Lall, and A. J. Goldsmith. Modeling and optimization of transmission schemes in energy-constrained wireless sensor networks. *IEEE/ACM Trans. Networking.*, 15(6):1359–1372, 2007.

[74] D. Maksimovic. Power management model and implementation of power management ICs for next generation wireless applications. In *Proc. IEEE Int. Symposium on Circuits and Systems*, May 2002.

[75] C. Martin and B. Ottersten. Asymptotic eigenvalue distributions and capacity for MIMO channels under correlated fading. *IEEE Trans. Wireless Commun.*, 3:1350–1359, Jul. 2004.

[76] P. Youssef Massaad, M. Medard, and L. Zheng. Impact of processing energy on the capacity of wireless channels. In *Int. Symp. on Info. Theory and its App.*, Oct. 2004.

[77] L. Massoulie and J.W. Roberts. Arguments in favour of admission control for TCP flows. *Proc. ITC 16*, 1999.

[78] L. Massoulie and J.W. Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15:185–201, 2000.

[79] M. R. McKay, I. B. Collings, A. Forenza, and R. W. Heath, Jr. Multiplexing / beamforming switching for coded-MIMO in spatially-correlated rayleigh channels. *IEEE Trans. on Veh. Technol.*, 56:2555–2567, Sep. 2007.

[80] G. Miao, N. Himayat, Y. Li, and D. Bormann. Energy efficient design in wireless OFDMA. In *Proc. IEEE Int. Conf. on Comm. (ICC)*, May 2008.

[81] G. Miao, N. Himayat, Y. Li, and A. Swami. Energy-efficient link adaptation in frequency-selective channels. *to appear in IEEE Transactions on Communications.*

[82] G. Miao, N. Himayat, Y. Li, and A. Swami. Cross-layer optimization for energy-efficient wireless communications: A survey. *Wiley Journal Wireless Commun. and Mobile Computing*, 9, April 2009.

[83] D. S. Michalopoulos, A. S. Lioumpas, and G. K. Karagiannidis. Increasing power efficiency in transmitter diversity systems under error performance constraints. *IEEE Trans. Communications*, 56:2025–2029, Dec. 2008.

[84] K. Navaie and H. Yanikomeroglu. Downlink joint base-station assignment and packet scheduling for cellular CDMA/TDMA networks. In *Proc. IEEE Int. Conf. on Comm. (ICC)*, pages 4339–44, 2006.

[85] M. F. Neuts. Matrix-geometric solutions in stochastic models. *The Johns Hopkins University Press*, 1981.

[86] C. Oliveira, J. B. Kim, and T. Suda. An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks. *IEEE Jour. Select. Areas in Comm.*, 16(6):858–874, Aug. 1998.

[87] D. Park, H. Kwon H. Seo, and B. G. Lee. Wireless packet scheduling based on the cumulative distribution function of user transmission rates. *IEEE Trans. Communications*, 53:1919–29, Nov. 2005.

[88] S. Patil and G. de Veciana. Feedback and opportunistic scheduling in wireless networks. *IEEE Trans. Wireless Commun.*, pages 4227–38, Dec. 2007.

[89] S. Patil and G. de Veciana. Managing resource and quality of service in heterogeneous wireless systems exploiting opportunism. *IEEE/ACM Trans. Networking.*, 15:1046–58, Oct. 2007.

[90] S. Patil and G. de Veciana. Measurement-based opportunistic scheduling for heterogeneous wireless systems. *IEEE Trans. on Communications, to appear*, 2009.

[91] A. Paulraj, R. Nabar, and D. Gore. *Introduction to Space-Time Wireless Communications*. Cambridge University Press, 2003.

[92] S. Pollin, B Bougard, R. Mangharam, F. Catthoor, I. Moerman, R. Rajkumar, and L. V. der Perre. Optimizing transmission and shutdown for energy-efficient real-time packet scheduling in clustered ad hoc networks. *EURASIP Jour. on Wireless Commun. and Networking*, 2005.

[93] S. Pollin, R. Mangharam, B Bougard, L. V. der Perre, I. Moerman, R. Rajkumar, and F. Catthoor. MEERA: Cross-layer methology for energy efficient resource allocation in wireless networks. *IEEE Trans. Wireless Commun.*, 6:617–628, Feb. 2007.

[94] WiMAX power amplifier ADL5570 and 5571. http://www.analog.com/uploadedfiless /data_sheets/adl5570.pdf and adl5571.pdf. *Analog Device*, Sep. 2007.

[95] B. Prabhakar, E.U. Biyikoglu, and A. E Gamal. Energy-efficient transmission over a wireless link via lazy packet scheduling. *Proc. IEEE INFOCOM*, 1:386–393, 2001.

[96] J. G Proakis. *Digital Communications, 3rd Edition*. 1995.

[97] D. Rajan, A. Sabharwal, and B. Aazhang. Delay-bounded packet scheduling of bursty traffic over wireless channels. *IEEE Trans. Infomation Theory*, 50:125–144, 2004.

[98] B. Rengarajan and G. de Veciana. Architecture and abstraction for environment and traffic aware system-level coordination of wireless networks: the downlink case. *Proc. IEEE INFOCOM*, pages 1175–1183, 2008.

[99] B. Rengarajan and G. de Veciana. Architecture and abstractions for environment and traffic aware system-level coordination of wireless systems. *submitted to IEEE/ACM Trans. Networking*, June 2009.

[100] B. Rengarajan and G. de Veciana. Load balancing is not optimal in wireless systems with dynamic intereference. *International Workshop on Network Control and Optimization, Submitted*, 2009.

[101] D.G. Sachs, I. Kozintsev, M. Yeung, and D.L. Jones. Hybrid ARQ for robust video streaming over wireless LANs. In *International Conference on Information Technology: Coding and Computing*, 2001.

[102] A. Sang, M. Madihian, X. Wang, and R. D. Gitlin. Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems. In *ACM Mobicom*, Sep. 2004.

[103] C. Schurgers, O. Aberthorne, and M. Srivastava. Modulation scaling for energy aware communication systems. *International Symposium on Low Power Electronics and Design*, pages 96–99, 2001.

[104] C. Schurgers, V. Raghunathan, and M. B. Srivastava. Power management for energy-aware communication systems. *ACM Trans. Embedded Computing Systems*, 2:431–447, Aug. 2003.

[105] S. Shakkottai. Effective capacity and QoS for wireless scheduling. *submitted for journal publication*, 2004.

[106] S. Shakkottai and A. Stolyar. Scheduling for multiple flows sharing a time-varying channel: the exponential rule. *American Mathematical Society Translations, Series 2*, 207, 2002.

[107] S. Shakkottai and A. Stolyar. Effective capacity and QoS for wireless scheduling. *IEEE Trans. Automatic Control*, Feb. 2008.

[108] K. Son, S. Chong, and G. de Veciana. Dynamic association for load balancing and interference avoidance in multi-cell networks. *IEEE Trans. Wireless Commun.*, to appear.

[109] E. Uysal-Biyikoglu, A. El Gamal, and B. Prabhakar. Adaptive transmission of variable-rate data over a fading channel for energy-efficiency. In *Proc. IEEE Glob. Telecom. Conf.*, volume 1, pages 97–101, 2002.

[110] E. Uysal-Biyikoglu and A. El Gamal. On adaptive transmission for energy efficiency in wireless data networks. *IEEE Trans. Infomation Theory*, 50(12):3081–3094, Dec. 2004.

[111] E. Uysal-Biyikoglu, B. Prabhakar, and A. El Gamal. Energy-efficient packet transmission over a wireless link. *IEEE/ACM Trans. Networking.*, 10(4):487–499, Aug. 2002.

[112] H. Velayos, V. Aleo, and G. Karlsson. Load balancing in overlapping wireless LAN cells. In *Proc. IEEE Int. Conf. on Comm. (ICC)*, pages 3833–36, Jun. 2004.

[113] P. Viswanath, V. Anantharam, and D. Tse. Optimal sequences, power control and capacity of synchronous CDMA systems with linear MMSE multiuser receivers. *IEEE Trans. Infomation Theory*, 45, Sept. 1999.

[114] J. Walrand. *An introduction to queueing networks.* Prentice Hall, 1998.

[115] A. Y. Wang, S. Chao, C. G. Sodini, and A. P. Chandrakasan. Energy efficient modulation and MAC for asymmetric RF microsensor system. In *Int. Symp. Low power electronics and Design*, pages 106–111, 2001.

[116] H. Wang and N. B. Mandayam. Opportunistic file transfer over a fading channel under energy and delay constraints. *IEEE Trans. Communications*, 53(4):632–644, Apr. 2005.

[117] C.-K. Wen, K.-K. Wong, P. Ting, and C.-L. I. Optimal power-saving input covariance for MIMO wireless systems exploiting only channel spatial correlations. In *Proc. IEEE Int. Conf. on Comm. (ICC)*, pages 2011–2015, 2005.

[118] M. Xiao, N. B. Shroff, and E. K. P. Chong. A utility-based power-control scheme in wireless cellular systems. *IEEE/ACM Trans. Networking.*, 11(2):210–221, Apr. 2003.

[119] S.-C. Yang and G. de Veciana. Size-based adaptive bandwidth allocation: Optimizing the average QoS for elastic flows. In *Proc. IEEE INFOCOM*, 2002.

[120] Y. Yao and G. B. Gianakis. Energy-efficient scheduling for wireless sensor networks. *IEEE Trans. Communications*, 53(8):1333–1342, Aug. 2005.

[121] R. Yates. A framework for uplink power control in cellular radio systems. *IEEE Jour. Select. Areas in Comm.*, 13:1341–1347, 1995.

[122] R. Yates and C.Y. Huang. Integrated power control and base station assignment. *IEEE Trans. on Veh. Technol.*, 44, 1995.

[123] Y. Yu, B. Krishnamachari, and V. K. Prasanna. Energy-latency tradeoffs for data gathering in wireless sensor networks. In *Proc. IEEE INFOCOM*, 2004.

[124] M. Zafer and E. Modiano. A calculus approach to energy-efficient data transmission with quality-of-service constraints. *IEEE/ACM Transactions on Networking, submitted*, 2007.

[125] J. Zander. Distributed cochannel interference control in celluar radio systems. *IEEE Trans. on Veh. Technol.*, 41, 1992.

[126] J. Zander. Performance of optimum transmitter power control in celluar radio systems. *IEEE Trans. on Veh. Technol.*, 41, 1992.

[127] A. Zemlianov and G. de Veciana. Load balancing in wireless systems supporting end nodes with dual mode capabilities. In *Proc. Conference on Information Sciences and Systems (CISS)*, Mar. 2005.

# Vita

Hongseok Kim received his B.S. and M.S. degree in Electrical Engineering from Seoul National University in 1998 and 2000, respectively. From 2000 to 2005 he was a member of technical staff at Korea Telecom Network Laboratory. He participated in FSAN and ITU-T SG16 FS-VDSL standardization from 2000 to 2003. In 2005 he entered the University of Texas at Austin and advanced to Ph.D. candidacy in 2007. He was a student intern at Intel Corporation during June to December 2008, and Qualcomm Flarion Technology during June to August 2009. He was a visiting scholar at Princeton University from September to October 2009 and has become a post doctoral researcher since November 2009. He was the recipient of Korea Government Overseas Scholarship in 2005.

*In his heart a man plans his course, but the LORD determines his steps (Proverb 16:9).*

Permanent address: 3487 Lake Austin Blvd. Apt D.
Austin, Texas 78703

This dissertation was typeset with LaTeX[†] by Hongseok Kim.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.