# Leveraging Dynamic Spare Capacity in Wireless Systems to Conserve Mobile Terminals' Energy

Hongseok Kim, *Student Member*, *IEEE* and Gustavo de Veciana, *Senior Member*, *IEEE*

***Abstract*—In this paper we study several ways in which mobile terminals can backoff on their uplink transmit power in order to extend battery lifetimes. This is particularly effective when a wireless system is underloaded as the degradation in user's perceived quality of service 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. The energy-optimal transmission strategy for real-time sessions is determined by solving a convex optimization. An iterative approach exhibiting superlinear convergence achieves substantial amount energy savings, e.g., more than 50% when the session blocking probability is 0.1% or less. The case of file transfers is more subtle because power backoff changes the system dynamics. We study energy-efficient transmission strategies that realize energy-delay tradeoff. The proposed mechanism achieves a 35–75% in 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.**

***Index Terms*—Energy-efficiency, flow-level dynamics, idling/ circuit power, wireless systems.**

## I. Introduction

Wireless cellular systems such as WiMAX are evolving to support mobile broadband services [1]. 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) [2]–[8], 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 [4]). Other energy consumption such as display or microprocessor, etc., are not considered in this paper.

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% [9]. Similarly a substantial fraction of Wi-Fi hotspot capacity is unused [10]. 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 [1], [11]. 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 [12]. This further contributes to underutilization of the system, even when the loads are heavy. The central premise of this paper 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

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

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

Manuscript submitted May 12, 2008. The authors are with the Wireless Networking and Communications Group (WNCG), The University of Texas at Austin, TX 78712 USA (e-mail:hkim4@ece.utexas.edu; gustavo@ece.utexas.edu). This work was supported in part by a gift from the Intel Research Council and NSF Award CNS-0721532.

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 paper, 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 sometimes referred to as flow-level dynamics [13], see Fig. 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] [14], [15]. 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 energy 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* [5], [15].

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 [16]–[18]. Energy-efficient power control was first explored in the context of sensor networks [2], [3]. 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 [19], [20], [4] 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 [21]–[24]; i.e., when the channel is good, transmit power is increased. However, the above work neglects circuit power, idling power and flow-level dynamics.

[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 [14].

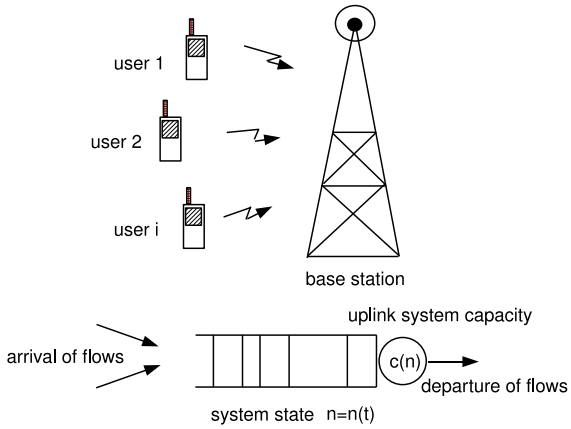



Fig. 1. Flow-level model for uplink transmission in a dynamic system. One user corresponds to one flow.

Recent results show that if circuit power is taken into account, circuit energy consumption increases monotonically as the transmission time grows [5], [7], [25], [26]. 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 [5] focuses on the physical modulation techniques with a single sender and receiver pair for sensor 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 [6] and in wireless LANs [7], and further up-to the routing layer [8]. Energy-efficient transmission strategy for orthogonal frequency division multiple access (OFDMA) system considering circuit power was proposed in [27]. 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 [15].

***Contributions.*** We highlight the contributions of this paper 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.*** The paper is organized as follows. In Section II, we describe our system model and assumptions. Section III 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 IV and conclude the paper with Section V.

## II. System model

### A. 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 [15]. 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. 1. However, our framework is also applicable to downlink transmissions. 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.

### B. 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. We will separately consider the case where a flow corresponds to real-time session or a file transfer, in Section III and Section IV respectively. The system dynamics are captured by a *flow-level queueing model* shown in Fig. 1 which tracks the arrival and departure process of users (or flows), see e.g., [13]. 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.

### C. Minimizing energy consumption in a stationary system

Our objective is to minimize the energy consumption of a *typical*[2] 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 $T$. Letting $J$ denote the energy consumption of a typical flow, our goal will be to minimize

$$E[J] = E\left[\int_0^T F(s)ds\right] \tag{2}$$

subject to either sustaining minimum rate requirements for real-time sessions or achieving an average throughput for file transfers. Minimizing (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 1 (**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,

$$E[P] = \lambda E[J]. \tag{3}$$

*Proof:* This result is intuitive and can be shown via Brumelle's theorem [28], which is a generalized version of Little's law. ■

Based on Theorem 1, we below focus on minimizing the average system power consumption which in turn minimizes the average energy consumed by a typical mobile terminal.

### D. 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 of the RF power amplifier and associated circuit blocks [5], [15]. It also depends on classes of power amplifiers, modulation schemes and power-saving mechanisms [29]. 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 [5]. Then, the power equation $f(x)$ at transmission rate $x$ can be derived from (1) to give

$$f(x) \quad = \quad \begin{cases} \left(\exp(\frac{x}{w}) - 1\right)\frac{\sigma^2}{\eta g} + \alpha & (\text{active},\ x > 0) \\ \beta & (\text{idling},\ x = 0), \end{cases} \tag{4}$$

where $\eta$ is the drain efficiency, which is defined as the ratio of the output power and the power consumed in the power amplifier; $\alpha$ is the circuit power; and $\beta$ is the idling power [5], [15]. To simplify our notation, we let $\gamma = \frac{\eta g}{\sigma^2}$, i.e., the

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

TABLE I
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 |
| $\alpha$ | circuit power |
| $\beta$ | idling power |
| $\xi :=$ | $\alpha - \beta$ |
| $\lambda$ | 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 |

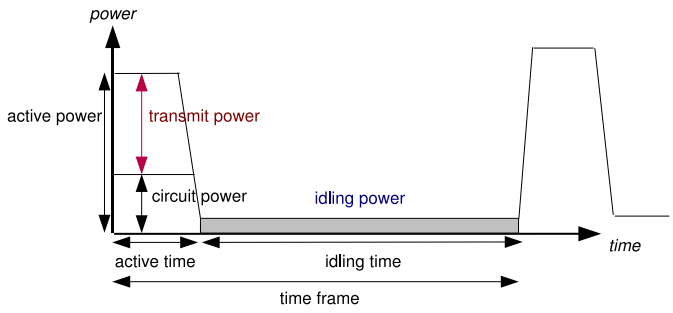



Fig. 2. Transmission power model in TDMA systems.

signal-to-noise ratio (SNR) when the transmit power, defined below, is 1. We summarize our terminologies as follows.

*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 2.

*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) Circuit power $\alpha$:* The circuit power $\alpha$ includes several circuit blocks in the transmission chain and remains almost constant irrespective of the transmission rate $x$. It is modeled in [5], [15], by $\alpha = p_{\text{dac}} + p_{\text{mix}} + p_{\text{filt}} + p_{\text{syn}}$, where $p_{\text{dac}}$, $p_{\text{mix}}$, $p_{\text{filt}}$, $p_{\text{syn}}$ stand for the power consumption from a digital-to-analog converter, a mixer, a filter, a frequency synthesizer, respectively.

*4) Idling power $\beta$:* 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. 2, idling users turn off their transmission circuits and power amplifier to save energy, but they still consume idling power $\beta$, ranging from a few to tens of mW, due to leakage currents [14]. Even though $\beta$ could be negligible in a static system, it remains non-negligible in a dynamic system [15]. We will see the impact of idling power, particularly for the case of file transfers in Section IV. Power-related parameters and notations are summarized in Table II and Table I.

TABLE II
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 |
| $\alpha$ = 115.9 mW | $N_0$ = $-174$ dBm |
| $\beta$ = 25 mW | $\mu^{-1}$ = 60 kbytes |

## III. 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.

### A. 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 $\mu_s^{-1}$. 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.,

$$\sum_{i \in \mathcal{A} \cup \{k\}} \frac{r_i}{c_i} \leq 1 \tag{5}$$

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. The stationary distribution for the number of users $\pi_s(n)$ is then given by

$$\pi_s(n) = \pi_s(0)\frac{\rho_s^n}{n!} \tag{6}$$

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 $\pi_s(n^*)$ [30].

From Theorem 1, 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*.[3] Thus, the problem reduces to one

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

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

$$\begin{aligned}\min_{\mathbf{x}\succ 0} \quad & \sum_{i\in\mathcal{A}} \frac{r_i}{x_i}\left(\frac{\exp(\frac{x_i}{w})-1}{\gamma_i}+\alpha_i+\sum_{j\in\mathcal{A}\setminus\{i\}}\beta_j\right) \\ & +\left(1-\sum_{i\in\mathcal{A}}\frac{r_i}{x_i}\right)\sum_{i\in\mathcal{A}}\beta_i \\ s.t. \quad & \sum_{i\in\mathcal{A}}\frac{r_i}{x_i}\le 1, \end{aligned} \tag{7}$$

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 put the subscript $i$ for $\alpha$, $\beta$, $g$ 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 $(\exp(x_i/w)-1)/\gamma_i+\alpha_i+\sum_{j\in\mathcal{A}\setminus\{i\}}\beta_j$. 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 $\sum_{i\in\mathcal{A}}\beta_i$.

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

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

$$\min_{\mathbf{x}\succ 0} \quad \sum_{i\in\mathcal{A}}\frac{r_i}{x_i}\left(\frac{\exp(\frac{x_i}{w})-1}{\gamma_i}+\xi_i\right) \tag{8}$$

$$s.t. \quad \sum_{i\in\mathcal{A}}\frac{r_i}{x_i}\le 1, \tag{9}$$

where $\xi_i := \alpha_i - \beta_i$. 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 $\xi_i \ge 0$. When $\xi_i \le 0$, Problem $\mathbb{O}1$ can be further simplified to have an equality constraint and $\xi_i$ can be dropped.

### B. 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 [31], i.e., for all $i \in A$

$$\frac{\partial L}{\partial x_i^*}=0 \text{ and } \kappa^*\left(\sum_{i\in\mathcal{A}}\frac{r_i}{x_i^*}-1\right)=0 \tag{10}$$

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

$$\kappa^*=\frac{1}{\gamma_i}\left(\exp\left(\frac{x_i^*}{w}\right)\left(\frac{x_i^*}{w}-1\right)+1\right)-\xi_i,\ \forall i\in\mathcal{A}. \tag{11}$$

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 (11). Unlike the previous work which approximated the solution assuming high transmission rate [32] 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 [34]

$$W(z)e^{W(z)}=z, \tag{12}$$

and a concave, monotone increasing and differentiable function. We assume that mobile terminals have tabulated or can compute $W(z)$. The solution to (11) is then given by

$$x_i^*=\left(W\left(\frac{(\kappa^*+\xi_i)\gamma_i-1}{e}\right)+1\right)w,\ i\in\mathcal{A} \tag{13}$$

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

$$p_i^*=\left(\exp\left(W\left(\frac{(\kappa^*+\xi_i)\gamma_i-1}{e}\right)+1\right)-1\right)\frac{\sigma^2}{g_i}. \tag{14}$$

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

*Example 1 (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 2 (Homogeneous Case 2):* Suppose still that $\gamma_i = \gamma$, but now that $\xi_i = \xi > 0$. In this case (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.

### C. 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

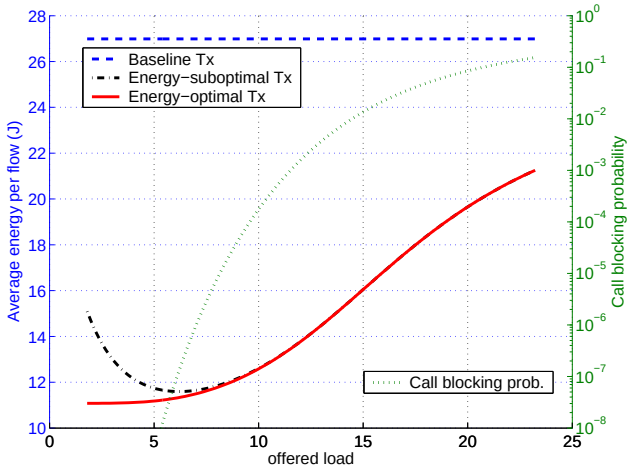



Fig. 3. Energy saving for real-time sessions under various loads. $r_i$ = 150 kbps for all users, $n^* = 23$, $c_{\max} = 3.49$ Mbps, $\mu_s^{-1} = 180$ sec, received SNR with full power transmit = 15 dB, other parameters are shown in Table II.

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 (13). Let $p(n)$ denote the system power consumption in state $n$; it is given by

$$p(n) = \frac{nr}{x}\left(f(x) + (n-1)\beta\right) + \left(1 - \frac{nr}{x}\right) n\beta. \qquad (15)$$

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 (6). From Theorem 1 and considering the call blocking probability $\pi_s(n^*)$, the average per-flow energy is given by

$$E[J] = \frac{E[P]}{\lambda_s(1 - \pi_s(n^*))}. \qquad (16)$$

Representative results for the three policies are shown in Fig. 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 3.1:* The second policy $x = nr$ (dash-dot line) exhibits an interesting behavior in Fig. 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 2 demonstrated this effect in a static system; here we see the analogous effect for the *dynamic* system.

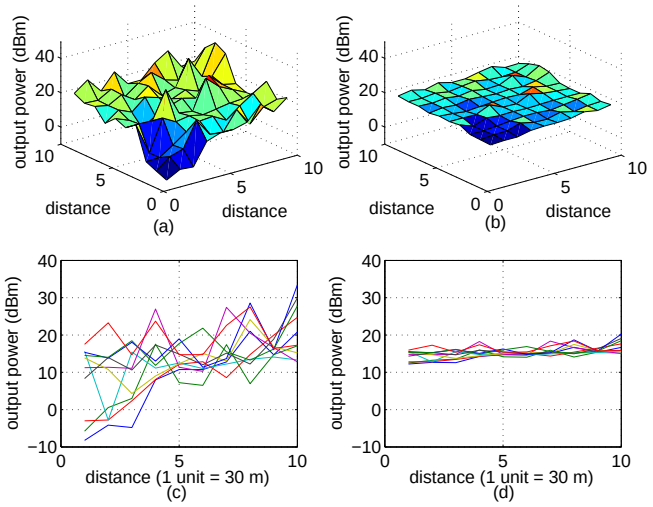



Fig. 4. Spatial power smoothing, (a) Equal time fraction allocation (b) Optimized rate and time fraction (c) Side view of (a), (d) Side view of (b): $r_i$ = 50 kbps, path loss exponent = 3, cell radius = 300 m, 100 users, carrier frequency = 1 GHz, other parameters are shown in Table II.

### D. Spatial power smoothing

A further gain of our energy-optimal transmission policy is that the output power levels 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. 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 base station. Fig. 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. 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 in this paper, reduced output power at the cell boundary suggests that our energy-saving mechanism could contribute to reducing inter-cell interference in multiple cell scenario.

## IV. 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

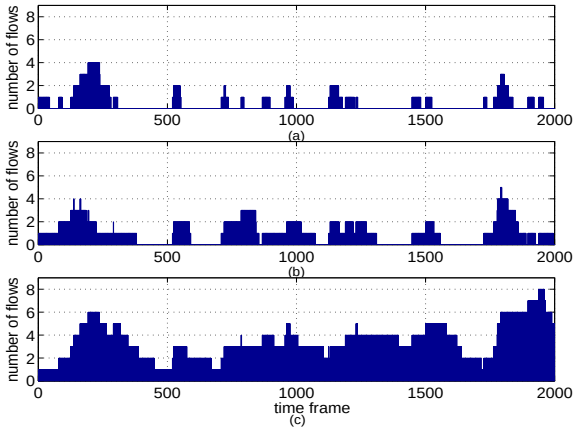



Fig. 5. Time varying number of users in a dynamic system with offered load 30%. Individual target throughput is (a) 5.10 Mbps (b) 1,275 kbps (c) 318 kbps, and the arrival processes are identical. Simulation setup is given in Section IV-H.

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.

### A. 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.

*Example 3 (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.[4]) 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 4 (Short-term time scale):* Fig. 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. 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. 5 (b) and (c) when such a strategy is used the system utilization increases, yet energy may be conserved.

[4]System capacity in this paper is not the same notion as the information theoretic capacity.

### B. Problem formulation

Let us go back to the system model shown in Fig. 1 to formulate the problem in a dynamic system. Our objective is to minimize $E[J]$, see (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

$$c(n) = c_{\max}, \tag{17}$$

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

$$c(n) = \min(nq, c_{\max}), \tag{18}$$

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

***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 [35] would

$$c(n) = E[\max(R_1, \cdots, R_n)] \tag{19}$$

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.

### C. 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 $\mu^{-1}$. 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

$$\pi(n) = \pi(0) \frac{\rho^n}{\Pi_{m=1}^n c(m)}, \tag{20}$$

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. 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

$$p(n) = f(c(n)) + (n-1)\beta, \tag{21}$$

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

$$E[J] = \frac{1}{\lambda} \sum_{n=1}^{\infty} \Big( f(c(n)) + (n-1)\beta \Big) \pi(n). \tag{22}$$

Note that $\pi(n)$ depends on the system capacity $c(n)$, i.e., these are *coupled* together, see (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.

### D. Energy-delay tradeoffs: Numerical results

Next, we investigate how changing the tuning parameter $q$ in (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. 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 ◯ in Fig. 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

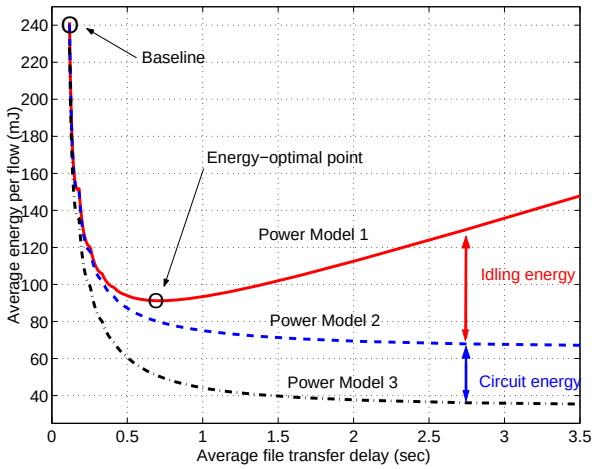



Fig. 6. Energy-delay tradeoff for various throughput $q$. ($\lambda$ = 3.65/sec, $c_{\max}$ = 5.84 Mbps, offered load = 30%, received SNR with maximum rate transmission = 17.5 dB. Model 1: $\alpha$ = 115.9 mW, $\beta$ = 25 mW, Model 2: $\alpha$ = 115.9 mW, and $\beta$ = 0 mW, Model 3: $\alpha = \beta = 0$ mW. Other parameters are given in Table II.)

$q$ 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. 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 $\frac{\alpha}{\lambda}$.

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

*Theorem 2 (**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 $\frac{\alpha}{\lambda}$ is decreasing as $\lambda$ grows, and thus the circuit energy becomes asymptotically negligible as the load grows. ■

Although Lemma 1 and Theorem 2 are simple, they demonstrate a key difference between static and dynamic systems. Here are two supporting examples.

*Example 5:* 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. 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 $\frac{\alpha}{\lambda}$. 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 [5], [7], [25], [26].

*Example 6:* To have an insight on diminishing impact of

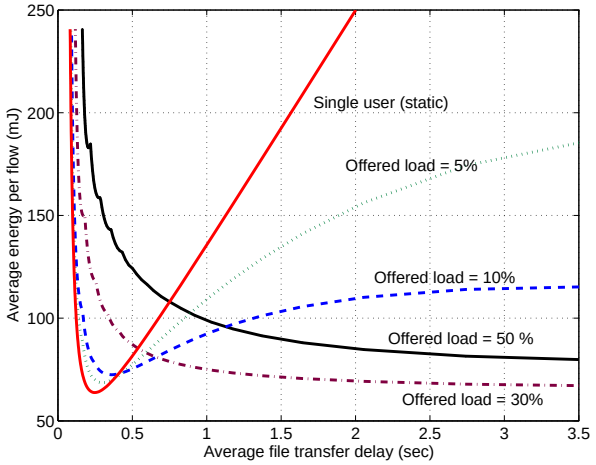



Fig. 7. The weak impact of circuit power in energy-delay tradeoff: $\alpha = 115.9$ mW, $\beta = 0$ mW, $c_{\max}$ = 5.84 Mbps, received SNR with maximum rate transmission = 17.5 dB. Other parameters are given in Table II.

circuit energy, we plot the energy consumption for Model 2 for various offered loads. In Fig. 7, we exhibit the energy and delay in the case of single user; the energy increases linearly when the delay is large (and the slope becomes identical to circuit power $\alpha$). 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.

### E. 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 (22) can be simplified using the approximation $c(n) \approx nq$. The queue's stationary distribution $\pi(n)$ in (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.,

$$\phi(\lambda,q) := \sum_{n=1}^{\infty} \left( \frac{\exp(\frac{nq}{w}) - 1}{\gamma} + \alpha + (n-1)\beta \right) \frac{e^{-\frac{\lambda}{\mu q}} (\frac{\lambda}{\mu q})^n}{\lambda n!}.$$

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

$$\begin{aligned} \phi(\lambda,q) &= \frac{\exp\left(\frac{\lambda}{\mu q}(\exp(\frac{q}{w}) - 1)\right) - 1}{\lambda\gamma} + \\ &\quad \alpha \frac{1 - \exp(-\frac{\lambda}{\mu q})}{\lambda} + \beta \left( \frac{1}{\mu q} - \frac{1 - \exp(-\frac{\lambda}{\mu q})}{\lambda} \right). \end{aligned} \tag{23}$$

Note that, as $\lambda \to 0$, (23) also captures the energy expenditure for a single user which sees no other flows than itself:

$$\lim_{\lambda \to 0} \phi(\lambda, q) = \frac{1}{\mu q} \left( \frac{\exp(\frac{q}{w}) - 1}{\gamma} + \alpha \right). \tag{24}$$

The first term in (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 (23) accounts for circuit energy. As mentioned in Lemma 1 and Theorem 2, as $q$ goes to zero, the circuit energy goes to $\frac{\alpha}{\lambda}$. Furthermore as the load grows, it becomes asymptotically negligible.

The third term in (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

$$e := \operatorname*{argmin}_{q>0} \phi(\lambda, q). \tag{25}$$

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., $\phi(\lambda, q) \approx s_1 q + s_2 + \frac{\beta}{\mu q}$, 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:

$$e \approx w\sqrt{2\exp\left(-\frac{\rho}{w}\right)\beta\gamma}. \tag{26}$$

*Remark 4.1 (**Throughput region**):* Eq. (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 $\beta$ is high; otherwise accumulated users will consume too much idling energy.

### F. 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 (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 (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 $\left(\frac{1}{n(t)} f_i(n(t)x) + \frac{n(t)-1}{n(t)}\beta_i\right)/x$ where $f_i(\cdot)$ is a user-indexed version of (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

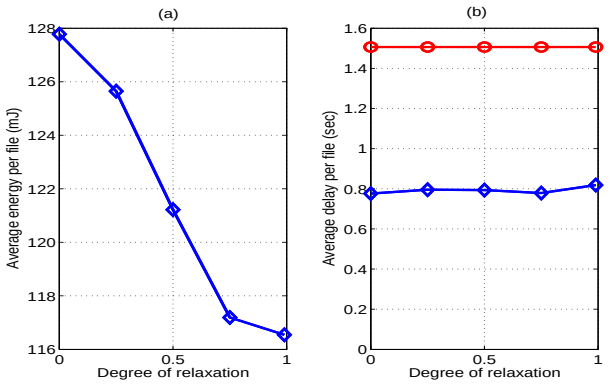



Fig. 8. Additional energy saving by relaxed target rate in Rayleigh fading channels. $q_i$ = 320 kbps (1.5 second delay for 60 kbyte file), $c_{\max}$ = 5.1 Mbps, 30 % offered load. (a) average energy per file, (b) target delay ($\circ$), and the achieved delay ($\diamond$). Parameters are same to the simulations in Section IV-H

determined based on

$$e_i(t) \quad := \quad \underset{x\geq 0}{\operatorname{argmin}} \left\{ \frac{f_i(n(t)x) + (n(t)-1)\beta_i}{x} \right\} \quad (27)$$

i.e., the throughput that minimizes the average energy per bit. Since (27) is differentiable and convex, $e_i(t)$ is given by simple calculus such as

$$e_i(t) = \frac{w}{n(t)} \left[ W\left( \frac{\alpha_i + (n(t)-1)\beta_i}{e} \gamma_i(t) - \frac{1}{e} \right) + 1 \right]. \quad (28)$$

Using (28), each mobile can determine its own energy efficient rate $e_i(t)$ given $n(t)$.

*Remark 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 [21].

***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

$$r_i(t) \quad = \quad \min \left[ \max \left[ e_i(t), q_i \right], \frac{c_i(t)}{n(t)} \right], \; i \in \mathcal{A}(t). \quad (29)$$

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

$$\bar{r}_i(t) = \nu \bar{r}_i(t-1) + (1-\nu) r_i(t), \; i \in \mathcal{A}(t) \quad (30)$$

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

$$q_i(t) \quad = \quad \frac{q_i - \nu \bar{r}_i(t-1)}{1-\nu}, \; i \in \mathcal{A}(t). \quad (31)$$

This relaxes the time scale over which the performance target should be met and contributes to further energy savings. Fig. 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

$$r_i(t) \quad = \quad \min \left[ \max \left[ e_i(t), q_i(t) \right], \frac{c_i(t)}{n(t)} \right], \; i \in \mathcal{A}(t). \quad (32)$$

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 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 (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 following results are shown in the Appendix II.

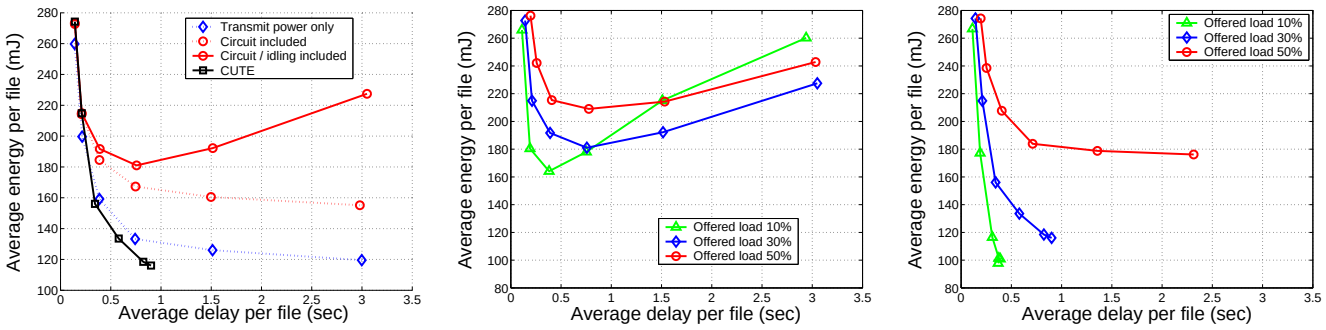
*Theorem 3 (**Convergence of CUTE**):* Suppose that the number of users and channel gains are *fixed*, and consequently $e_i(t) = e_i$ and $\frac{c_i(t)}{n(t)} = \frac{c_i}{n}$ are fixed. Then, the average throughput $\bar{r}_i(t)$ and the transmission rate $r_i(t)$ both converge to $\min(\max(q_i, e_i), \frac{c_i}{n})$. Thus, if feasible, CUTE converges to the greater of $q_i$ and $e_i$, otherwise, to $\frac{c_i}{n}$.

*Theorem 4 (**Convergence speed**):* Both of $\bar{r}_i(t)$ and $r_i(t)$ converge to the equilibrium rate at least exponentially fast.

### G. 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 [23], 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 [35]–[39]. 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 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.




(a) (b) (c)

Fig. 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.

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

$$r_i(t) = \min\left(\max\left(e_i(t), q_i(t)\right), c_i(t)\right) \mathbf{1}_{\{i=s_\theta(t)\}}, \qquad (33)$$

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

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

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

$$q_i(t) \quad = \quad \frac{n(t)q_i - \bar{r}_i(t-1)\nu}{1-\nu}, \, i \in \mathcal{A}(t) \qquad (34)$$

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

### *H. 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 [13]. 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 II. 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 $q_i = (1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}) \times 5.1$ Mbps to show how the user's preference on energy savings against fast transmission impacts the energy-delay tradeoff.

Fig. 9 demonstrates energy-delay tradeoffs under round-robin scheduling. Fig. 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 $\square$). 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 $\frac{\alpha}{\lambda} = 36.2$ mW as stated in Lemma 1. Comparing solid $\circ$ line with solid $\square$ 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 4.2 transmitting at rate $e_i(t)$ is energy-opportunistic.

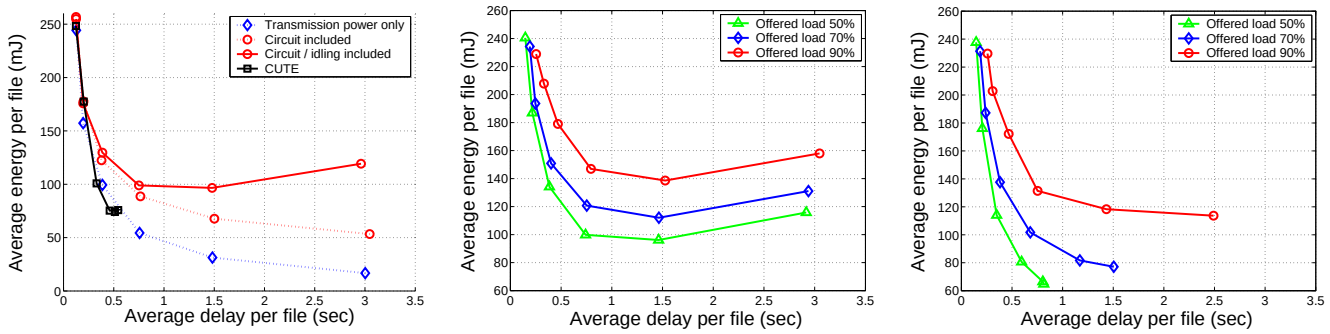
Fig. 9 (b) shows the average energy and delay when

$$r_i(t) = \min\left[q_i(t), \frac{c_i(t)}{n(t)}\right], \, i \in \mathcal{A}(t) \qquad (35)$$

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 2 and Example 6.

Finally, Fig. 9 (c) shows the performance of CUTE when (35) is replaced by (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. 9 (b) moves to (0.9 sec, 116 mJ) in Fig. 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 4.3.

Results for the case where opportunistic scheduling is used are shown in Fig. 10. 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




(a) (b) (c)

Fig. 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.

of round-robin scheduling. For example, comparing Fig. 9 (a) and Fig. 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. 9 (c) and Fig. 10 (c) with offered load 50% also shows that both of the energy and delay become less than half.

## V. 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 paper, our approach appears to be quite promising.

This work is not the final word on this topic. As mentioned earlier we expect the approach to be suitable for a broader set of multiple access technologies, e.g., beyond TDMA, FDMA to OFDMA, and extended to multiple cell scenario. 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 interferences. Thus one might expect to achieve even better energy savings, or in the case of file transfers to see an improved energy-delay tradeoff.

## 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.,

$$\delta = \sum_{i \in \mathcal{A}} \frac{r_i}{x_i}. \tag{36}$$

By substituting (13) into (36), we have

$$\delta(\kappa) = \frac{1}{w} \sum_{i \in \mathcal{A}} \frac{r_i}{W\left(\frac{(\kappa+\xi_i)\gamma_i - 1}{e}\right) + 1}. \tag{37}$$

Note that $\delta(\kappa)$ is a convex and monotone decreasing function of $\kappa$. From KKT conditions in (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 (13) and (14). This is the case for Example 2.

***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.,

$$\kappa_{m+1} = \max\left[\kappa_m - \frac{\delta(\kappa_m) - 1}{\delta'(\kappa_m)}, \kappa^{\min}\right] \tag{38}$$

where

$$\delta'(\kappa) = -\frac{1}{w} \sum_{i \in \mathcal{A}} \frac{r_i W'\left(\frac{(\kappa+\xi_i)\gamma_i - 1}{e}\right)}{\left(W\left(\frac{(\kappa+\xi_i)\gamma_i - 1}{e}\right) + 1\right)^2} \frac{\gamma_i}{e} \tag{39}$$

and $W'(z) = \frac{W(z)}{z(1+W(z))}$ if $z \neq 0$, and $W'(0) = 1$ [34]. Although $\kappa_m$ converges to $\kappa^*$ superlinearly (because it is Newton's method [31]), a good initial value further reduces the number of iterations. In particular we start the iteration at $\kappa^{\min}$ where

$$\kappa^{\min} = \left[\min_i \left(\frac{\left(\exp(v)(v-1)+1\right)}{\gamma_i} - \xi_i\right)\right]^+ \tag{40}$$

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 3:* If $\bar{r}_i(t)$ converges, then, from (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 (32) into (30),

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

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. 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. 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 4:* 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. ■

## REFERENCES


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

[2] B. Prabhakar, E. Uysal Biyikoglu, and A. El Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *Proc. IEEE INFOCOM*, 2001, vol. 1, pp. 386–394.

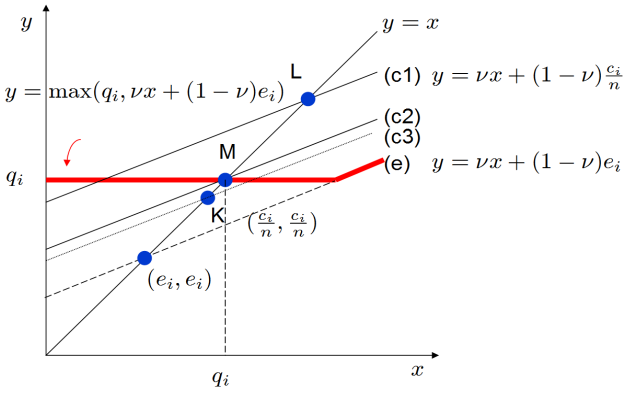
[3] A. El Gamal, C. Nair, B. Prabhakar, E. Uysal-Biyikoglu, and S. Zahedi, "Energy-efficient scheduling of packet transmissions over wireless networks," in *Proc. IEEE INFOCOM*, 2002, vol. 3, pp. 1773–1782.

[4] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay-bounded packet scheduling of bursty traffic over wireless channels," *IEEE Trans. Infomation Theory*, vol. 50, pp. 125–144, 2004.

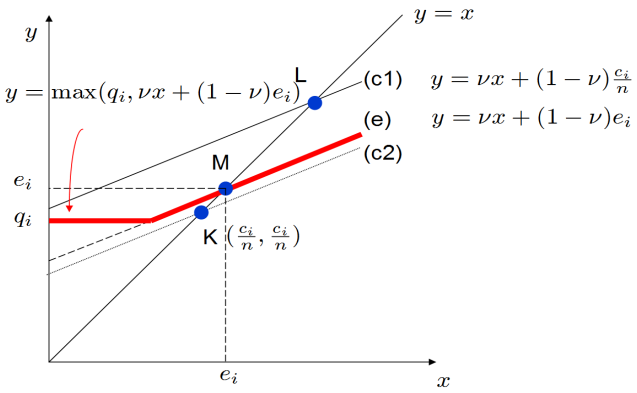
[5] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wireless Comm.*, vol. 4, pp. 2349–32360, Sep. 2005.

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

[7] 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 Comm.*, vol. 6, pp. 617–628, Feb. 2007.


(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 $\min(e_i, \frac{c_i}{n})$.

Fig. 11. Geometric proof of convergence theorem.


[8] 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.*, vol. 15, no. 6, pp. 1359–1372, 2007.

[9] M. Kodialam, T. V. Lakshman, and S. Sengupta, "Traffic-oblivious routing for guaranteed bandwidth performance," *IEEE Comm. Mag.*, pp. 46–51, Apr. 2007.

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

[11] B. Rengarajan and G. de Veciana, "Architecture and abstraction for environment and traffic aware system-level coordination of wireless networks: the downlink case," in *Proc. IEEE INFOCOM*, 2008, to apprear.

[12] 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.*, vol. 16, no. 6, pp. 858–874, Aug. 1998.

[13] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. IEEE INFOCOM*, 2003.

[14] WiMAX power amplifier ADL5570 and 5571, "http://www.analog.com/uploadedfiless/data_sheets/adl5570.pdf, http://www.analog.com/uploadedfiless/data_sheets/adl5571.pdf," *Analog Device*, Sep. 2007.

[15] 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.

[16] G. J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Trans. on Veh. Technol.*, vol. 42, no. 4, pp. 641–646, Nov. 1993.

[17] R. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Jour. Select. Areas in Comm.*, vol. 13, pp. 1341–1347, 1995.

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

[19] 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.

[20] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Infomation Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[21] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Infomation Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[22] H. Wang and N. B. Mandayam, "Opportunistic file transfer over a fading channel under energy and delay constraints," *IEEE Trans. Communications*, vol. 53, no. 4, pp. 632–644, Apr. 2005.

[23] K-K. Leung and C. W. Sung, "An opportunistic power control algorithm for cellular network," *IEEE/ACM Trans. Networking.*, vol. 14, no. 3, pp. 470–478, Jun. 2006.
[24] 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.
[25] Y. Yu, B. Krishnamachari, and V. K. Prasanna, "Energy-latency tradeoffs for data gathering in wireless sensor networks," in *Proc. IEEE INFOCOM*, 2004.
[26] 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.
[27] 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.
[28] A. Kumar, D. Manjunath, and J. Kuri, *Communication Networking*, Elsevier, 2003.
[29] 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.
[30] D. Bertsekas and R. Gallager, *Data Networks*, 1992.
[31] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.
[32] Y. Yao and G. B. Gianakis, "Energy-efficient scheduling for wireless sensor networks," *IEEE Trans. Communications*, vol. 53, no. 8, pp. 1333–1342, Aug. 2005.
[33] S. Cui, A. Goldsmith, and A. Bahai, "Joint modulation and multiple access optimization under energy constraints," in *Proc. IEEE Glob. Telecom. Conf.*, 2004, pp. 151–155.
[34] 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*, vol. 5, pp. 329–359, 1996.
[35] R. Knopp and P.A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, 1995, pp. 331–335.
[36] S. Patil and G. de Veciana, "Measurement-based opportunistic scheduling for heterogeneous wireless systems," *IEEE/ACM Trans. Networking, submitted*, 2006.
[37] 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*, vol. 53, pp. 1919–29, Nov. 2005.
[38] X. Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, 2003.
[39] S. Shakkottai, "Scheduling for multiple flows sharing a time-varying channel: the exponential rule," *American Mathematical Society Translations, Series 2*, vol. 207, 2002.