Cross-layer Wireless Resource Allocation: Fundamental Performance Limits

Randall A. Berry[*]and Edmund M. Yeh[‡]

1 INTRODUCTION

A fundamental problem in networking is the allocation of limited resources among the users of the network. In a traditional layered network architecture, the resource to be allocated at the medium access control (MAC) and network layers is the use of communication links, viewed as "bit-pipes" that deliver data at a fixed rate with occasional random errors. This bit pipe is a simple abstraction of the underlying physical and data link layers. This abstraction has, in some ways, caused the research community to split into two distinct groups, which we shall refer to as the networking and communication communities. Research in the networking community has focused on allocating these bit pipes among different streams of randomly arriving traffic using approaches such as packet scheduling and collision resolution. The goal here is to efficiently utilize the bit pipes while providing acceptable quality of service (QoS) in terms of delay and throughput to each user. In contrast, the communication community has focused on building better bit pipes, *i.e.* improving the transmission rate or spectral efficiency for a given channel through improved detection, modulation and coding. The random arrivals and departures of traffic are typically ignored and delay is not addressed. Though this separation has many advantages, both practically and conceptually, there is growing awareness that this simple bit-pipe view is inadequate, particularly in the context of modern wireless data networks. Indeed, as highlighted throughout this issue, significant performance gains can be achieved by various *cross-layer* approaches, *i.e.* approaches that jointly consider physical layer and higher networking layer issues in an integrated framework.

In this paper, we consider several basic cross-layer resource allocation problems for wireless fading channels. Here, the resources to be allocated include the transmission power and rate assigned to each user. In modern

wireless systems, a variety of link adaptation techniques such as adaptive modulation and coding or variable-rate spreading are employed which enable a user's data rate to be adapted over time based in part on time-varying channel fading. This results in a physical layer that is no longer well modeled as a fixed-rate bit pipe; instead, a much richer abstraction is required. In this setting, our focus is on characterizing fundamental performance limits, taking into account both network layer QoS and physical layer performance. We note that at the physical layer, fundamental communication limits established by information theory are, in many cases, well understood. However, when higher-layer objectives such as delay are taken into account, much less is known about fundamental performance trade-offs.The problems surveyed in this paper are attempts to address such basic questions. Their solutions serve to establish some benchmarks regarding the achievable performance of cross-layer schemes.

There are several reasons why cross-layer approaches are particularly well-suited for wireless data networks. First, a wireless channel is inherently a shared medium. Efficient resource sharing mechanisms in this setting depend strongly on both the stochastic nature of user activity as well as the selection of physical-layer coding and modulation schemes [1, 2]. For instance, consider a multiaccess problem where a group of distributed users are accessing a common channel. Assuming a simple collision model (*i.e.* only one user can successfully transmit at any time) leads naturally to the classic ALOHA and CSMA algorithms [3],whereas a more CDMA-like model (allowing multiple users to be decoded simultaneously) has very different implications (*e.g.* [4]). An information-theoretic multiaccess model implies still another set of conclusions [2, 5–8]. Second, in wireless networks, where channel quality can vary dramatically in both time and frequency, knowledge of the channel state can be exploited by the system to significantly improve performance. For example, at the physical layer, in a single user fading channel, the transmission scheme that maximizes the long-term throughput results in transmitting more information in good channel states and less in poor conditions [9]. However, when packet delay is taken into account it may not be feasible to delay transmission until channel conditions improve. In a multi-user setting, another important characteristic is that channel quality varies across the user population. This results in the phenomenon of *multi-user diversity* [10],whereby as the number of users in a system increases, the probability that some user has a very good channel also increases. Exploiting this diversity results in a total system capacity that is increasing with the number of users. However, once again, this must be balanced with network layer issues such as fairness and delay. Finally, the efficient use of energy in mobile devices is of paramount concern in wireless networks. This turns out to be an issue which cuts across almost every protocol layer. In particular, reducing the transmission energy

used at the physical layer may result in higher error rates or lower transmission rates, which again affects network layer performance. All of the above coupling effects demonstrate the need to consider network-layer quality of service issues such as throughput and delay jointly with physical-layer issues such as channel fading, coding, and modulation.

In this paper, we focus primarily on multiaccess (uplink) models, *i.e.* communication from mobile users to a single base-station or access point. We will also point out several issues that apply to broadcast (downlink) models as well. We consider a situation where randomly arriving data is buffered until it is transmitted and resources are allocated as a function of each flow's buffer occupancy and channel state. We are primarily interested in the case where a centralized controller makes all resource allocation decisions, though some comments about distributed approaches are also included. In order to characterize fundamental performance limits, we address these problems within an information theoretic framework. Specifically, when allocating resources, such as rate and power, these quantities are constrained by the appropriate capacity region, which depends on the current channel state. Since information theoretic capacity regions characterize asymptotic limits, requiring arbitrary long coding lengths, a careful reader may argue that such results are not applicable in a setting where delays are important. We address this issue in two ways. First, no matter what code lengths are used in practice, information theory provides an upper bound to all achievable rates. For example, for each channel considered here, a corresponding converse coding theorem [11] establishes that reliable communication is impossible outside the capacity region, *for all coding lengths.* Second, the gap between information-theoretic limits and the performance of practical codes with reasonable complexity has narrowed considerably in recent years, due to rapid advances in coding technology.For fading channels, as long as the coherence time is reasonably long, as is the case in typical situations[1], it is not unreasonable to assume that powerful codes with manageable block lengths can be employed to approximate information-theoretic limits. Moreover, channel coherence times are typically much smaller than the relevant time-scales at the network layer. Hence, there is no need to consider using shorter codes to further reduce delays. Finally, we note that the framework presented here is quite general and can accommodate other physical layer models, such as specific coding and modulation schemes.

---

[1]For instance, for a user traveling at urban speeds, the coherence time is typically on the order of tens of milliseconds, while the bandwidth is on the order of megahertz, implying that the coherence time corresponds to a coding length of at least several thousand symbols.

2 MULTIACCESS FADING CHANNELS

We consider the multiaccess (uplink) wireless communication setting, where multiple transmitters send to a single receiver, in the same time and frequency locality. Consider an M-user slowly-varying, flat-fading[2] Gaussian multiaccess channel with bandwidth W, given by the model

$$Y(t) = \sum_{i=1}^{M} \sqrt{H_i(t)} X_i(t) + Z(t). \tag{1}$$

Here, $H_i(t)$ is the fading process of the ith user, $X_i(t)$ is the transmitted signal of the ith user, $Z(t)$ is white Gaussian noise with noise density $N_0/2$, and $Y(t)$ is the received signal. Assume that transmitter i has a long-term average power constraint $\overline{P}_i$ and a short-term peak power constraint $\hat{P}_i$.[3] Next, assume that the channel coherence times are sufficiently long as to allow for long code lengths at a fixed joint fading level $\boldsymbol{h}$.

Given that the ith transmitter experiences a fixed channel fading level h_i and employs a fixed power level p_i, the information-theoretic multiaccess capacity region $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$ specifies the set of all transmission rates $\boldsymbol{r}$ (in bits per second) at which reliable communication is possible *under any coding and modulation scheme*. This capacity region [11] is the set of all non-negative vectors $\boldsymbol{r}$ such that

$$\sum_{i \in S} r_i \leq W \log \left(1 + \frac{\sum_{i \in S} h_i p_i}{N_0 W} \right) \text{ for all } S \subseteq \{1, \ldots, M\}. \tag{2}$$

In the two-user case, $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$ is a pentagon, as shown in Figure 1. For the M-user case, it is a bounded convex polyhedron defined by $2^M - 1$ linear inequalities and M non-negativity constraints.

An important observation is that in order to achieve all rates in the capacity region, *joint multi-user coding techniques* must be employed. Indeed, CDMA-like strategies, whereby the receiver decodes each user regarding the transmissions of all other users as noise, and simple time-sharing or scheduling strategies, whereby only one user transmits to the receiver at a time, can typically achieve only a proper subset of the rates in the information-theoretic capacity region (see Figure 1) [1]. To achieve all rates in $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$, a procedure called *successive decoding* can be used. For instance, the corner point $\boldsymbol{r}_A$ in Figure 1 is not achievable by pure time-sharing or a

---

[2]For a slowly-varying channel, the symbol duration T_s is much smaller than the channel coherence time T_{coh}, the time interval over which the fading is roughly constant. Flat fading channels are non-frequency-selective, in the sense that the signal bandwidth W is much smaller than the channel coherence bandwidth B_{coh}, the band over which fading is roughly constant.

[3]The average power constraint may correspond to a battery energy constraint, while the peak power constraint may correspond to a regulatory constraint.
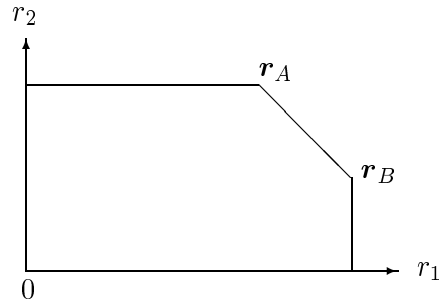
Figure 1: Illustration of $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$ for the two-user case. The extreme point $\boldsymbol{r}_A$ can be achieved by decoding the users successively in the order (1 2), while $\boldsymbol{r}_B$ is achieved by successively decoding in the order (2 1).

CDMA-like strategy, but is achievable by successively decoding user 1 (regarding user 2 as interference in addition to background noise), and then (after subtracting the estimate for user 1 from the received signal), decoding user 2 (facing, with high probability, only background noise). To achieve $\boldsymbol{r}_B$, the receiver implements successive decoding in the opposite order, decoding user 2 first, and then user 1. The successive decoding strategy is generalizable to M users, and it turns out that $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$ has precisely $M!$ extreme points, one corresponding to each possible permutation of $\{1, \ldots, M\}$ [11].

2.1 Random Arrivals and Resource Allocation

Our focus in this paper is on systems where packets for each user arrive to be transmitted at random instants in time. We follow the formulation of [5–8]. Specifically, we model the ith data source as generating packets according to an ergodic counting process $A_i(t)$, where $A_i(t)$ is the number of packet arrivals up to time t. The packet lengths for source i are i.i.d. according to distribution function $F_{Z_i}(\cdot)$ with $\mathsf{E}[Z_i] < \infty$ and $\mathsf{E}[Z_i^2] < \infty$. Next, assume that each source i, $i = 1, \ldots, M$, has its own (infinite-capacity) buffer into which its packets arrive. Packets for the ith source are stored in the ith buffer until they are served by transmitter i. The transmission power $P_i(t)$ and rate $R_i(t)$ used by transmitter i at time t are to be dynamically allocated so as to optimize throughput and delay. At the physical layer, we assume that at any time t, any set of powers and rates from the instantaneous multiaccess information-theoretic capacity region can be allocated to the transmitters, as long as average and peak power constraints are satisfied.

We now explicitly pose the dynamic resource allocation problem. Let $U_i(t)$ be the *number of untransmitted bits, or the amount of unfinished work* in queue i at time t. Consider a stationary controller which at any time $t \geq 0$ takes as inputs $\boldsymbol{H}(t)$ and $\boldsymbol{U}(t)$ and outputs a power allocation $\boldsymbol{P}(t)$, and a rate allocation $\boldsymbol{R}(t)$, to transmitters 1 to
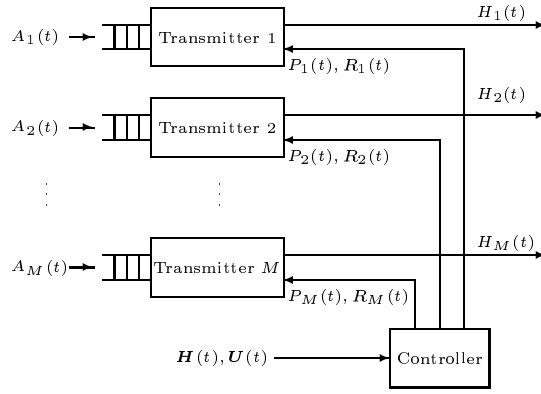
5

Figure 2: Power and rate allocation for multiaccess fading channels.

M. The controller does this by first choosing a *power control policy* $\boldsymbol{p} = \mathcal{P}(\boldsymbol{h}, \boldsymbol{u})$ satisfying $\mathsf{E}[\mathcal{P}_i(\boldsymbol{H}, \boldsymbol{U})] \leq \bar{P}_i$ for all i, where the expectation is taken with respect to the steady-state distribution induced by the controller, and $\mathcal{P}_i(\boldsymbol{h}, \boldsymbol{u}) \leq \hat{P}_i$ for all $(\boldsymbol{h}, \boldsymbol{u})$, for all i. Here, $p_i = \mathcal{P}_i(\boldsymbol{h}, \boldsymbol{u})$ is the power allocated to transmitter i in response to fading state $\boldsymbol{h}$ and queue state $\boldsymbol{u}$. Next, the controller chooses a *rate allocation policy*[4] $\boldsymbol{r} = \mathcal{R}(\boldsymbol{h}, \boldsymbol{u}) \in \mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$ where $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$ is given by (2). That is, the controller is allowed to adopt stationary power policies $\mathcal{P}$ which satisfy the average power constraints $\bar{\boldsymbol{P}}$ and peak power constraints $\hat{\boldsymbol{P}}$, and given $\mathcal{P}$, the controller is allowed to allocate any rate in the multiaccess capacity region induced by the power policy $\mathcal{P}$. The set-up is illustrated in Figure 2. Our formulation assumes that all transmitters as well as the receiver have access (possibly through side communication channels) to global channel and queue state information $\boldsymbol{H}(t)$ and $\boldsymbol{U}(t)$. As described in [12], this setting can often be approximated via feedback in practical wireless systems.

2.2 *Stability Region and Throughput Optimal Resource Allocation*

The first significant question for the multiaccess queueing system concerns the *stability region*, *i.e.* the set all bit arrival rates for which no queue "blows up." First, some definitions. Let $\lambda_i = \lim_{t \to \infty} A_i(t)/t$ denote the packet arrival rate to queue i, and let $\rho_i = \lambda_i \mathsf{E}[Z_i]$ be the bit arrival rate to queue i. We define stability as in [13]. Consider the "overflow" function $f_i(\xi) = \limsup_{t \to \infty} \frac{1}{t} \int_0^t 1_{[U_i(\tau) > \xi]} d\tau$, where 1_A is the indicator of the event A. We say that the multiaccess system is *stable* for a particular resource allocation policy if $f_i(\xi) \to 0$ as $\xi \to \infty$ for all i. The *stability region* of the multiaccess system is the set of all bit arrival rate vectors $\boldsymbol{\rho}$ for which there exist some a feasible power control policy and a rate allocation policy under which the system is stable.

---

[4]Note that due to the nature of the constraints, there is no loss of optimality in choosing $\mathcal{P}$ and $\mathcal{R}$ in a two-stage manner.

It is established in [6] that the stability region is equal to the information-theoretic capacity region under power control, defined in [14]. This region is given by $\mathcal{C}_{MAC}(\bar{\boldsymbol{P}}, \hat{\boldsymbol{P}}) = \bigcup_{\mathcal{P} \in \mathcal{F}} \mathcal{C}_{MAC}(\mathcal{P})$ [14]. Here, $\mathcal{P}$ is a power control policy *depending only on the fading state* $\boldsymbol{h}$ ($\mathcal{P}(\boldsymbol{h}, \boldsymbol{u}) = \mathcal{P}(\boldsymbol{h})$), and $\mathcal{F}$ is the set of all feasible power control policies depending only on the fading state which satisfy all peak and average power constraints. Finally, $\mathcal{C}_{MAC}(\mathcal{P})$ is the set of all non-negative $\boldsymbol{r}$ such that $\sum_{i \in S} r_i \leq \mathsf{E}\left[W \log \left(1 + \frac{\sum_{i \in S} H_i \mathcal{P}_i(\boldsymbol{H})}{N_0 W}\right)\right]$ for all $S \subseteq \{1, \ldots, M\}$. That is, $\mathcal{C}_{MAC}(\mathcal{P})$ is the average capacity region (averaged over all fading states) corresponding to a particular power policy $\mathcal{P} \in \mathcal{F}$. Suppose joint arrival process $\{\boldsymbol{A}(t)\}$ and joint fading process $\{\boldsymbol{H}(t)\}$ are modulated by a finite-state ergodic Markov chain. Then, it is shown in [6] that the multiaccess queueing system can be stabilized by some power control and rate allocation policy if $\boldsymbol{\rho} \in \text{int}(\mathcal{C}_{MAC}(\bar{\boldsymbol{P}}, \hat{\boldsymbol{P}}))$. Conversely, the multiaccess fading channel cannot be stabilized if $\boldsymbol{\rho} \notin \mathcal{C}_{MAC}(\bar{\boldsymbol{P}}, \hat{\boldsymbol{P}})$, as long as the average and peak power constraints are satisfied.

The stability result states that if $\boldsymbol{\rho} \in \text{int}(\mathcal{C}_{MAC}(\bar{\boldsymbol{P}}, \hat{\boldsymbol{P}}))$, then the queues can be stabilized. In general, however, this may require knowing the value of $\boldsymbol{\rho}$. In reality, the arrival rates $\boldsymbol{\rho}$ can be learned only over time. One would prefer to find *adaptive* resource allocation policies which can stabilize the system *without* knowing $\boldsymbol{\rho}$, as long as $\boldsymbol{\rho} \in \text{int}(\mathcal{C}_{MAC}(\bar{\boldsymbol{P}}, \hat{\boldsymbol{P}}))$, *i.e.* the system is stabilizable. Such a resource allocation policy is referred to as *throughput optimal*. Throughput optimal scheduling for fading channels has been examined in [13, 15–19]. These papers, while offering many valuable insights, do not consider information-theoretic optimal coding at the physical layer, and do not account for the effect of power control subject to long-term constraints. These important considerations are taken into account in [7], where it is shown that an adaptive version of the power and rate allocation algorithm derived by Tse and Hanly [14] is throughput optimal for the multiaccess queueing system.

In [14], Tse and Hanly consider the problem of to maximizing a weighted combination of long-term transmission rates in a multiaccess channel where all transmitters have *infinite backlogs of bits*, and both the transmitters and receivers have access to the channel state. This problem can be stated as

$$\max \sum_{i=1}^{M} \mu_i r_i \quad \text{subject to } \boldsymbol{r} \in \mathcal{C}_{MAC}(\bar{\boldsymbol{P}}, \hat{\boldsymbol{P}}) \tag{3}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$ is a vector of nonnegative weights. Using a Lagrangian formulation of (3) and the underlying polymatroidal structure of $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$, they show that (3) is equivalent to solving a family of optimization problems over parallel Gaussian multiple access channels, one for each fading state $\boldsymbol{h}$. Their analysis

yields a feasible power control policy (satisfying peak and average power constraints) and a rate allocation policy (satisfying capacity constraints) which solve (3). Notice that for a given direction $\boldsymbol{\mu}$, the Tse-Hanly power control policy $\mathcal{P}_{TH}(\boldsymbol{h}, \boldsymbol{\mu})$ and rate allocation policy $\mathcal{R}_{TH}(\boldsymbol{h}, \boldsymbol{\mu})$ are *functions of $\boldsymbol{h}$ only.*

In [7], it is proved that a throughput optimal resource allocation policy for the multiaccess system with *random packet arrivals* is given by the Tse-Hanly solution, *with the direction $\boldsymbol{\mu}$ chosen to correspond to the queue state $\boldsymbol{u}$.* Specifically, the throughput optimal policy is given by

$$\mathcal{P}^*_{MAC}(\boldsymbol{h}, \boldsymbol{u}) = \mathcal{P}_{TH}(\boldsymbol{h}, \boldsymbol{\alpha} * \boldsymbol{u}), \ \mathcal{R}^*_{MAC}(\boldsymbol{h}, \boldsymbol{u}) = \mathcal{R}_{TH}(\boldsymbol{h}, \boldsymbol{\alpha} * \boldsymbol{u}) \tag{4}$$

where $\boldsymbol{u}$ is the queue state, $\boldsymbol{\alpha}$ is any vector of positive numbers, and $\boldsymbol{\alpha} * \boldsymbol{u}$ is the vector whose ith component is $\alpha_i u_i$. The vector $\boldsymbol{\alpha}$ can be seen as a set of weights representing the relative priorities of the various users. The proof of the throughput optimality results in [7] makes use of the Foster-Lyapunov Criterion for the stability of Markov chains [13].

To interpret the throughput optimal resource allocation policies $\mathcal{P}^*_{MAC}$ and $\mathcal{R}^*_{MAC}$, let $\boldsymbol{v} = \boldsymbol{\alpha} * \boldsymbol{u}$, where $v_i = \alpha_i u_i$, be the vector of weighted queue sizes. In the case of one user ($M = 1$), it can be shown [14] that $(\mathcal{P}^*_{MAC}, \mathcal{R}^*_{MAC})$ reduces to the well-known water-filling scheme [9], whereby more power is allocated to favorable channel states, and less or no power is allocated to unfavorable channel states. In the case of multiple users ($M > 1$) where all weighted queue sizes are the same: $v_1 = v_2 = \cdots = v_M$, and the fading conditions are symmetric, $(\mathcal{P}^*_{MAC}, \mathcal{R}^*_{MAC})$ reduces to the "multi-user waterfilling" scheme of Knopp and Humblet [10], whereby when all channel states are sufficiently unfavorable, no one transmits. Otherwise, only the user with the best channel condition transmits. In the general case of many users and unequal weighted queue lengths, more than one user typically transmit. Little in general can be said about the optimal power policy $\mathcal{P}^*_{MAC}$. The optimal rate allocation policy $\mathcal{R}^*_{MAC}$, however, satisfies a general principle we refer to as *Longest Weighted Queue Highest Possible Rate* (LWQHPR). This principle holds for any given feasible power control policy $\mathcal{P}$, and is described as follows. Given $\mathcal{P}$, $\mathcal{R}^*_{MAC}(\boldsymbol{h}, \mathcal{P}(\boldsymbol{h}, \boldsymbol{u}), \boldsymbol{u})$ is given by maximizing $\sum_i v_i r_i$ over $\mathcal{C}_{MAC}(\boldsymbol{h}, \mathcal{P}(\boldsymbol{h}, \boldsymbol{u}))$. Due to the polymatroidal nature of $\mathcal{C}_{MAC}(\boldsymbol{h}, \mathcal{P}(\boldsymbol{h}, \boldsymbol{u}))$ [14], the solution is explicitly given as follows. Let $v_{[1]} \geq v_{[2]} \geq \cdots \geq v_{[M]}$ denote

the components of $\boldsymbol{v}$ in *decreasing* order. Then, $\boldsymbol{r}^* = \mathcal{R}^*_{MAC}(\boldsymbol{h}, \mathcal{P}(\boldsymbol{h}, \boldsymbol{u}), \boldsymbol{u})$ is given by

$$r^*_{[i]} = W \log \left(1 + \frac{h_{[i]}\mathcal{P}_{[i]}(\boldsymbol{h}, \boldsymbol{u})}{\sum_{j<i} h_{[j]}(t)\mathcal{P}_{[j]}(\boldsymbol{h}, \boldsymbol{u}) + N_0 W} \right). \tag{5}$$

It can be verified that $\boldsymbol{r}^*$ is the extreme point of $\mathcal{C}_{MAC}(\boldsymbol{h}, \mathcal{P}(\boldsymbol{h}, \boldsymbol{u}))$ corresponding to successively decoding the users in the order $[M], [M-1], \ldots, [1]$. That is, the smallest component of $\boldsymbol{v}$ being decoded first, and the largest component of $\boldsymbol{v}$ being decoded last. Alternatively, $\boldsymbol{r}^*$ is given by a greedy rate allocation procedure where longer weighted queues are given preference over shorter weighted queues.[5] Hence, the name LWQHPR [7]. To illustrate the LWQHPR policy, we refer to the two-user example of Figure 1, where LWQHPR assigns $\boldsymbol{r}_A$ (corresponding to decoding order (1 2) if $v_1 < v_2$ and assigns $\boldsymbol{r}_B$ (corresponding to decoding order (2 1)) if $v_1 \geq v_2$.

We now compare the performance of the throughput optimal policies $(\mathcal{P}^*_{MAC}, \mathcal{R}^*_{MAC})$ to those of widely-used alternative resource allocation policies. Consider an example in which there are two users observing i.i.d. fading processes. For each user, $\Pr(H = h_0) = \Pr(H = 1/h_0)$ for some fixed $h_0 > 0$. The fading state remains constant for a period of T seconds, and then changes to a new independent fading state. The arrival processes are independent Poisson with $\lambda_1 = \lambda_2 = \lambda$ and packets lengths are i.i.d. exponential with parameter 1. We focus on the parameters $h_0 = 10$, $\bar{P} = \bar{P}_1 = \bar{P}_2 = 1$,[6] $T = 0.4$, $N_0 W = 1$, and equal weights ($\alpha_1 = \alpha_2$). Figure 3 shows the simulated performance of the throughput optimal strategy $(\mathcal{P}^*_{MAC}, \mathcal{R}^*_{MAC})$ relative to those of four other strategies. The sum of the average queue sizes is plotted versus the arrival rate λ for the five strategies described below. The *Throughput Optimal* strategy is given by $(\mathcal{P}^*_{MAC}, \mathcal{R}^*_{MAC})$. The strategy of *Knopp-Humblet* [10] maximizes the sum rate assuming an infinite backlog, which corresponds to the throughput optimal strategy with $u_1 = u_2$. The *Scheduling* algorithm allocates power $2\bar{P}$ to the user with the better fade and zero power to the other user. The *Constant Power LQHPR* strategy uses constant power ($\mathcal{P}_i(\boldsymbol{h}, \boldsymbol{u}) = \bar{P}$ for all i, $\boldsymbol{h}$ and $\boldsymbol{u}$) and allocates rates according to (5). The *Constant Power BCHPR* strategy also uses constant power, and (ignoring the queue size) gives the Best Channel the Highest Possible Rate (BCHPR). The experimental results demonstrate the superior performance of the throughput optimal resource allocation policy, in that its total average queue size is considerably smaller than those of the competitors.

---

[5]Note that the order of decoding is the *opposite* of the order of preference.

[6]For simplicity, we assume the peak power constraints are large enough to be ineffective.
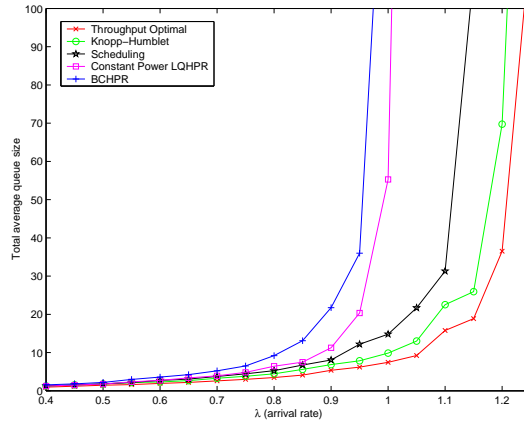
Figure 3: Total average queue size vs. arrival rate for the multiaccess fading channel under five control strategies.

2.3 Delay Optimal Resource Allocation

We have thus far concentrated on stability and throughput optimality. Stability in a queueing system implies that the queue sizes do not "blow up," but it does not indicate how large the queue sizes can be. In order to minimize the average packet delay/latency and other related QoS measures, it is necessary to keep the queue sizes *as short as possible*. The general problem of finding delay optimal joint power control and rate allocation policies to minimize average delay for multiaccess channels is still open. In [6], the problem of finding the delay optimal rate allocation policy for a given power control policy is addressed. The main result is that in a symmetric multiaccess queueing system, the symmetric version of the LWQHPR rate allocation policy given by (5) (with $\alpha_i = 1$ for all i) minimizes the average packet delay in a very strong sense.

Consider the case where all arrival processes are non-homogeneous Poisson with rate function $\lambda(t)$, and all arriving packets are i.i.d. exponential with common parameter $\mu > 0$. Due to the memoryless nature of the system, the vector $\boldsymbol{Q}(t) = (Q_1(t), \ldots, Q_M(t))$, where $Q_i(t)$ is the *number of packets* in queue i at time t, constitutes a state. Thus, we focus on resource allocation policies of the form $\mathcal{P}(\boldsymbol{h}, \boldsymbol{q})$ and $\mathcal{R}(\boldsymbol{h}, \mathcal{P}(\boldsymbol{h}, \boldsymbol{q}), \boldsymbol{q})$, where $\boldsymbol{q} = (q_1, \ldots, q_M)$ is the vector of queue lengths. We assume that the fading process $\boldsymbol{H}(t)$ is symmetric (or exchangeable) in the following sense: for all t, and all $\boldsymbol{a}$ in the fading state space $\mathcal{H}$, $\Pr\left(H_1(t) = a_1, \ldots, H_M(t) = a_M\right) = \Pr\left(H_1(t) = a_{\pi(1)}, \ldots, H_M(t) = a_{\pi(M)}\right)$ for any permutation π on the set $\{1, \ldots, M\}$. Beyond this symmetry, we do not make any other assumptions on the fading process. We focus on power policies $\mathcal{P}$ which are functions of the fading state only, *i.e.* $\mathcal{P}(\boldsymbol{h}, \boldsymbol{q}) = \mathcal{P}(\boldsymbol{h})$. We say a power policy is *symmetric* if for all $\boldsymbol{a} \in \mathcal{H}$, $\mathcal{P}_i(a_1, \ldots, a_M) = \mathcal{P}_{\pi^{-1}(i)}\left(a_{\pi(1)}, \ldots, a_{\pi(M)}\right)$ for any permutation π. That is, under a symmetric power control policy, the power allocated to a given user is determined by the fading level experienced by that user (and not on

10

the identity of that user) relative to the fading levels experienced by all other queues. For instance, suppose $M = 2$ and $a_1 > a_2$, then if $\mathcal{P}$ is symmetric, $\mathcal{P}_1(a_1, a_2) = \mathcal{P}_2(a_2, a_1)$. An example of a symmetric power control policy is the "multi-user water-filling" policy given by Knopp and Humblet [10].

Consider the version of the LWQHPR policy where $\alpha_i = 1$ for all i. We refer to this as the *Longest Queue Highest Possible Rate (LQHPR) policy*. The main result on delay optimality from [6] is the following. For an M-user symmetric multiaccess queueing system, let $\mathcal{P} : \mathcal{H} \mapsto \mathsf{R}_+^M$ be a given symmetric power control policy. Let $\boldsymbol{q}_0$ be the vector of queue sizes at time 0. Let $\boldsymbol{Q}(t)$ be the queue evolution under the LQHPR rate allocation policy, and $\boldsymbol{Q}'(t)$ be the queue evolution under any feasible rate allocation policy. Then, $\mathsf{E}\left[\varphi(\boldsymbol{Q}(t))\right] \leq \mathsf{E}\left[\varphi(\boldsymbol{Q}'(t))\right]$ for all $t \geq 0$, for all *increasing and Schur-convex*[7] functions $\varphi : \mathbb{R}^M \mapsto \mathbb{R}$. As a main example, the result holds for all symmetric[8], increasing, convex functions on $\mathbb{R}^M$. More specific examples include $\varphi(\boldsymbol{x}) = \max_{i_1 < i_2 < \cdots < i_k}(|x_{i_1}| + \cdots + |x_{i_k}|)$ for $1 \leq k \leq M$, $\varphi(\boldsymbol{x}) = \sum_{i=1}^M |x_i|^r$ for $r \geq 1$ or $r \leq 0$, and $\varphi(\boldsymbol{x}) = (\sum_{i=1}^M |x_i|^r)^{1/r}$ for $r \geq 1$. Thus, the LQHPR policy minimizes the expected value of a large class of functions of $\boldsymbol{Q}(t)$, of which $\mathsf{E}[\sum_i Q_i(t)]$ is one example. By Little's Law, this implies that LQHPR minimizes the average system delay of packets. The main technique used in proving the result of [6] earlier related results in [5] is *stochastic coupling*, a method relying directly on probabilistic intuition which is capable of generating powerful and elegant results when given certain symmetry conditions. Finally, the delay optimality of the LQHPR policy has been partially extended to the case where arriving packets are i.i.d. according to a general distribution function $F_Z(\cdot)$ with finite first and second moments [8]. This investigation requires a new analytical technique combining finite-horizon dynamic programming and renewal process theory.

3   BROADCAST FADING CHANNELS

The analytical framework established in Section 2 for multiaccess networks can be extended to broadcast (downlink) wireless networks, where a single transmitter sends separate *independent* information to M receivers, where the ith receiver has fading process $H_i(t)$ and receiver noise power density $N_{0i}/2$. For a fixed transmit power p and fading states $h_1, \ldots, h_M$, the broadcast capacity region $\mathcal{C}_{BC}(\boldsymbol{h}, p)$ [11] is very different from the multiaccess region $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$. As in the multiple-access case, however, simple time-sharing (whereby the transmitter sends to one receiver at a time) cannot achieve all points in $\mathcal{C}_{BC}(\boldsymbol{h}, p)$. To accomplish the latter, a process called

---

[7]Schur convex functions are functions which preserve an ordering called *majorization* [20].

[8]A function φ is *symmetric* on $\mathcal{A} \subset \mathbb{R}^M$ if $\varphi(\boldsymbol{x}) = \varphi(\boldsymbol{x}\mathsf{P})$ for any $\boldsymbol{x} \in \mathcal{A}$ and any M by M permutation matrix P.

superposition coding combined with the successive decoding technique discussed in Section 2 can be used [11].

In [7], a broadcast communication system with random packet arrivals is considered. It is shown that the queueing system considered in Figure 2 can be directly carried over to the broadcast case, with the following caveats. First, unlike the multiple-access case, there is only one actual transmitter in a broadcast network with long-term average power constraint $\overline{P}$ and short-term peak power constraint $\hat{P}$. There are, however, still multiple arrival processes and queues at the transmitter, corresponding to the information streams of the respective receivers. Since the transmitter uses superposition coding and allocates separate powers and rates to sub-codes for the various receivers, we can associate a "virtual transmitter" with each of the M receivers (sub-codes). Second, in the broadcast case, the power control policy $\mathcal{P}$ determines the *total power* p used by the *actual transmitter* as a function of $\boldsymbol{h}$ and $\boldsymbol{u}$ to satisfy $\mathsf{E}[\mathcal{P}(\boldsymbol{H}, \boldsymbol{U}) \leq \overline{P}$. The rate allocation policy then decides what fraction of power $\gamma_i p$ (with $\sum_{i=1}^{M} \gamma_i = 1$) and rate r_i to assign to each virtual transmitter or sub-code.

For the broadcast network, we consider the same throughput and delay questions as in the multiaccess case. The results for stability and throughput optimality are the exact analogues of those for multiaccess channels. In [7], it is shown that stability region for the broadcast network is the same as the information-theoretic capacity region $\mathcal{C}_{BC}(\overline{P}, \hat{P})$ under power control defined in [21,22]. Moreover, parallel to the multiaccess case, an adaptive version of the power and rate allocation policies designed to maximize transmission rates in a broadcast channel with infinite backlogs of bits, is throughput optimal for a system with random packet arrivals. In [21, 22], Tse, Li, and Goldsmith use a Lagrangian formulation similar to the multiaccess case to obtain the optimal power policy $\mathcal{P}_{TLG}(\boldsymbol{h}, \boldsymbol{\mu})$ and rate policy $\mathcal{R}_{TLG}(\boldsymbol{h}, \boldsymbol{\mu})$ which maximizes $\sum_i \mu_i r_i$ subject to $\boldsymbol{r} \in \mathcal{C}_{BC}(\bar{P}, \hat{P})$, where $\boldsymbol{\mu}$ is a vector of nonnegative weights.[9] In [7], it is shown that a throughput optimal policy for the broadcast channel with random packet arrivals is $\mathcal{P}^*_{BC}(\boldsymbol{h}, \boldsymbol{u}) = \mathcal{P}_{TLG}(\boldsymbol{h}, \boldsymbol{\alpha} * \boldsymbol{u}), \mathcal{R}^*_{BC}(\boldsymbol{h}, \boldsymbol{u}) = \mathcal{R}_{TLG}(\boldsymbol{h}, \boldsymbol{\alpha} * \boldsymbol{u})$, where $\boldsymbol{\alpha}$ is any vector of positive numbers, $\boldsymbol{\alpha} * \boldsymbol{u}$ is the vector whose ith component is $\alpha_i u_i$, and u_i is the number of untransmitted bits in the queue for the ith receiver. An interesting consequence of this policy is the following: if there exists a user i such that $\alpha_i u_i \geq \alpha_j u_j$ for all $j \neq i$, and such that $\frac{\alpha_i u_i h_i}{N_{0i}} \geq \frac{\alpha_j u_j h_j}{N_{0j}}$ for all $j \neq i$, then the throughput optimal policy transmits only to user i. In particular, if $\alpha_j = 1$ and $N_{0j} = N_0$ for all j, then the throughput optimal policy *transmits to the user with best fading when all queues are equal*, and *transmits to the user with longest queue when all fading levels are equal* [7]. Numerical experiments indicate that the performance of the policy

---

[9]Even though the structure of $\mathcal{C}_{BC}(\boldsymbol{h}, p)$ and $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$ are very different, greedy algorithms can be used to solve the family of optimization problems, one for each fading state $\boldsymbol{h}$, in both cases.

$(\mathcal{P}_{BC}^*(\boldsymbol{h}, \boldsymbol{u}), \mathcal{R}_{BC}^*(\boldsymbol{h}, \boldsymbol{u}))$ is superior to those of competing resource allocation policies, in that the resulting total average queue size is substantially lower [7].

Finally, we come to the problem of delay optimal resource allocation for broadcast channels. Here, very little progress has been made. Even the problem of finding the delay optimal rate allocation policy for a given power control policy has not been successfully tackled. The delay problem for broadcast channels appears to be more difficult than that for multiaccess channels, mainly because the region $\mathcal{C}_{BC}(\boldsymbol{h}, p)$ lacks the many desirable structural properties of $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$. Much more work is needed in this area.

4 ENERGY/DELAY TRADE-OFFS

Energy efficiency is a key concern for mobile wireless devices that must rely on limited battery resources. As transmission power is one of the main energy consumers in wireless devices, there has been much interest in approaches for efficiently utilizing this resource. In this context, a fundamental metric is the average energy per bit used in communication. At the physical layer, the minimum energy per bit needed to reliably communicate at a given rate R can be related to the channel's capacity. Specifically, let $C(P)$ denote the capacity of a single user channel as a function of the transmission power. Since for reliable communication $R < C(P)$, it follows that the energy per bit, E_b must satisfy $E_b > \frac{C^{-1}(R)}{R}$, where C^{-1} denotes the inverse of $C(\cdot)$. For any channel $\frac{C^{-1}(R)}{R}$ is a decreasing convex function of R; hence, energy can be conserved by transmitting at a lower rate. For example, if $C(P) = \ln(1 + P)$ as in a Gaussian noise channel, then $\frac{C^{-1}(R)}{R} = \frac{1}{R}\left(e^R - 1\right)$. Asymptotically, as $R \to 0$, the minimum energy per bit is the reciprocal of a channel's *capacity per unit cost* [23], with "cost" given by energy per channel use.

Transmitting at a lower rate conserves energy, but increases delay at the network layer. This illustrates a fundamental trade-off between energy efficiency and delay. In fading channels, reducing packet delay may further increase the required energy by forcing users to transmit when channel conditions are not favorable. The energy-delay trade-off necessitates resource allocation algorithms which optimally balance these two important concerns. A number of approaches have been applied to this problem including [24–27]. Interestingly, these approaches are useful even in a single user setting [24, 27], and even in a setting with random arrivals but no channel fading [25].

4.1 Optimal Power/Delay Resource Allocation - Single User Case

We first discuss the case where a single user is transmitting data over the fading channel in (1) (with $M = 1$). Once again, we assume that data randomly arrives and is buffered until it is transmitted. For simplicity, we consider a

discrete-time *block fading* model where during the nth time-slot the channel gain is constant with value $H[n]$, *i.e.*
$H(t) = H[n]$ for $(n-1)\Delta \leq t < n\Delta$ where Δ is the length of a time-slot. Denote the unfinished work in the
buffer at the start of the nth time-slot by $U[n]$ and let $I[n]$ be the number of bits (amount of work) arriving during
the nth time slot. The unfinished work evolves according to

$$U[n+1] = \max(U[n] + I[n] - R[n]\Delta, 0), \tag{6}$$

where $R[n]$ is the transmission rate during the nth time slot. At each time n, the rate is again specified by a sta-
tionary rate allocation policy $r = \mathcal{R}(h, u)$ that depends on the current queue state and fading state. In this section,
we view the system as first specifying a rate allocation policy and then incurring a "power cost" $P(R[n], H[n])$
that depends on this policy and the current channel state. Specifically, let $P(r, h)$ represent the minimum power
required such that the rate r is less than the corresponding channel capacity. For the channel in (1), this is given by

$$P(r, h) = \frac{N_0 W}{h}\left(2^{r/W} - 1\right), \tag{7}$$

which is an increasing convex function of the transmission rate.

For a given policy $\mathcal{R}$, let $\bar{P}(\mathcal{R}) = \lim_{n \to \infty} \mathsf{E}\left[P(\mathcal{R}(H[n], U[n]), H[n])\right]$ denote the steady-state average trans-
mission power consumed. Assuming the system is stable, the average energy per bit is given by $\bar{P}(\mathcal{R})/\bar{I}$, where $\bar{I}$
indicates the average bit arrival rate per time-slot.[10] Also, for a given policy $\mathcal{R}$, let $\bar{D}(\mathcal{R}) = \lim_{n \to \infty} \mathsf{E}\left[U[n]/\bar{I}\right]$
indicate the steady-state average queueing delay. We define the *optimal power/delay* trade-off curve, $P^*(D)$, as

$$\mathsf{P}^*(D) = \inf\{\bar{P}(\mathcal{R})|\mathcal{R}(\cdot) \text{ such that } \bar{D}(\mathcal{R}) \leq D\}, \tag{8}$$

i.e. $P^*(D)$ characterizes the minimum average power required for any policy with an average delay no greater
than D. Conversely, we also define the delay/power trade-off:

$$\mathsf{D}^*(P) = \inf\{\bar{D}(\mathcal{R})|\mathcal{R}(\cdot) \text{ such that } \bar{P}(\mathcal{R}) \leq P\}. \tag{9}$$

Assuming $\mathsf{P}^*(D)$ is strictly monotonic, $\mathsf{D}^*(P)$ will simply be its inverse. These quantities, like a rate-distortion

---

[10]Using our previous notation, $\bar{I} = \lambda \Delta \mathsf{E}[Z]$, where $\mathsf{E}[Z]$ is the expected length of a packet.
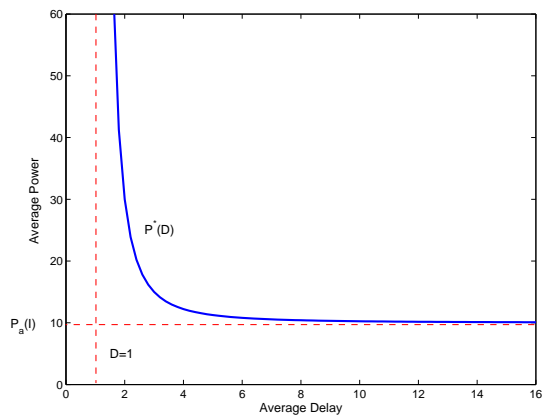
Figure 4: Example of a power/delay trade-off.

curve in source coding, provide a characterization of the fundamental power delay trade-off that can be achieved by any scheduling policy.

In [27], it is shown that $\mathsf{P}^*(D)$ is a strictly decreasing, convex function of D. From this it follows, that any point on this curve can be found by solving an average cost dynamic programming problem with a per stage cost given by $J[n] = P(R[n], H[n]) + \beta\left(U[n]/\bar{I}\right)$, where β corresponds to a Lagrange multiplier for an average delay constraint. By varying β, different points on $\mathsf{P}^*(D)$ can be found. An example of $\mathsf{P}^*(D)$ is shown in Fig. 4.

For a given β, the optimal rate allocation can be found numerically, and the optimal β for a given delay constraint can be found via standard convex programming techniques. General structural properties of the optimal policy can also be characterized [26,27]. For example, assuming the fading is i.i.d., it can be shown that the optimal transmission rate is non-decreasing in the unfinished work. In [24], a version of this problem was considered where $P(r, h)$ is linear in r for each h, as would be reasonable in the wide-band limit. This simplifies the dynamic programming problem and more precise structural results can be shown. In addition to considering average delay versus the long-term average power, there have been a number of papers which have examined related problems in a dynamic programming context e.g., [28]. In these papers, the object is again to minimize a cost related to energy, subject to various delay constraints, such as a deadline by which all packets must be sent.

4.2 Asymptotic Power Delay Trade-offs

The behavior of the power/delay trade-off can be explicitly characterized in various asymptotic regimes. In [27] it is shown that as the average delay increases, $\mathsf{P}^*(D)$ approaches an asymptotic value of $P_a(\bar{I})$ (see Fig. 4). The limiting value, $P_a(\bar{I})$, is the minimum average power needed to support the average arrival rate, which corresponds

15

to the average power level required for the channel to have a "long-term" capacity equal to $\bar{I}/\Delta$. Additionally, the rate at which $\mathsf{P}^*(D)$ converges to this asymptotic limit is given by $\frac{1}{D^2}$ under any reasonable fading distribution. That $\mathsf{P}^*(D)$ can converge no faster than this is shown using drift arguments and the convexity of $P(r, h)$. This bound is shown to be tight by showing that a simple "threshold water-filling" policy [27] is *order optimal*, *i.e.* it converges to the asymptotic limit at the optimal rate of $\frac{1}{D^2}$. This policy has only a weak dependence on the buffer occupancy via a threshold rule. When the unfinished work exceeds this threshold, the transmitter uses a water-filling policy with an average rate that is greater than the arrival rate; when the unfinished work is less than the threshold, a water-filling policy with an average rate less than the arrival rate is used. The thresholds and the average rates used in each portion are adjusted depending on the average delay. It can also be shown that some dependence on the buffer occupancy is required for a policy to be order optimal [27]. In other words, the power required by any family of policies that do not depend on the buffer occupancy can not converge to $P_a(\bar{I})$ at the rate of $\frac{1}{D^2}$ as the average delay increases. This buffer dependence is needed for a policy to maintain a backlog of packets in the buffer. This backlog enables the transmitter to better exploit good channel conditions and smoothes out some of the burstiness in the arrival process.

The behavior of $\mathsf{P}^*(D)$ can also be characterized in the regime of asymptotically small delays [29]. From the buffer dynamics in (6), the minimum possible average delay is 1 time-unit. By the small delay regime we mean the behavior of of $\mathsf{P}^*(D)$ as $D \to 1$. The asymptotic value of $P^*(D)$ in this regime depends on the fading distribution. Two distinct cases can be identified. The first case corresponds to channels such a Rayleigh fading channel where $\mathsf{P}^*(1)$ is infinite. This corresponds to a channel having a *delay-limited capacity* of zero [14]; i.e. using finite power it is impossible to send at a non-zero rate in every channel state. For these channels, $\mathsf{D}^*(P)$ will be greater than 1 for all finite P, but as $P \to \infty$, $\mathsf{D}^*(P)$ will approach this value. The second case corresponds to those channels with $\mathsf{P}^*(1) < \infty$. In this case, $\mathsf{P}^*(1)$ corresponds to the minimum average power needed for the channel to have a delay-limited capacity of $\bar{I}/\Delta$. Once again, the rate at which $\mathsf{D}^*(P)$ approaches its asymptotic limit can be characterized [29]. This also depends on the fading distribution, in particular on its behavior near zero. For example, in the case of Rayleigh fading, $\mathsf{D}^*(P) - 1$ decreases at a rate of $e^{-\alpha P}$, where α is a constant depending on the fading distribution near zero. This is much faster than the optimal rate in the large delay regime, which implies that the reduction in power by a slight increase in tolerable delay is very significant in the small delay regime. An order optimal policy in this regime is a "channel-threshold policy" which transmits at a fixed

rate, whenever the channel gain is greater than a threshold. Interestingly, this policy requires no dependence on the buffer state.

These results are robust to many variations in the model. For example, the power cost in (7), may be replaced by any convex increasing function; this can be used to reflect the power required for specific modulation and coding schemes. The results can be extended to models that allow the possibility of packet errors and retransmissions as well as models with finite buffer sizes [30].

4.3 Multiaccess Model

The above approach can be generalized to a multi-user setting, where a centralized control policy specifies the transmission rate and power for each user [30]. For example, consider the multiaccess channel with M users as in (1). In this case, let $\mathcal{R}(\boldsymbol{h}, \boldsymbol{u})$ denote a rate allocation policy that specifies the transmission rate of each user as a function of the joint fading states and buffer occupancies. Now, the corresponding power cost will be given by

$$P(\boldsymbol{r}, \boldsymbol{h}) = \inf \left\{ \sum_{i=1}^{M} w_i P_i \text{ such that } \boldsymbol{r} \in \mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p}) \right\}, \tag{10}$$

where w_i are given weights, and $\mathcal{C}_{MAC}(\boldsymbol{h}, \boldsymbol{p})$ denotes the multiaccess capacity region in (2) corresponding to power allocation $\boldsymbol{p} = (p_1, \ldots, p_M)$. A solution to the optimization can always be found such that the rate vector $\boldsymbol{r}$ lies at one of the $M!$ extreme points of the capacity region.

In this case, a natural generalization of the optimal power/delay trade-off is to define $\mathsf{P}^*(D_{sum})$ to be the minimum average weighted sum power required for the (possibly weighted) sum of the queueing delays to be no greater than D_{sum}. Once again, the power delay trade-off can evaluated via a dynamic programming formulation. The asymptotic analysis also carries through to this case. For example, as the average weighted sum delay, D_{sum} increases, $\mathsf{P}^*(D_{sum})$ can be shown to decrease to its asymptotic limit at a rate of $\frac{1}{D_{sum}^2}$. An order optimal sequence of polices can again be specified that require only a weak dependence on each user's queue state. Specifically, a centralized controller only requires one bit of information about each user's queue to implement these policies. This information will again indicate whether the queue size is above or below a threshold. Given this information the controller can identify one of 2^M quadrants within which the joint buffer state lies. The controller then implements a policy for each quadrant that depends only on the channel state $\boldsymbol{h}$. This provides insight into the amount of control information that must be shared among users in a distributed setting to implement an order optimal policy.

4.4 Off-line and Look-ahead Scheduling Algorithms

We briefly mention another approach for energy efficient scheduling from [25]. A finite horizon problem is considered with a deadline of T seconds. During time $[0, T)$ packets randomly arrive and all packets that arrive in this interval must be transmitted over a channel without fading by time T. The goal is to accomplish this using the minimum energy. This is done by specifying the transmission time per packet, τ. For a given transmission time τ, the required energy is given by $e(\tau)$, where $e(\tau)$ is a decreasing convex function. Again, $e(\tau)$ can be related to the channel capacity. In [25], an optimal "off-line" scheduling algorithm is first considered, where all packet arrivals times are known *a priori*. This results in a convex optimization problem, which admits a simple solution. Approximate "on-line" algorithms based on a look-ahead buffer are developed which exploit the structure of the off-line algorithm. These on-line algorithms are shown to require an average energy quite close to the optimal off-line algorithm via simulations. This approach has the advantage that it does not require detailed knowledge of the arrival statistics.

REFERENCES

[1] R. Gallager, "A perspective on multiaccess channels," *IEEE Trans. Information Theory*, vol. 31, no. 2, pp. 124–142, 1985.

[2] I. E. Telatar and R. Gallager, "Combining queueing theory with information theory for multiaccess," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 963–969, 1995.

[3] D. Bertsekas and R. Gallager, *Data Networks*. NJ: Prentice Hall, 1992.

[4] Q. Zhao and L. Tong, "A multi-queue service room MAC protocol for wireless networks with multipacket reception," *IEEE/ACM Transactions on Networking*, feb 2003.

[5] E. Yeh, "An inter-layer view of multiaccess communications," in *Proc. of the 2002 ISIT*, (Lausanne, Switzerland), p. 112, 2002. See Also, "Delay optimal rate allocation in multiple access communications: A cross layer approach," in *Proc. of the 2002 International Workshop on Multimedia Signal Processing*, 2002.

[6] E. Yeh and A. Cohen, "Throughput and delay optimal resource allocation in multiaccess fading channels," in *Proceedings of the International Symposium on Information Theory*, (Yokohama, Japan), p. 245, 2003.

[7] E. Yeh and A. Cohen, "Information theory, queueing, and resource allocation in multi-user fading communications," in *Proc. of the Conference on Information Sciences and Systems*, (Princeton, NJ), 2004. See also,

"Throughput optimal power and rate control for queued multiaccess and broadcast communications," in *Proc. of ISIT*, 2004.

[8] E. Yeh, "Delay optimal multiaccess communication for general packet length distributions," in *Proceedings of the International Symposium on Information Theory*, (Chicago, IL), 2004.

[9] A. Goldsmith and P. Varaiya, "Capacity of Fading Channels with Channel Side Information," *IEEE Trans. Inf. Th. vol. 43, no. 6*, pp. 1986–1992, Nov 1997.

[10] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proceedings of International Conference on Communications*, (Seattle, WA), 1995.

[11] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[12] R. Laroia, S. Uppala, and J. Li, "Designing a broadband mobile wireless network: an integrated approach," This issue.

[13] M. J. Neely, E. Modiano, and C. E. Rohrs, "Power and server allocation in a multi-beam satellite with time varying channels," in *Proceedings of Infocom 2002*, (New York City), 2002.

[14] D. Tse and S. Hanly, "Multi-access fading channels: Part I and Part II," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2796–2831, 1998.

[15] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Transactions on Information Theory*, vol. 39, no. 2, pp. 466–478, 1993.

[16] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–154, 2001.

[17] R. Leelahakriengkrai and R. Agrawal, "Scheduling in multimedia wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 52, pp. 226–239, 2003.

[18] S. Shakkottai and A. L. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule." To appear in *Translations of the AMS, A volume in memory of F. Karpelevich*, 2002.

[19] A. Eryilmaz, R. Srikant, and J. Perkins, "Stable scheduling policies for broadcast channels," in *Proc. ISIT*, p. 382, 2002.

[20] A. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Press, 1979.

[21] D. Tse, "Optimal power allocation over parallel gaussian broadcast channels," in *Proceedings of the 1997 International Symposium on Information Theory*, (Ulm, Germany), p. 27, 1997.

[22] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels: Part I: Ergodic capacity," *IEEE Trans. Information Theory*, vol. 47, no. 3, pp. 1083–1102, 2001.

[23] S. Verdu, "On channel capacity per unit cost," *IEEE Trans. on Information Theory*, vol. 32, pp. 1019–1030, 1990.

[24] B. Collins and R. Cruz, "Transmission Policies for Time Varying Channels with Average Delay Constraints," in *Proc. 1999 Allerton Conf. on Commun. Control, & Comp.*, (Monticello, IL), 1999.

[25] B. Prabhakar, E. Uysal-Biyikoglu, and A. E. Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *Proc. IEEE Infocom 2001*, 2001.

[26] V. Goyal, A. Kumar, and V. Sharma, "Power constrained and delay optimal policies for scheduling transmission over a fading channel," in *Proc. of IEEE INFOCOM 2003*, (San Francisco), March 30-April 3 2003.

[27] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. on Information Theory*, vol. 48, pp. 1135–1149, May 2002.

[28] A. Fu, E. Modiano, and J. Tsitsiklis, "Optimal energy allocation and admission control for communicaitons satellites," *IEEE Trans. on Networking*, vol. 11, pp. 488–501, June 2003.

[29] R. Berry, "Optimal power/delay trade-offs in wireless communications - small delay asymptotics," in *Proc. of IASTED CIIT*, Scottsdale, AZ, Nov 2003. See also "Order Optimal Energy Efficient Transmission Policies in the Small Delay Regime," *Proc. of 2004 IEEE ISIT*, Chicago, IL, June 2004.

[30] R. Berry, "Communication over fading channels with finite buffer constraints - single user and multiple access cases," in *Proc. 2001 IEEE International Symposium on Info. Th.*, (Washington, DC), June 24-29 2001.

Biographies:

Randall Berry received the B.S. degree in Electrical Engineering from the University of Missouri-Rolla in 1993 and the M.S. and PhD degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology in 1996 and 2000 respectively. He is currently an assistant professor in the Department of Electrical and Computer Engineering at Northwestern University. In 1998 he was on the technical staff at MIT Lincoln Laboratory in the Advanced Networks Group. His primary research interests include wireless communication, data networks, and information theory. He is the recipient of a 2003 NSF CAREER award.

Edmund Yeh received his Bachelor of Science in Electrical Engineering with Distinction from Stanford University in 1994. In the same year, he was awarded one of ten Winston Churchill Scholarships for overseas study in science and engineering at Churchill College, University of Cambridge, in the United Kingdom. He received his Master of Philosophy in Electrical Engineering from the University of Cambridge in 1995. As a doctoral student, he studied under Professor Robert Gallager at the Laboratory for Information and Decision Systems at the Massachusetts Institute of Technology. He received his Ph.D. in Electrical Engineering from MIT in 2001. Since July 2001, he has been an Assistant Professor of Electrical Engineering and Computer Science at Yale University, New Haven, Connecticut. Dr. Yeh is a recipient of the Army Research Office (ARO) Young Investigator Program (YIP) Award. He received the National Science Foundation and Office of Naval Research Fellowships for graduate study, and is a member of Phi Beta Kappa, Tau Beta Pi, and IEEE. He has spent a number of summers working at industrial labs, including the Mathematical Sciences Research Center at Bell Laboratories, Murray Hill, New Jersey.