Opportunistic Beamforming using Dumb Antennas

Pramod Viswanath, David N.C. Tse and Rajiv Laroia*

September 11, 2001

Abstract

Multiuser diversity is a form of diversity inherent in a wireless network, provided by independent time varying channels across the different users. The diversity benefit is exploited by tracking the channel fluctuations of the users and scheduling transmissions to users when their instantaneous channel quality is near the peak. The diversity gain increases with the dynamic range of the fluctuations and is thus limited in environments with little scattering and/or slow fading. In such environments, we propose the use of multiple transmit antennas to artificially induce large and fast channel fluctuations so that multiuser diversity can still be exploited. The scheme can be interpreted as *opportunistic beamforming* and we show that *true* beamforming gains can be achieved when there are sufficient users, even though very limited channel feedback is needed. Furthermore, in a cellular system, the scheme plays an important and dual role of *opportunistic nulling* of the interference created on users of adjacent cells. We discuss the design implications of implementing this scheme in a complete wireless system.

1 Introduction

A fundamental characteristic of the wireless channel is the fading of the channel strength due to constructive and destructive interference between multipaths. An important means to cope with channel fading is the use of *diversity*. Diversity can be obtained over time (interleaving of coded bits), frequency (combining of multipaths in CDMA systems) and space (multiple antennas). The basic idea is to improve performance by creating several independent signal paths between the transmitter and the receiver.

These diversity modes pertain to a point-to-point link. Recent results point to another form of diversity, inherent in a wireless network with multiple users. This *multiuser diversity* is

*Pramod Viswanath is with the Department of Electrical and Computer Engineering at the University of Illinois, Urbana-Champaign, IL 61801. (Email: pramodv@uiuc.edu) David Tse is with the Department of EECS, University of California, Berkeley, CA 94720. (Email: dtse@eecs.berkeley.edu) Rajiv Laroia is with Flarion Technologies, Bedminster, NJ 07921. This work was initiated when the first author was with Flarion Technologies and the second author was visiting there.

best motivated by an information theoretic result of Knopp and Humblet [10]. They focused on the uplink in the single cell, with multiple users communicating to the basestation via time-varying fading channels, which is assumed to be tracked at the receiver and information fed back to the transmitters. To maximize the total information theoretic capacity, they showed that the optimal strategy is to schedule at any one time only the user with the best channel to transmit to the basestation. Diversity gain arises from the fact that in a system with many users whose channels have the same statistics but vary *independently*, there is likely to be a user whose channel is near its peak at any one time. Overall system throughput is maximized by allocating at any time the common channel resource to the user that can best exploit it. It can also be thought of as a form of *selection diversity*. Similar results are obtained for the downlink from the basestation to the mobile users [20]. A scheduling algorithm exploiting the multiuser diversity benefits is implemented in the downlink of IS-856 [18] (previously known as HDR: High Data Rate) system, where each user measures its downlink signal-to-interference ratio (SIR) based on a common pilot and feeds back the information to the basestation [21].

Traditionally, channel fading is viewed as a source of *unreliability* that has to be *mitigated*. In the context of multiuser diversity, however, fading can instead be considered as a source of *randomization* that can be *exploited*. This is done by scheduling transmissions to users only when their channels are near their peaks. The larger the dynamic range of the channel fluctuations, the higher the peaks and the larger the multiuser diversity gain. In practice, such gains are limited in two ways. First, there may be a line-of-sight path and little scattering in the environment, and hence there is a small dynamic range of channel fluctuations. Second, the channel may fade very slowly compared to the delay constraints of the application so that transmissions cannot wait until the channel reaches its peak. Effectively, the dynamic range of channel fluctuations is small within the time scale of interest. Both are important sources of hindrance to implementing multiuser diversity in a real system.

In this paper, we propose a scheme which artificially induces random fading when the environment has little scattering and/or the fading is slow. We focus on the downlink of a cellular system. We use multiple antennas at the basestation to transmit the same signal from each antenna modulated by a gain whose phase and magnitude is changing in time in a controlled but pseudo random fashion. The gains in the different antennas are varied independently. Channel variation is artificially induced through the constructive and destructive addition of signal paths from the multiple transmit antennas to the (single) receive antenna of each user. The overall (time varying) channel SIR is tracked by each user and is fed back to the basestation to form a basis for scheduling. The channel tracking is done via a single pilot signal which is repeated at the different transmit antennas, just like the data.

If the magnitudes and phases of the channel gains from all of the transmit antennas to the user can be tracked and fed back, then *transmit beamforming* can be performed by matching the powers and phases of the signals sent on the antennas to the channel gains in order to maximize the received SIR at the mobile. With a much more limited feedback of only the overall channel SIR, true beamforming cannot be performed. However, in a large system with many independently fading users, there is likely to be a user whose instantaneous channel gains are close to matching the current powers and phases allocated at the transmit

antennas. Viewed in this light, our scheme can be interpreted as performing *opportunistic beamforming*: the transmit powers and phases are randomized and transmission is scheduled to the user which is close to being in the beamforming configuration.

Recently there has been significant amount of work in the use of multiple transmit antennas in wireless communications (also called space-time codes, eg. [6, 14, 1, 16]). Performance gain over single-antenna system is achieved by *smart* coding and signal processing at the transmitter and the receiver. In contrast, our scheme uses the multiple transmit antennas in a *dumb* way: no additional processing at the transmitter nor the receiver is needed beyond that in a single antenna system. There is no need to change the modulation format nor have additional pilots to measure the channels from individual transmit antennas. In fact, the receiver is oblivious to the existence of multiple transmit antennas. This makes it particularly easy to upgrade existing systems to implement such a scheme, since only additional antennas has to be placed at the basestation but the mobile handsets need not be changed at all. The opportunistic beamforming scheme does need tight feedback of overall channel SIR measurements and rate adaptation, but we note that such mechanisms already exist in third-generation systems and beyond.

Earlier works have proposed the use of intentional frequency offset at the transmit antennas to create a fast fading environment [7, 8, 9]. The goal is to increase the time diversity of slow fading point-to-point links, but in that context this scheme has been shown to be inferior compared to other space-time coding techniques such as orthogonal design [1, 16]. Our work, in contrast, shifts the point-to-point view to the multiuser view, and we show that when such channel randomization is used in conjunction with multiuser diversity scheduling, the achieved performance can significantly surpass space-time codes.

The outline of the paper is as follows. In Section 2 we review the multiuser diversity concept and discuss its implementation in the downlink of IS-856 system. We introduce the idea of opportunistic beamforming in Section 3 and study its performance in slow and fast fading environments. In Section 4, we compare the opportunistic beamforming technique with other proposed ways to use multiple transmit antennas. An information theoretic comparison is undertaken in Appendix B. In Sections 5 and 6, we explore the role of opportunistic beamforming in wide band and cellular environments. It turns out that in the cellular context, the proposed technique plays an important and dual role of *opportunistic nulling* of interference caused in adjacent cells. Section 7 discusses various system and implementation issues. We distill some of the key ideas of this paper into Section 8 which also contains our conclusions.

2 Multiuser Diversity and Fair Scheduling

2.1 Multiuser Diversity

We begin with a simple model of the downlink of a wireless communication system. There is a single base-station (transmitter) communicating with K users (receivers). The baseband

time-slotted block fading channel model is given by:

$$y_k(t) = h_k(t)x(t) + z_k(t) \qquad k = 1, 2, \ldots, K \tag{1}$$

where $x(t) \in \mathcal{C}^T$ is the vector of transmitted symbol at time-slot slot t, $y_k(t) \in \mathcal{C}^T$ is the received signal of user k at time-slot, $h_k(t) \in \mathcal{C}$ is the fading channel gain from the transmitter to receiver k at in time-slot t, and $\{z_i(t)\}_t$ is a i.i.d. sequence of zero mean circular symmetric Gaussian random vectors $\mathcal{CN}(0, \sigma^2 I_T)$. We are assuming here that the bandwidth is narrow enough so that the channel response is flat across the whole band. We are also assuming a block fading model where the channel is assumed to be constant over time-slots of length T samples. We are also assuming that the transmit power level is fixed at P at all times, i.e. $E[|x(t)|^2] = PT$. This is a reasonable power constraint for the base-station.

If we assume that both the transmitter and the receivers can perfectly track the fading processes $\{h_k(t)\}$, then we can view this downlink channel as a set of parallel Gaussian channels, one for each fading state. The sum capacity of this channel, defined by the maximum achievable sum of long-term average data rates transmitted to all the users, can be achieved by a simple TDMA strategy: at each fading state, transmit to the user with the strongest channel [20].

In Fig. 1, we plot the sum capacity (in total number of bits per second per Hz) of the downlink channel as a function of the number of users, for the case when users undergo independent Rayleigh fading with average received SNR = 0 dB. We observe the the sum capacity increases with the number of users in the system. In contrast, the sum capacity of a non-faded downlink channel, where each user has a *fixed* AWGN channel with SNR = 0 dB, is constant irrespective of the number of users. Somewhat surprisingly, with moderate number of users, the sum capacity of the fading channel is greater than the that of a non-faded channel. This is the *multiuser diversity* effect: in a system with many users with independently varying channels, it is likely that at any time there is a user with channel much stronger than the average SNR. By transmitting to users with strong channel at all times, the overall spectral efficiency of the system can be made high, significantly higher than that of a non-faded channel with the same average SNR.

The system requirements to extract such multiuser diversity benefits are:

- each receiver tracking its own channel SIR, through say a common downlink pilot, and feeding back the instantaneous channel quality to the base-station;
- the ability of the base-station to schedule transmissions among the users as well as to adapt the data rate as a function of the instantaneous channel quality.

These features are already present in the designs of many 3G systems, such as IS-856 ([2]).

2.2 Proportional Fair Scheduling

To implement the idea of multiuser diversity in a real system, one is immediately confronted with two issues: fairness and delay. In the ideal situation when users' fading *statistics* are

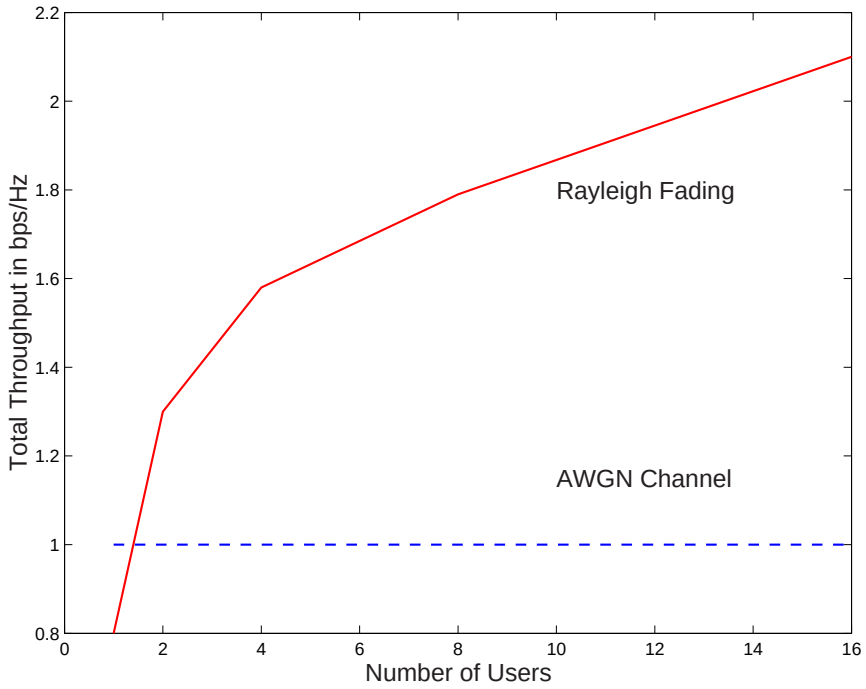

Figure 1: Sum Capacity of two channels, Rayleigh fading and AWGN, with average SNR = 0dB.

the same, the strategy above maximizes not only the total capacity of the system but also the throughput of individual users. In reality, the statistics are not symmetrical; there are users who are closer to the base-station with a better average SNR; there are users who are stationary and some that are moving; there are users which are in a rich scattering environment and some with no scatterers around them. Moreover, the strategy is only concerned with maximizing long-term average throughputs; in practice there are latency requirements, in which case the average throughputs over the delay time-scale is the performance metric of interest. The challenge is to address these issues while at the same time exploiting the multiuser diversity gain inherent in a system with users having independent, fluctuating channel conditions.

A simple scheduling algorithm has been designed to meet this challenge [21]. This work is done in the context of the downlink of IS-856 system, operating on a 1.25 MHz IS-95 bandwidth. In this system, the feedback of the channel quality of user k at time slot t to the base-station is in terms of a requested data rate $R_k(t)$: this is the data rate that the kth user's channel can currently support. The scheduling algorithm works as follows. It keeps track of the average throughput $T_k(t)$ of each user in a past window of length t_c. (This can be done using an exponential weighted low-pass filter, for example.) At time slot t, the scheduling algorithm simply transmits to the user k^* with the largest

$$\frac{R_k(t)}{T_k(t)}$$

among all active users in the system.

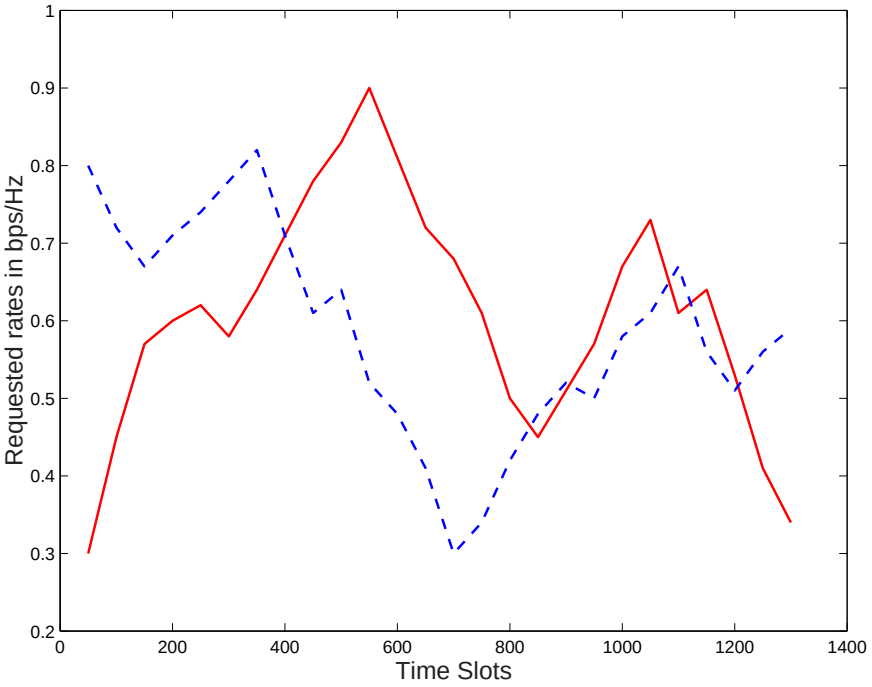

Figure 2: For symmetric channel statistics of users, the scheduling algorithm reduces to serving each user with the largest requested rate.

One can get an intuitive feel of how this algorithm works by inspecting Figures 2 and 3. We plot the sample paths of the requested data rates of two users as a function of time slots (each time slot is 1.67ms in IS-856). In Fig 2, the two users have identical fading *statistics*. If the scheduling time-scale t_c is much larger than the correlation time-scale of the fading dynamics, then by symmetry the throughput of each user $T_k(t)$ converges to the same quantity. The scheduling algorithm reduces to always picking the user with the highest requested rate. Thus, each user is scheduled when their channel is good and at the same time the scheduling algorithm is perfectly fair on the long term. In Fig 3, one user's channel is much stronger than the other user on the average, although both channels fluctuate due to multipath fading. Always picking the user with the highest requested rate means giving all the system resources to the statistically stronger user, and would be highly unfair. In contrast, under the proposed scheduling algorithm, users compete for resources not directly based on their requested rates but only after normalization by their respective average throughputs. The user with the statistically stronger channel will have a higher average throughput. Thus, the algorithm schedules a user when its instantaneous channel quality is high *relative* to its own average channel condition over the time-scale t_c. In short, data is transmitted to users when their channels are *near the peaks*. Multiuser diversity benefit can still be extracted because channels of different users fluctuate independently so that if there is a sufficient number of users in the system, there will likely be a user near its peak at any one time.

The parameter t_c is tied to the latency time-scale of the application. Peaks are defined with respect to this time-scale. If the latency time-scale is large, then the throughput is averaged

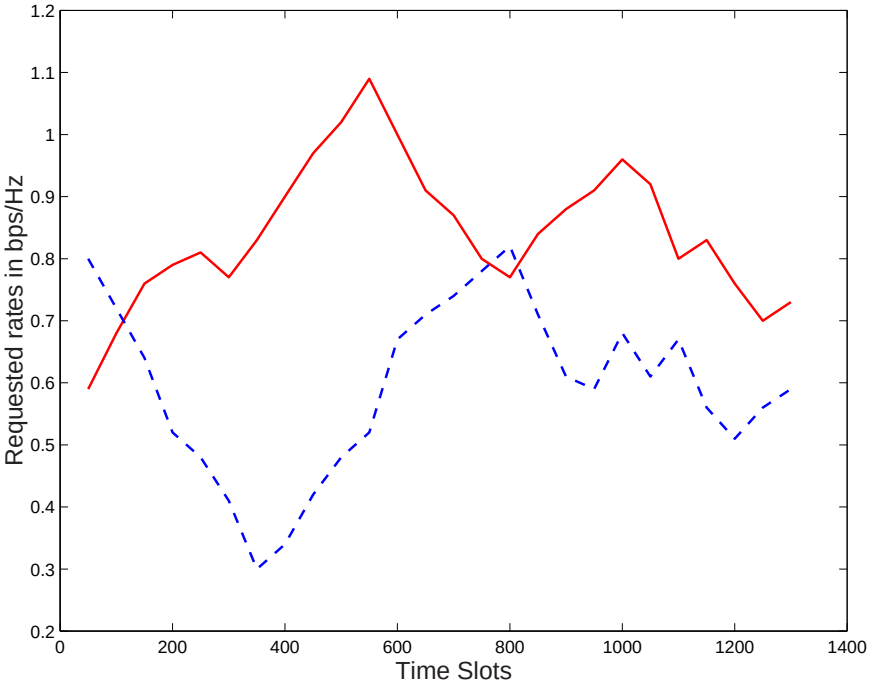

Figure 3: In general, with asymmetric user channel statistics, the scheduling algorithm serves each user when it is near its peak within the latency time-scale t_c.

over a longer time-scale and the scheduler can afford to wait longer before scheduling a user when its channel hits a really high peak.

The theoretical properties of this scheduling algorithm are further explored in [21]. There it is shown that this algorithm guarantees a fairness property called *proportional fair*. This property is further discussed in the appendix.

2.3 Limitation of Multiuser Diversity Gain

Figure 4 gives some insights on the issues involved in realizing multiuser diversity benefits in practice. The plot shows the total throughput of the HDR downlink under the proportional fair scheduling algorithm in two simulated environments:

- fixed: users are fixed but there are movements of objects around them (2 Hz Rician, ($\kappa \stackrel{\text{def}}{=} E_{\text{direct}}/E_{\text{specular}} = 5$);
- mobile: users move at walking speeds (3 km/hr, Rayleigh) ;

The total throughput increases with the number of users in both the fixed and mobile environments, but the increase is more dramatic in the mobility case. While the channel fades in both cases, the dynamic range and the rate of the variations is larger in the mobile environment than in the fixed one. This means that over the latency time-scale (1.67s in

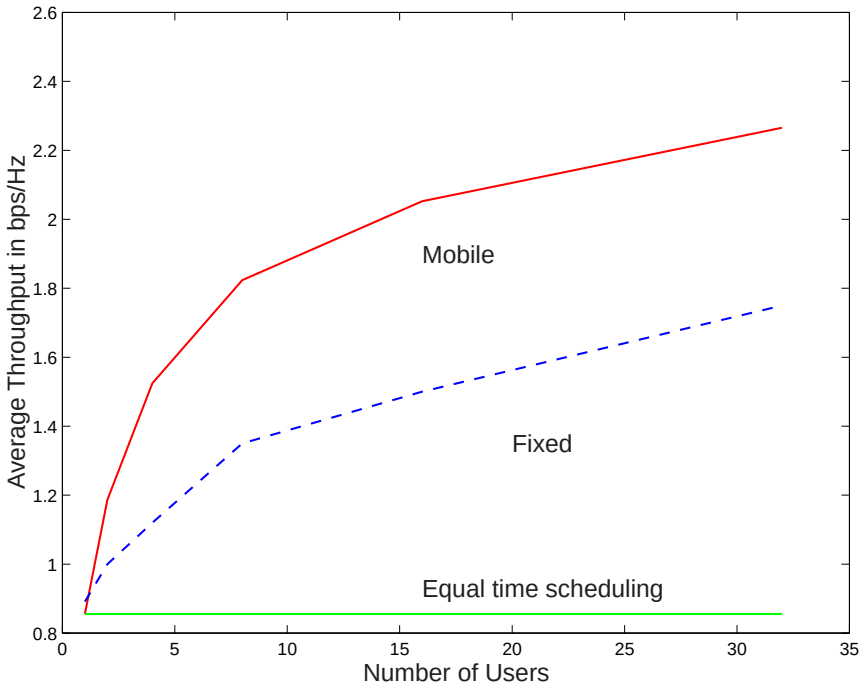

Figure 4: Multiuser diversity gain in fixed and mobile environments.

these examples) the peaks of the channel fluctuations are likely to be higher in the mobile environment, and the peaks are what determine the performance of the scheduling algorithm. Thus, the inherent multiuser diversity is more limited in the fixed environment.

3 Opportunistic Beamforming

The amount of multiuser diversity depends on the rate and dynamic range of channel fluctuations. In environments where the the channel fluctuations are small, a natural idea comes to mind: why not amplify the multiuser diversity gain by *inducing* faster and larger fluctuations *artificially*? Our technique is to use multiple transmit antennas at the base station as illustrated in Figure 5.

Consider a system with N transmit antennas at the base-station. Let $h_{nk}(t)$ be the complex channel gain from antenna n to the kth user at time t. At time slot t, the same block of symbols $x(t)$ is transmitted from all of the antennas except that it is multiplied by a complex number $\sqrt{\alpha_n(t)}e^{j\theta_n(t)}$ at antenna n, for $n = 1 \ldots N$, such that $\sum_{n=1}^{N} \alpha_n(t) = 1$, preserving the total power transmitted. The received signal at user k (recall (1) for a comparison) is given by :

$$y_k(t) = \left(\sum_{n=1}^{N} \sqrt{\alpha_n(t)} e^{j\theta_n(t)} h_{nk}(t) \right) x(t) + z_k(t) \tag{2}$$

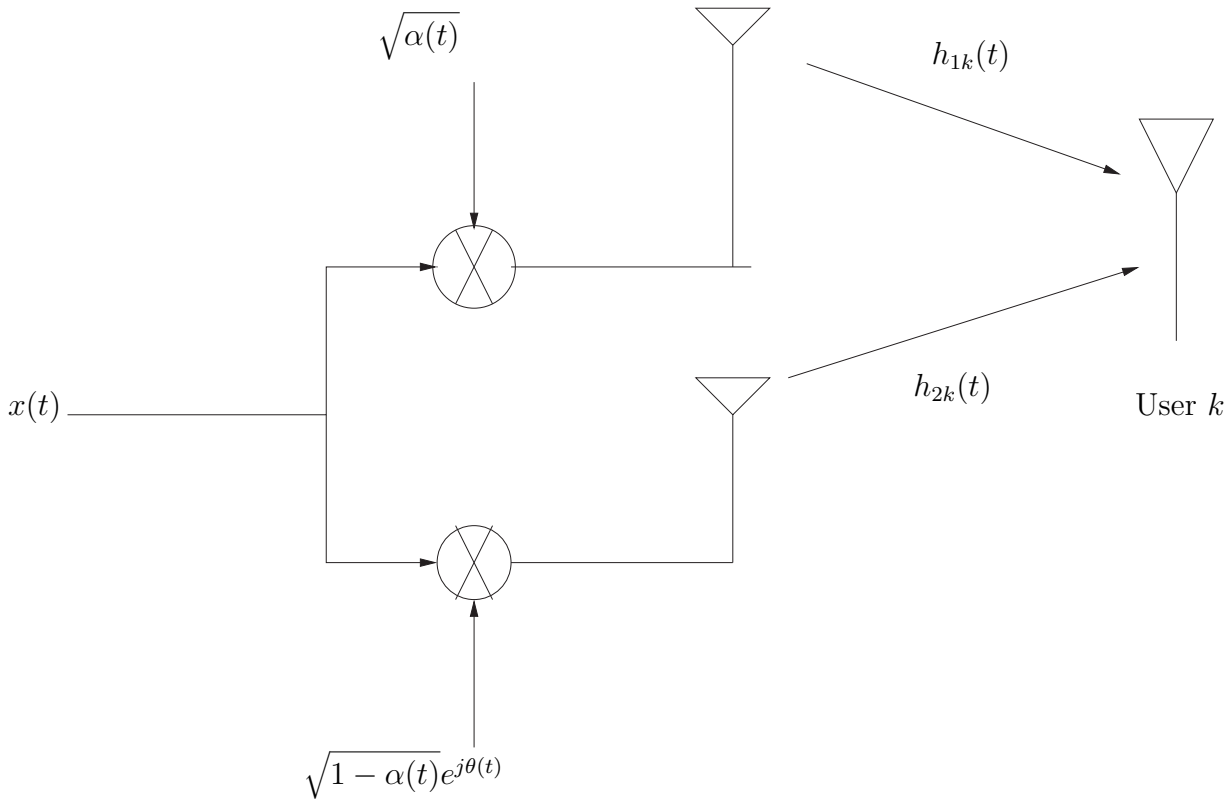

Figure 5: Same signal is transmitted over the two antennas with time varying phase and powers.

Thus, the overall channel gain seen by receiver k is now

$$h_k(t) := \sum_{n=1}^{N} \sqrt{\alpha_n(t)} e^{j\theta_n(t)} h_{nk}(t)$$

The $\alpha_n(t)$'s denote the fractions of power allocated to each of the transmit antennas, and the $\theta_n(t)$'s the phase shifts applied at each antenna to the signal. By varying these quantities over time ($\alpha_n(t)$'s from 0 to 1 and $\theta_n(t)$'s from 0 to 2π) , fluctuations in the overall channel can be induced even if the physical channel gains $h_{nk}(t)$ have very little fluctuations.

As in the single transmit antenna system, each receiver k feeds back the overall SNR $|h_k(t)|^2/\sigma^2$ of its own channel to the base-station (or equivalently the data rate that the channel can currently support) and the base-station schedules transmissions to users accordingly. There is no need to measure the individual channel gains $h_{nk}(t)$ (phase or magnitude); in fact, the existence of multiple transmit antennas is completely transparent to the receiver. Thus only a single pilot signal is needed for channel measurement (as opposed to a pilot to measure each antenna gain). The pilot symbols are repeated at each transmit antenna, exactly like the data symbols.

The rate of variation of $\{\alpha_n(t)\}$ and $\{\theta_n(t)\}$ in time is a design parameter of the system. We would like it to be as fast as possible to provide full channel fluctuations within the latency time scale of interest. On the other hand, there is a practical limitation to this fast variation of phase and power. The variation should be slow and happen at a time scale that allows the channel to be reliably estimated by the users and the SNR fed back. Further, the variation should be slow enough to ensure that the channel seen by the users does not change abruptly and thus maintains stability of the channel tracking loop.

To get more insights into the performance of this scheme, we will study the cases of slow fading and fast fading separately. In the analysis below, we will assume that the variations in $\{\alpha_n(t)\}$ and $\{\theta_n(t)\}$ are performed in such a way that the overall channel can be tracked and fed back perfectly by the receivers to the transmitter.

3.1 Slow Fading

Consider the case of slow fading where the channel gain of each user $h_{nk}(t) = h_{nk}$ remain constant for all t. (In practice, this means for all t over the latency time-scale of interest.) Then the received SNR for this user would have remained constant if only one antenna were used. If all users in the system experiences such slow fading, no multiuser diversity gain could have been exploited. Under the proposed scheme, on the other hand, the overall channel gain $h_k(t)$ for each user k varies in time and provides opportunity for exploiting multiuser diversity gain.

Let us focus on a particular user k. Now, if each $\alpha_k(t)$ is varied in time from 0 to 1 and $\theta_k(t)$ from 0 to 2π, the amplitude squared of the channel $|h_k(t)|^2$ seen by user k varies from 0 to $\sum_{n=1}^{N} |h_{nk}|^2$. The peak value occurs when the power and phase values are in the *beamforming*

configuration:

$$\begin{aligned}\alpha_n &= \frac{\mid h_{nk} \mid^2}{\sum_{n=1}^{N} \mid h_{nk} \mid^2}, \qquad n = 1, \ldots, N \\ \theta_n &= -\arg(h_{nk}), \qquad n = 1, \ldots, N.\end{aligned}$$

To be able to beamform to a particular user, the base-station needs to know individual channel amplitude and phase responses from all the antennas, much more information to feedback than just the overall SNR. However, if there are many users in the system, the proportional fair algorithm will schedule transmission to a user only when its overall channel SNR is near its peak. Thus, it is plausible that in a slow fading environment, our proposed technique can approach the performance of coherent beamforming but with only overall SNR feedback. In this context, the technique can be interpreted as *opportunistic beamforming*: phases and power allocated at the transmit antennas are varied in a pseudo random manner, and at any time transmission is scheduled to the user which is currently closest to being in its beamforming configuration. The following formal result justifies our intuition.

Suppose the data rate achieved per time slot is a monotonically increasing function of the instantaneous SNR of a user. We assume that the power and phase variation processes $\{(\alpha_1(t), \ldots, \alpha_N(t))\}_t$ and $\{(\theta_1(t), \ldots, \theta_N(t))\}_t$ are stationary and ergodic. It is easily seen that under the proportional fair scheduling algorithm with $t_c = \infty$, the long-term average throughput of each user exists [21]. Denote the average throughput of user k in a system with K users to be $T_k^{(K)}$. Note that in general $T_k^{(K)}$ depends on the slow fading states $(h_{1i}, \ldots h_{Ni})$, $i = 1, \ldots K$ of all users, as well as the statistics of the power and phase variation processes. However, if the power and phase variation processes "match" the slow fading distribution of the users, we have the following asymptotic result for a large system with many users. We assume a discrete set of slow fading states to minimize the technicality of the proof, but extension to the continuous case should be possible.

Theorem 1 *Suppose the slow fading states of the users are independent and identically distributed and are discrete, and the joint stationary distribution of*

$$(\alpha_1(t), \ldots, \alpha_N(t), \theta_1(t), \ldots, \theta_N(t))$$

is the same as that of

$$\left(\frac{\mid h_{1k} \mid^2}{\sum_{n=1}^{N} \mid h_{nk} \mid^2}, \ldots, \frac{\mid h_{Nk} \mid^2}{\sum_{n=1}^{N} \mid h_{nk} \mid^2}, -arg(h_{1k}), \ldots, -arg(h_{Nk}) \right)$$

for the slow fading state of any individual user k. Then, almost surely, we have

$$\lim_{K \to \infty} K T_k^{(K)} = R_k^{\text{bf}}$$

for all k. Here, R_k^{bf} is the data rate that user k achieves when it is in the beamforming configuration, i.e., when its instantaneous SNR is

$$P \sum_{n=1}^{N} |h_{nk}|^2 / \sigma^2.$$

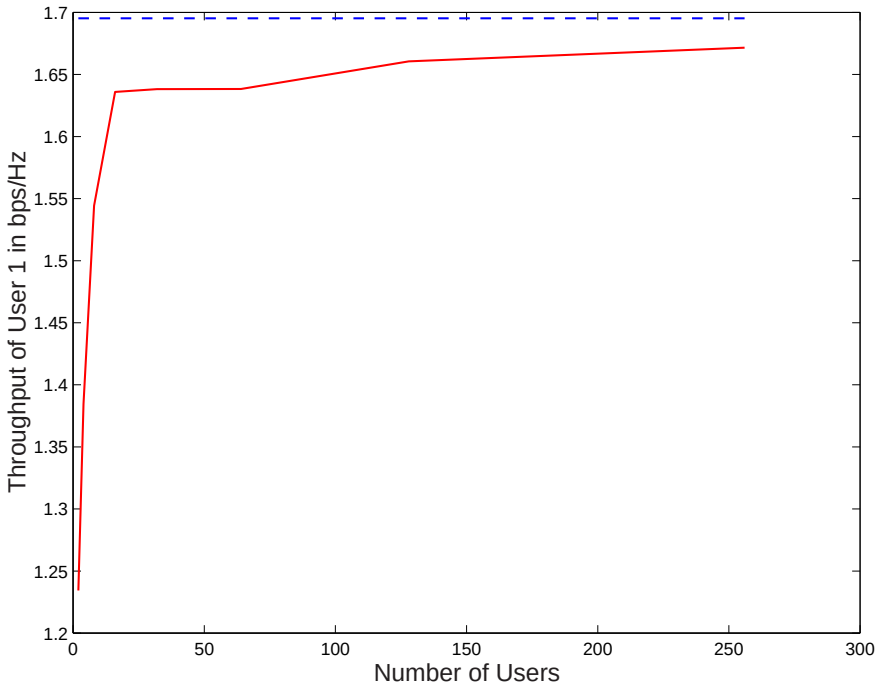

Figure 6: Throughput in bps/Hz for user 1 multiplied by number of users scheduled for slow Rayleigh fading at 0 dB SNR with the proportional fair scheduling algorithm. Performance of coherent beamforming for user 1 and scheduled at all time is plotted as a dotted line. We have chosen two antennas.

Proof

See Appendix A.

∘

This result implies that when there are many users, with high probability the proportional fair algorithm always schedules the users when they are in their respective beamforming configurations, and moreover allocates equal amount of time to each user.

To see how large the number of users has to be for this result to be valid, we have simulated the performance of the opportunistic beamforming scheme for two transmit antennas under a slow Rayleigh fading environment with average SNR = 0 dB. For this simulation experiment, we have assumed that reliable data rate depends on the SNR value as $\log(1 + \text{SNR})$. We perform two separate experiments and rotate the phases uniformly in $[0, 2\pi]$ and powers uniformly in $[0, 1]$ (and scaled such that they sum to unity). In the first, we generate the slow fading realizations (as i.i.d. Rayleigh distributed) for a large number of users (256 in the simulation example) and run the proportional fair scheduling algorithm on subsets of the users (2,4,8,16,64,128,256) and plot the throughput of user 1 (who is contained in each of the subsets) scaled by the number of users participating in that round of the scheduling algorithm. Fig 6 plots this throughput of user 1 for 2 antennas. Also plotted is the eventual limit promised by Theorem 1. In Fig 7 we repeat this experiment for 10 diversity antennas.

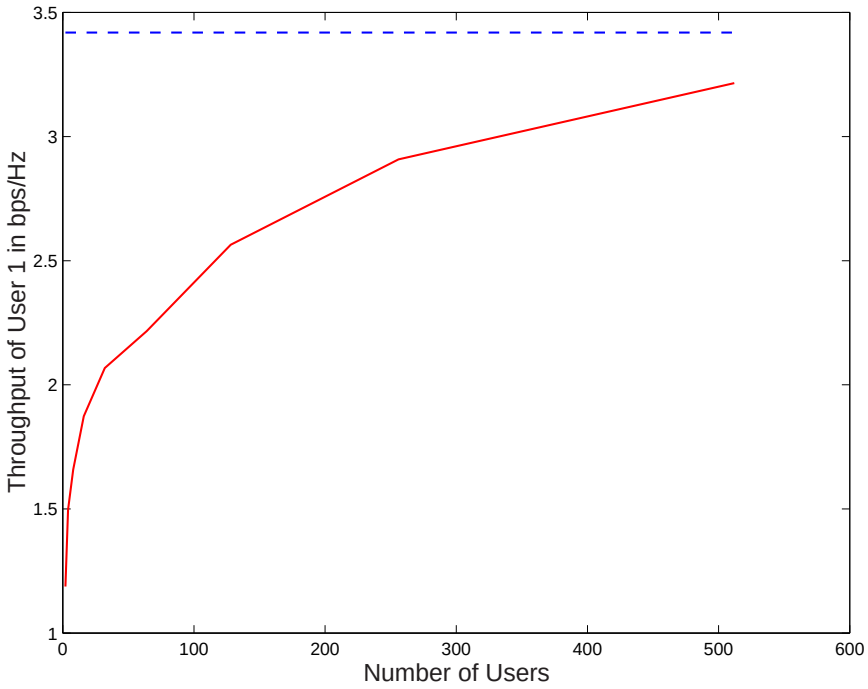

Figure 7: Throughput in bps/Hz for user 1 multiplied by number of users scheduled for slow Rayleigh fading at 0 dB SNR with the proportional fair scheduling algorithm. Performance of coherent beamforming for user 1 and scheduled at all time is plotted as a dotted line. There are 10 antennas in this experiment

The observation is that the convergence of the scaled throughput to the limit slows down as the number of antennas grows. In this experiment, the scaled throughput of user 1 is 40% away from its eventual limit even with 100 users in the system. Thus, to achieve close to asymptotic performance the number of users required grows rapidly with the number of antennas (the proof of Theorem 1 suggests that the number of users required grows exponentially with the number of antennas). We resume this topic and that of choosing the power and phase variation processes $\{\alpha_n(t)\}$ and $\{\theta_n(t)\}$ in Section 7 along with considerations of the impact on the system design.

In the second experiment, i.i.d. Rayleigh distributed realization of the slow fading coefficients are generated for each user. The average SNR is 0 dB. The power and phase allocations are then generated i.i.d. over the time slots. The total throughput of all users under the proportional fair algorithm with is noted. (Here we are assuming the use of powerful enough codes such that the data rate achieved in each time slot is given by the Shannon limit $\log(1 + SNR)$ per degree of freedom.) The total throughput is a function of the realization of the slow fading coefficients. The average total throughput is obtained by averaging over 300 realizations. This is plotted as a function of the number of users in the system in Figure 8. Note that there is almost a 100% improvement in throughput going from 1 user to 16 users. Also plotted is the performance under coherent beamforming and equal-time round robin scheduling. The total throughput is independent of the number of users in the system (the effective channel doesn't change, so there is no multiuser diversity gain.). We see

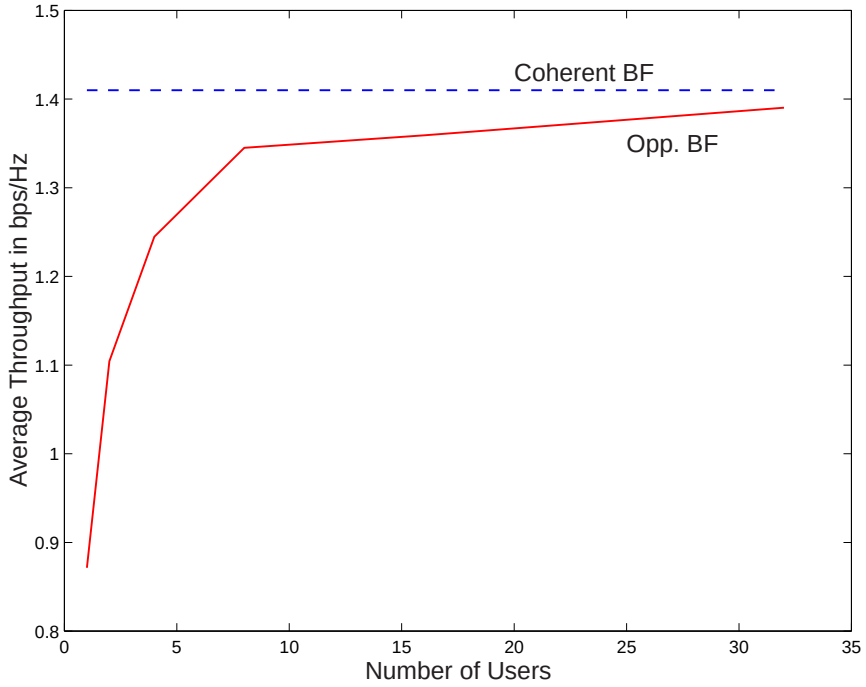

Figure 8: Total Throughput in bps/Hz averaged over slow Rayleigh fading at 0 dB SNR with the proportional fair scheduling algorithm. Performance of coherent beamforming is also plotted.

that for 16 users, opportunistic beamforming already achieves the performance of coherent beamforming.

Theorem 1 says that the asymptotic performance in a large system does not depend on the distribution of these processes as long as there is a non-zero probability of being close to the beamforming configuration of each user. We saw that with a small number of transmit antennas (two in the simulation example of Fig. 8) the asymptotic performance is achieved with a small number of users (16 users in the simulation example of Fig. 8).

3.2 Fast Fading

We see that opportunistic beamforming can significantly improve performance in slow fading environments by adding fast time-scale fluctuations on the overeall channel quality. The rate of channel fluctuation is artificially sped up. Can opportunistic beamforming help if the underlying channel variations are already fast (fast compared to the latency time-scale)?

For simplicity, let us focus on the symmetric case when the fading statistics of the users are identical. (The situation in the asymmetric case is similar.) Suppose the channel gains $\{(h_{1k}(t), \ldots, h_{Nk}(t)\})$ are stationary and ergodic over time for each user k and independent across users. Let us assume that the power and phase variation processes are stationary and

ergodic as well. The overall channel gain process

$$h_k(t) := \sum_{n=1}^{N} \sqrt{\alpha_n(t)} e^{j\theta_n(t)} h_{nk}(t)$$

has the same statistics for all users and at time t the proportional fair scheduling algorithm simply transmits to the user with the highest $|h_k(t)|^2$. (Here we are assuming that the latency time-scale t_c is set to be ∞.). The throughput achieved is

$$E\left[f\left(\max_{k=1\ldots K} |h_k|^2/\sigma^2\right)\right] \tag{3}$$

where f denotes the function that maps the received SNR value to the reliable data rate of transmission (in the simulation example earlier, we had taken $f(x)$ to be $\log(1+x)$). Here the integration is over the stationary distribution of the process $\{(h_1(t), \ldots, h_K(t))\}$. The impact of opportunistic beamforming in the fast fading scenario then depends on how the stationary distributions of the overall channel gains can be modified by power and phase randomization. Intuitively, better multiuser diversity gain can be exploited if the dynamic range of the distribution of h_k can be increased, so that the maximum SNRs can be larger. We consider a few examples of common fading models.

3.2.1 Independent Rayleigh fading

In this model appropriate for an environment where there is full scattering and the transmit antennas are spaced sufficiently apart, the channel gains $h_{1k}(t), \ldots, h_{Nk}(t)$ are independent, identically distributed circular symmetric Gaussian random variables. It can be seen that in this case, $h_n(t)$ has exactly the same distribution as each of the individual gains $h_{nk}(t)$. Thus, in an independent fast Rayleigh fading environment, the opportunistic beamforming technique does not provide any performance gain.

3.2.2 Independent Rician fading

Rician fading models the situation where there is a direct line-of-sight component which is not time-varying:

$$h_{nk}(t) = \sqrt{a}\exp(j\phi_{nk}) + b_{nk}(t),$$

where a is a constant, ϕ_{nk} are uniformly and independently distributed phases but fixed over time, and $b_{nk}(t)$ are i.i.d. $\mathcal{CN}(0, v)$ random variables representing the time-varying diffused component of the fading. The first term is the direct component, differing only in a shift of phases for each of the transmit antennas. (We are assuming that the received energy of the direct component is the same from all the transmit antenna to a given user.) The κ-factor is the ratio of the energy in the direct component to that in the diffused component:

$$a = \frac{\kappa}{1+\kappa}, \qquad v := E(|b_{nk}(t)|^2) = \frac{1}{1+\kappa}.$$

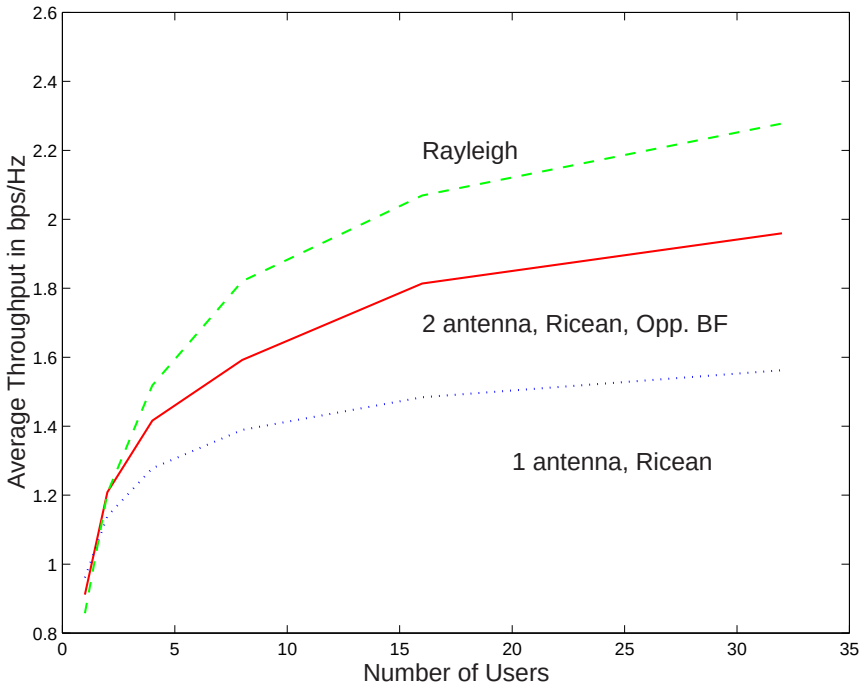

Figure 9: Total throughput as a function of the number of users under Rician fading, with and without opportunistic beamforming. The power allocation $\alpha_n(t)$'s are uniformly distributed in $[0, 1]$ and the phases $\theta_n(t)$'s uniform in $[0, 2\pi]$.

In contrast to the Rayleigh fading case, opportunistic beamforming has a significant impact in a Rician environment, particularly when the κ-factor is large. In this case, the scheme can significantly increase the dynamic range of the fluctuations. This is because the fluctuations in the underlying Rician fading process come from the diffused component, while with randomization of phase and powers, the fluctuations are from the coherent addition and cancellation of the direct path components in the signals from the different transmit antennas, in addition to the fluctuation of the diffused components. If the direct path is much stronger than the diffused part (large κ values), then much larger fluctuations can be created by this technique.

This intuition is substantiated in Figure 9, which plots the total throughput for Rician fading with $\kappa = 10$. We see that there is much improvement in performance going from the single transmit antenna case to dual transmit antennas with opportunistic beamforming. For comparison, we also plot the analogous curves for pure Rayleigh fading; as expected, there is no improvement in performance in this case. Figure 10 compares the stationary distributions of the overall channel gain $h_k(t)$ in the single antenna and dual antenna cases; one can see the increase in dynamic range due to opportunistic beamforming.

More insights into the nature of the performance gain can be obtained by an asymptotic analysis in the limit of large number of users. The key quantity of interest (c.f. eqn.3) is the random variable:

$$g_K := \max_{k=1\ldots K} |h_k|^2$$

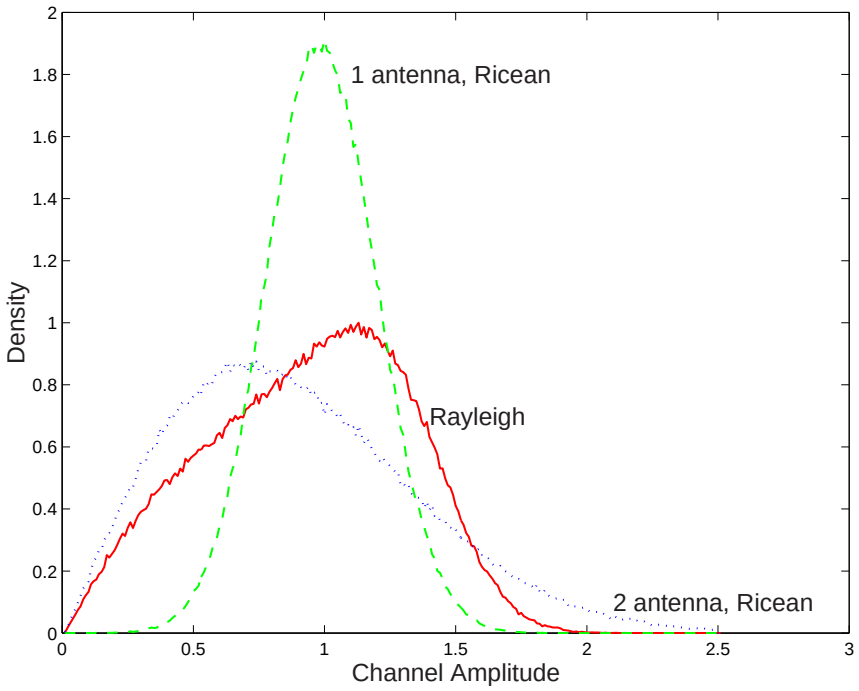

Figure 10: Comparison of the distribution of the overall channel gain with and without opportunistic beamforming using two transmit antennas, Rician fading.

where h_k is the overall channel gain to user k. For large k, the distribution of g_k depends only on the tail behavior of the distribution of the individual $|h_k|^2$. In all cases of interest, $|h_k|^2$ has an exponential tail, in which case the limiting distribution of g_k can be computed based on the following result. (p. 207 of [5])

Lemma 2 *Let $z_1, \ldots, z_K$ be i.i.d. random variables with a common cdf $F(\cdot)$ and pdf $f(\cdot)$ satisfying $F(z)$ is less than 1 for all z and is twice differentiable for all z, and is such that*

$$\lim_{z \to \infty} \left[\frac{1 - F(z)}{f(z)} \right] = c > 0 \tag{4}$$

for some constant c. Then

$$\max_{1 \le k \le K} z_k - l_K$$

converges in distribution to a limiting random variable with cdf

$$\exp(-e^{-x/c}).$$

In the above, l_K is given by $F(l_K) = 1 - 1/K$.

This result states that the maximum of K such i.i.d. random variables grows like l_K.

Let us first consider the case when the h_k's are i.i.d. Rayleigh, the magnitude $|h_k|^2$ is exponentially distributed with mean 1. Condition 4 is satisfied (and in fact $(1 - F(u))/f(u) = 1$

for every $u > 0$). The constants $l_K = \log K$, and hence the gain of the strongest user grows like $\log K$.

In the case when the h_k's are Rician (i.e, single transmit antenna case), the tail of the cdf and pdf of $|h_k|^2$ can be calculated to be:

$$\begin{aligned} 1 - F(z) &\sim \frac{1}{2}\sqrt{\frac{v}{\pi}}(az)^{-1/4}\exp\left(\frac{-(\sqrt{z}-\sqrt{a})^2}{v}\right), \\ f(z) &\sim \frac{1}{2}\frac{1}{\sqrt{v\pi}}(az)^{-1/4}\exp\left(\frac{-(\sqrt{z}-\sqrt{a})^2}{v}\right), \end{aligned}$$

where the approximations are in the sense that the ratio of the left and right hand sides approach 1 as $z \to \infty$. Hence

$$\lim_{z\to\infty}\left[\frac{1-F(z)}{f(z)}\right] = \frac{1}{v}$$

and condition (4) is satisfied. Solving $F(l_K) = 1 - 1/K$ yields

$$l_K = \left(\sqrt{v\log K} + \sqrt{a}\right)^2 + O(\log\log K). \tag{5}$$

Intuitively, this expression says that in a large system, the user who has the strongest gain is one whose diffused component magnitude is the strongest among all users *and* whose diffused and fixed components are in phase.

Comparing to the Rayleigh case, we see that the leading term in the gain of the strongest user is now $v\log K = \frac{1}{1+\kappa}\log K$ instead of $\log K$, reduced by a factor of $1/(1+\kappa)$.

What is the effect of opportunistic beamforming? The overall gain for user k is:

$$\begin{aligned} |h_k|^2 &= \left|\sqrt{a}\sum_{n=1}^{N}\sqrt{\alpha_n}\exp[j(\theta_n+\phi_{nk})] + \sum_{n=1}^{N}\sqrt{\alpha_n}\exp(j\theta_n)b_{nk}\right|^2 \\ &= \left|\sqrt{a}\sum_{n=1}^{N}\sqrt{\alpha_n}\exp[j(\theta_n+\phi_{nk})] + c_k\right|^2 \end{aligned}$$

where c_k is $\mathcal{CN}(0,v)$, the c_k's are independent and independent of α_n's and θ_n's. The largest possible value for the term

$$\left|\sqrt{a}\sum_{n=1}^{N}\sqrt{\alpha_n}\exp[j(\theta_n+\phi_{nk})]\right|$$

is $\sqrt{Na}$, when the power and phase allocations are in beamforming configuration with respect to the fixed component of the channel gain of a user. Assume the phase and power distributions are uniform. In a large system, for any fixed $\delta > 0$ and every $\epsilon \in (0,1)$, almost surely ϵ of users at every time instant (the users constituting the fraction could change from time to time) for which

$$\left|\sqrt{a}\sum_{n=1}^{N}\sqrt{\alpha_n}\exp[j(\theta_n+\phi_{nk})]\right| > \sqrt{Na} - \delta$$

These ϵK users can be thought of as experiencing Rician fading with norm of the fixed component close to $\sqrt{Na}$. Using eqn. (5), the maximum of the gains $|h_k|^2$ among these ϵK users grows at least as fast as:

$$\begin{aligned}
& \left(\sqrt{v\log(\epsilon K)} + \sqrt{Na} - \delta\right)^2 + O(\log\log K) \\
= & \left(\sqrt{v\log K + v\log\epsilon} + \sqrt{Na} - \delta\right)^2 + O(\log\log K) \\
= & \left(\sqrt{v\log K + v\log\epsilon} + \sqrt{Na} - \delta\right)^2 + O(\log\log K) \\
= & \left(\sqrt{v\log K} + \sqrt{Na} - \delta\right)^2 + O(\log\log K)
\end{aligned}$$

as $K \to \infty$, for fixed $\epsilon, \delta > 0$. Since this is true for any $\delta > 0$ and for a subset of the users, we conclude that a lower bound on the growth rate of $\max_{1\le k\le K} |h_k|^2$ is

$$\left(\sqrt{v\log K} + \sqrt{Na}\right)^2 + O(\log\log K) \tag{6}$$

This growth rate can be interpreted as attained by the ideal situation when all users are simultaneously at the beamforming configurations of their fixed component *and* the resulting fixed component is in phase with the diffused component for every user. Using this interpretation and by a simple coupling argument, (6) can also be shown to be an upper bound to the growth rate. Thus, the growth rate under opportunistic beamforming is given by

$$\left(\sqrt{v\log K} + \sqrt{Na}\right)^2 + O(\log\log K)$$

Intuitively, one can interpret this result as saying that the user with the strongest channel is the one simultaneously having the strongest diffused component among all users, the fixed component in a beamforming configuration, and the diffused and fixed components in phase. Compared to the case with single transmit antenna, opportunistic beamforming increases the effective magnitude of the fixed component from $\sqrt{a}$ to $\sqrt{Na}$. While this does not increase the leading term in the growth rate ($v\log K$), it does increase the second term, of order $\sqrt{\log K}$.

While the above analysis assumes that the fixed component is from a line-of-sight path, it is also applicable to the case when the fixed component arises from slow fading. This models for example the situation when part of the environment is fixed and part is time-varying.

3.2.3 Correlated Rayleigh Fading

When the transmit antennas are at close proximity or there is not enough scattering in the environment, the fading gains of the antennas are correlated. From a traditional diversity point of view, antennas with correlated fading is less useful than antennas with independent

fading. From the point of view of opportunistic beamforming in a fast fading environment, the opposite conclusion is true. We illustrate this phenomenon using the example of completely correlated Rayleigh fading:

$$h_{nk}(t) = l_k(t) \exp(j\phi_{nk}).$$

Here, the channel gains from all the transmit antennas to a user is the same except for a phase shift; $\{l_k(t)\}$ is a Rayleigh fading process. The phases ϕ_{nk} depend on the angle of the direct path to user k with respect to the antenna array, the actual placement (linear versus planar arrangement) of the antenna array, but fixed over time. We can write $\phi_{nk} = r(n, \psi_k)$ where θ_k is the angle of departure of the direct path to user k and r represents the function that decides the phases at the antennas abstracting the placement of the antenna array. For example, with linear arrays and uniform spacing of length d between the antennas we have

$$\phi_{nk} = \psi_k + \frac{2(n-1)d\pi \cos\psi_k}{\lambda} \quad , n = 1 \ldots N \ ,$$

where λ is the wavelength of the transmitted signal. Unlike the independent Rayleigh fading case, the overall channel gain $h_k(t)$ is no longer Rayleigh; instead, it is a mixture of Gaussian distributions with different variances. When the received signals from the transmit antennas add in phase, the overall received SNR is large; when the received signals add out of phase, the overall received SNR is small.

In the case of completely correlated fading, power randomization is not necessary, since the transmit antennas always have the same magnitude gain to each of the users. It suffices to allocate equal amount of power to each of the antennas ($\alpha_n(t) = 1/N$) and change the phases by rotating the single parameter: angle of departure. Denoting this single parameter by $\theta(t)$,we let $\theta_n(t) = -\theta(t) - r(n, \theta(t))$ where $\theta(t)$ is uniformly rotated. In Figure 11, we plot the distribution of the overall channel gains with opportunistic beamforming of four transmit antennas, and compare it to the case of one transmit antenna (Rayleigh fading). We assumed $d = \frac{\lambda}{8}$ for this simulation example. One can observe the increase in dynamic range due to opportunistic beamforming. Figure 12 shows the total throughput with and without opportunistic beamforming in the completely correlated fading case. There is a significant improvement in throughput, in contrast to the independent fading case.

An asymptotic analysis in the limit of large number of users provides some insight. The overall channel gain of the kth user under opportunistic beamforming is given by:

$$|h_k|^2 = \left| \sum_{n=1}^{N} \frac{1}{\sqrt{N}} \exp[j(\theta_n + \phi_{nk})] \right|^2 |l_k|^2.$$

Thus $|h_k|^2$ is a product of two independent random variables. The maximum value that the first random variable can take on is N, when user k is in the beamforming configuration. Consider a large system with many users. For a fixed $\delta > 0$, there will almost surely be a fraction $\epsilon \in (0, 1)$ of users for which

$$\left| \sum_{n=1}^{N} \frac{1}{\sqrt{N}} \exp[j(\theta_n + \phi_{nk})] \right|^2 > N - \delta.$$

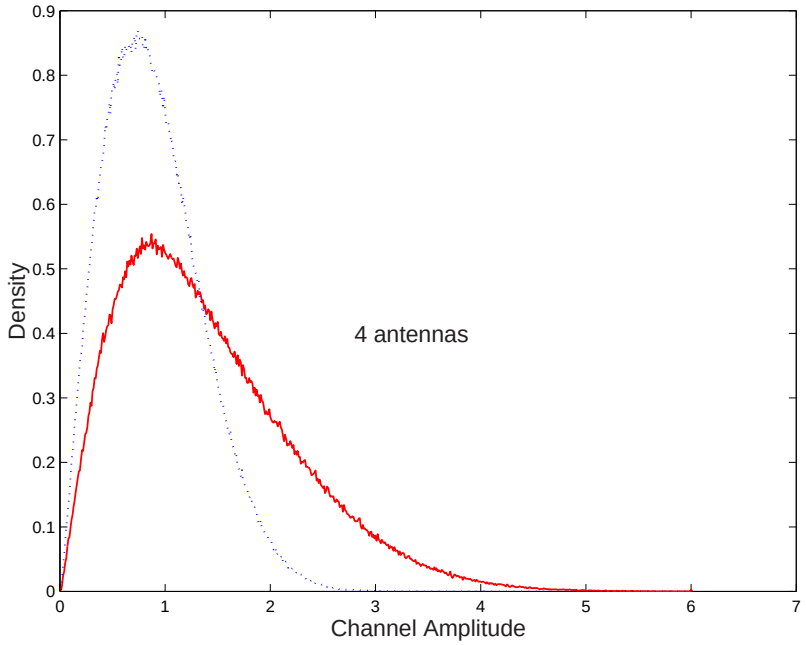

Figure 11: Comparison of the distribution of the overall channel gain with and without opportunistic beamforming using 4 transmit antennas, completely correlated Rayleigh fading.

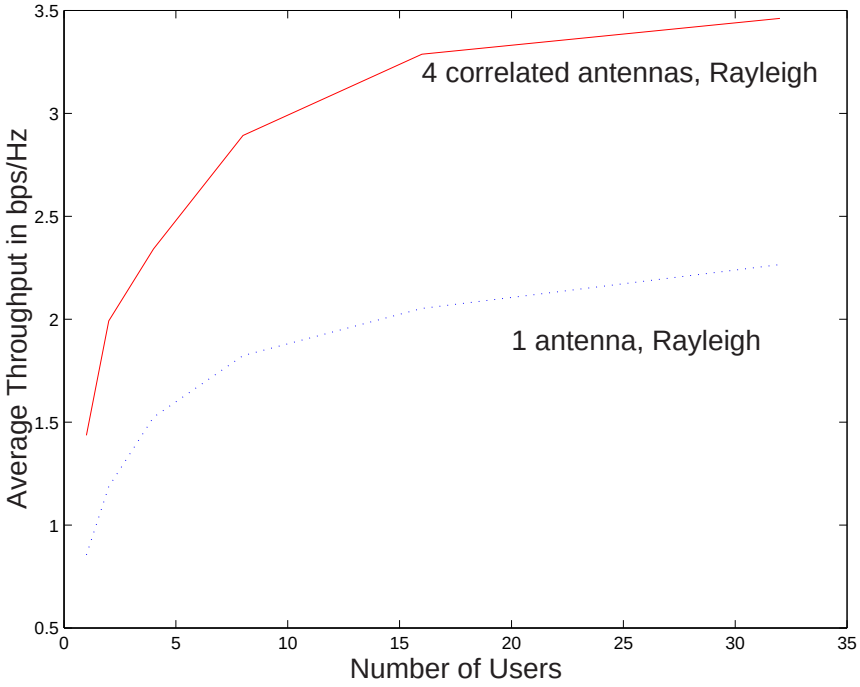

Figure 12: Total throughput as a function of the number of users under completely correlated Rayleigh fading, with and without opportunistic beamforming.

Now l_k are i.i.d. Rayleigh distributed random variables. Hence, among these ϵK users, the maximum of their $|h_k|^2$ grows at least as fast as $(N-\delta)\log(\epsilon K) = (N-\delta)\log K + O(1)$ as $K \to \infty$. This is true for every $\delta > 0$, and it gives a lower bound to the growth rate of $\max_{1\leq k\leq K} |h_k|^2$. Moreover, it is also clear that $N \log K$ is an upper bound to that growth rate. Thus, we see that opportunistic beamforming in a correlated fading environment yields approximately a factor of N improvement in SNR in a system with large number of users. The improvement is more dramatic than in the case of independent Rician fading considered earlier. Another important improvement is in the rate of convergence to the asymptotic performance, the asymptote being in number of users. Since the powers are not being varied, and the phase is varied in only one dimension, the number of users required to achieve close to asymptotic performance grows only *linearly* with the number of antennas.

4 Opportunistic Beamforming vs Space Time Codes: A Comparison

We have motivated the use of multiple transmit antennas to artificially induce an environment with larger and faster channel fluctuations. This channel fluctuation increases the multiuser diversity available in the system and is harnessed by an appropriate scheduler. This use of multiple transmit antennas to perform opportunistic beamforming was motivated by taking a multiuser communication system point of view. On the other hand, there exist schemes, referred to as *space time codes*, which use multiple transmit antennas in a point-to-point communication scenario. In this section, we will compare and contrast the opportunistic beamforming technique with a multiuser system using space time codes (designed for a point-to-point communication system) in terms of both system requirements and performance.

For concreteness, we will begin with a pair of transmit antennas at the base station. The best known space time code for this scenario is given by Alamouti [1], and has been accepted as an option in the 3G standards [19]. This scheme requires *separate* pilots (or training symbols) for each of the transmit antennas and the receivers track the channels (amplitude and phase) from both the transmit antennas. Consider the slow fading scenario where the channel does not vary over the time scale of communication. The Alamouti scheme creates essentially a single transmit antenna channel with effective SNR of user k given by:

$$\frac{P\left(\mid h_{1k} \mid^2 + \mid h_{2k} \mid^2\right)}{2\sigma^2} \tag{7}$$

where P is the total transmit power. Observe that unlike the opportunistic beamforming scheme, the effective channel of each user does not change with time in a slow fading environment. In this static environment, the proportional fair scheduling algorithm reduces to equal-time scheduling [21]. Comparing this performance with that under opportunistic beamforming, we see for large number of users, from Theorem 1, that users are also allocated equal time *but* the effective SNR when a user is transmitted to is

$$P\frac{\mid h_{1k} \mid^2 + \mid h_{2k} \mid^2}{\sigma^2}, \tag{8}$$

twice that in the Alamouti scheme. This is the so called "3 dB gain" achieved from transmit beamforming. *Actual* transmit beamforming requires feedback of the phases and amplitudes of both the channels to the transmitter. The opportunistic beamforming scheme achieves this performance using minimal feedback from each receiver: SNR of the overall channel.

We can also compare the outage performance of both schemes. This metric is relevant when the bit rate is to be maintained constant and we are interested in minimizing the probability of outage; outage is the event that the constant rate is not supportable by the random slow-fading channel condition. One way to characterize this performance is how fast the outage probability decays as a function of the average SNR (SNR $= P/\sigma^2$) for a given target rate. The outage probability for a scheme can be computed as the probability that the effective SNR falls below the target level. For independent Rayleigh fading gains h_{1k}, h_{2k}, the Alamouti scheme (c.f. eqn (7)) achieves an outage performance decay of $\frac{1}{\text{SNR}^2}$ (in contrast to the decay of $\frac{1}{\text{SNR}}$ when there is a single transmit antenna). Thus, Alamouti's scheme yields a *diversity gain* of 2. The opportunistic beamforming scheme with many users also has the same decay of outage probability with SNR with a further 3dB gain on the value of SNR (c.f. eqn. (8). Thus, in a multiuser system with enough users under proportional fair scheduling, the opportunistic beamforming scheme strictly outperforms the Alamouti scheme in terms of both throughput and outage performances at all SNR levels.

An important point to observe is that implicit in the comparison is the assumption that we are spending equal amount of time serving each user in the system. This is true if we use a proportional fair scheduling algorithm. If on the other hand another scheduling algorithm is used which spends a large fraction of time serving one user, then the Alamouti scheme could yield a better performance than opportunistic beamforming, for *that* user. This is because with so many time slots allocated to the user, it will not be able to always serve him near the peak under opportunistic beamforming. This may happen if there is one user with a very poor channel and the system has to allocate a disproportionate amount of resources just to meet a minimum rate requirement.

One should also compare the two schemes in terms of system requirements. The Alamouti scheme requires separate pilot symbols on both of the transmit antennas. It also requires all the receivers to track both the channels (amplitude and phase). To achieve the throughput in (9), a slow scale feedback of the current channel SNR is also required from the receivers to the transmitter. On the other hand, the opportunistic beamforming scheme does not require separate pilot symbols on the transmit antennas. The same signal (including pilot and data) goes over both the transmit antennas. The receivers track the channel and a tight feedback of the instantaneous SNR of the receivers to the transmitter is required. We point out that such feedback is a part of the system design of all 3G systems and appears to be a mild system requirement in view of the advantages it allows, particularly for data systems where the latency time-scale is not as tight as voice. Further, to implement space time codes in a system all the receivers have to implement a specific demodulating technique (that has complexity twice that of the single transmit antenna case). In contrast, the opportunistic beamforming scheme has no such requirement. In fact, the receivers are *completely ignorant* of the fact that there are multiple transmit antennas and the receiver is identical to that in the single transmit antenna case. It is in this context that we have termed our technique as

using "dumb antennas".

With more than two transmit antennas, the Alamouti scheme does not generalize; no full-rate designs exist [16]. The orthogonal designs in [16] achieve data rates which are fractions of that which can be supported by a channel with effective SNR:

$$\frac{P}{N}\frac{\sum_{n=1} \mid h_{nk} \mid^2}{\sigma^2}. \tag{9}$$

for user k. Comparing this quantity with the performance of opportunistic beamforming, we see from Theorem 1, for a large system that user k has SNR N times larger than that under full-rate orthogonal designs (even if they existed). From the point of view of outage, the diversity gain is N in both cases.

Let us now consider the fast Rayleigh fading scenario. In this case, we have observed that the opportunistic beamforming technique has no effect on the overall channel and the full multiuser diversity gain is realized. It is interesting to observe that with space time codes on the array of transmit antennas makes the time varying channel almost constant: by the law of large numbers, for any user k,

$$\frac{\sum_{n=1}^{N} \mid h_{nk} \mid^2}{N} \simeq 1 \ ,$$

as the number of antennas N grows. Thus, the space time codes turn the time varying channel into a less varying one and the inherently available multiuser diversity gain is reduced (cf. Figure 1). We conclude that the use of space time codes are actually harmful in the sense that even the naturally present multiuser diversity has been removed. Of course, to capture the inherent multiuser diversity gain, the transmitter has to be able to track the channels of the users. In scenarios when the fading is very fast or the delay requirement is very short, such tight feedback may not be possible. We will revisit this point in Section 7.

We have based our comparison in the case of coherent communication: when pilot symbols are inserted at the transmitter and the receiver tracks both the amplitude and phase of the channel. A noncoherent space time coding scheme has been proposed in [17] which has about a 3dB loss in SNR with respect to the performance of (9). The opportunistic beamforming also can be used in conjunction with a noncoherent communication scheme and the resulting performance will again be 3dB better in SNR when compared to the space time coding approach.

In this comparison, we have retained the assumption that only one user is scheduled at any time. We visit the issue of scheduling to multiple users simultaneously in Appendix B by taking an information theoretic view of the downlink broadcast channel.

5 WideBand Channel

The performance gain of opportunistic beamforming becomes more apparent when there are many users in the system. This suggests that the technique is particularly suited in

wide band channels shared by many users. In such a wide band channel, it is natural to consider *frequency-selective* fading. While multiuser diversity gain in flat fading channels is obtained by scheduling users when their overall channel SNR is good, multiuser diversity gain in frequency-selective fading channel is exploited by transmitting to the users on the frequency band where their channel SNR is good.

A simple model of a frequency selective wide band channel is a set of L parallel narrow band sub-channels with channel fluctuations in each of the narrow band channels being frequency flat. The transmit power is fixed to be P for each of the narrow band sub-channel.[1] The users measure the SNR on each of the narrow band sub-channels and feed back the SNRs (or equivalently, requested rates) to the base-station. Observe that this scheme requires L times more feedback than in the flat fading case where a single requested rate is fed back.[2] The scheduler allocates at each time a single user to transmit to for each of the narrow band sub-channels.

The proportional fair scheduling algorithm generalizes naturally from the flat fading to the frequency-selective fading scenario. For each user k, it keeps track of the average throughput $T_k(t)$ the user has been getting *across all narrow band sub-channels* in a past window of length t_c. For each narrow band sub-channel l, it transmits to the user $k^*(l)$, where

$$k^*(l) = \text{argmax}_k \frac{R_{kl}(t)}{T_k(t)},$$

where $R_{kl}(t)$ is the requested rate of user k in channel l at time slot t. Observe that the throughput $T_k(t)$ is averaged over all the narrow band sub-channels, not just the sub-channel l. This is because the fairness criterion pertains to the *total* throughput of the users across the entire wide band channel and not to each of the narrow band sub-channel. As in the flat fading scenario, when $t_c = \infty$ and the fading statistics are stationary and ergodic, this algorithm can be shown to maximize $\sum_k \log T_k$, where T_k is the average throughput of each user (across all bands) [21].

A natural generalization of the opportunistic beamforming technique is to generate independent powers and phase randomization processes in the different sub-channels. A performance analysis can be done in a similar way as in the flat fading scenario. In the fast fading case with symmetric stationary fading statistics among the users (and $t_c = \infty$), the steady-state throughput T_k is the same for every user. Hence, the proportional fair algorithm reduces to scheduling the user with the highest request rate in each of the narrow band sub-channel. Thus the throughput per user per channel scales exactly as in the flat fading case, already analyzed in Section 3.2, and the total throughput per user is just the sum of the throughputs over all the narrow band channels. The advantage of having a wider band channel in the fast fading scenario comes from the fact that all users share all bands, translating into more users per band for the opportunistic beamforming technique to capitalize on. (Recall that the throughput per band always grows with the number of users K.)

[1] In theory, performance can be further improved by allocating different amount of power for each of the narrow band sub-channels. In a system with large number of users this improvement is marginal because of a statistical effect.

[2] On a more practical note, users could only feedback the SNR value on the best of the sub-channels and the identity of that sub-channel. In this case the extra feedback increases only logarithmically in L.

Let us now consider the time-invariant slow fading scenario, where the gain of user k to antenna n in sub-channel l is given by $h_{nk}^{(L)}$ and does not change over time. We showed in Theorem 1 that for the flat fading case, opportunistic beamforming allows each user to be scheduled at its peak rate (i.e. when it is at its beamforming configuration) as long as there are sufficiently many users in the system and the stationary distribution of the power and phase rotation process matches that of the slow fading distribution of the users. A generalization to the wide band case can be obtained.

Theorem 3 *Suppose the slow fading states of the users are independent and identically distributed and are discrete, and that the slow fading state distribution for each user is symmetric across sub-channels. Assume also that for every l, the joint stationary distribution of the power and phase randomization process for sub-channel l:*

$$(\alpha_{1l}(t), \ldots, \alpha_{Nl}(t), \theta_{1N}(t), \ldots, \theta_{Nl}(t))$$

is the same as the distribution of user k's slow fading state

$$\left(\frac{\mid h_{1k}^{(l)} \mid^2}{\sum_{n=1}^{N} \mid h_{nk}^{(l)} \mid^2}, \ldots, \frac{\mid h_{Nk}^{(l)} \mid^2}{\sum_{n=1}^{N} \mid h_{nk}^{(l)} \mid^2}, -arg(h_{1k}^{(l)}), \ldots, -arg(h_{Nk}^{(l)}) \right),$$

conditional on the fact that

$$\sum_{n=1}^{N} \mid h_{nk}^{(l)} \mid^2 = \max_{i=1\ldots L} \sum_{n=1}^{N} \mid h_{nk}^{(i)} \mid^2 . \tag{10}$$

Then, almost surely, we have

$$\lim_{K\to\infty} KT_k^{(K)} = L \max_{l=1\ldots L} R_{kl}^{\text{bf}}$$

for all k. Here, R_{kl}^{bf} is the data rate that user k achieves in sub-channel l when it is in the beamforming configuration, i.e., when its instantaneous SNR is

$$P \sum_{n=1}^{N} |h_{nk}^{(l)}|^2 / \sigma^2.$$

Proof

The proof is along the lines of that of Theorem 1. See Appendix A.

∘

Thus in a system with large number of users, the proportional fair algorithm serves each user when it is at its peak over all degrees of freedom, i.e. as though each user is transmitted to only when it is perfectly beamformed and only in the sub-channel for which the beamforming gain is the highest. Moreover, the algorithm spends equal amount of time serving each user, but each user is served only in the sub-channel in which its channel is the strongest.

For the theorem to hold, there should be a match between the power and phase randomization processes and the slow fading state distribution of the users. In the case when the slow fading state distribution is Rayleigh and independent across all sub-channels, however, the conditioning (10) is unnecessary and hence the matching requirement is identical to that in the narrow band case. This is because for Rayleigh fading,

$$\left(\frac{\mid h_{1k}^{(l)} \mid^2}{\sum_{n=1}^{N} \mid h_{nk}^{(l)} \mid^2}, \ldots, \frac{\mid h_{Nk}^{(l)} \mid^2}{\sum_{n=1}^{N} \mid h_{nk}^{(l)} \mid^2}, -\arg(h_{1k}^{(l)}), \ldots, -\arg(h_{Nk}^{(l)}) \right)$$

and

$$\sum_{n=1}^{N} \mid h_{nk}^{(l)} \mid^2$$

are independent. Applying Theorem 3 to the Rayleigh case, we see that opportunistic beamforming asymptotically yields a NL-fold diversity gain for each user in the slow fading environment. This diversity gain is the product between the transmit antenna diversity gain and the frequency diversity gain.

6 Cellular Systems: Opportunistic Nulling

So far we have considered a single cell scenario, where the noise is assumed to be white Gaussian. For wide band cellular systems with full frequency reuse, it is important to consider the effect of inter-cell interference on the performance of the system, particularly in interference-limited scenarios. In a cellular system, the channel quality of a user is measured by the SINR, signal-to-interference-plus-noise ratio. In a fading environment, the energies in both the received signal and the received interference fluctuate over time. Since the multiuser diversity scheduling algorithm allocates resources based on the channel SINR (which depends on both the channel amplitude and the amplitude of the interference), it automatically exploits both the fluctuations in the energy of the received signal as well as that of the interference: the algorithm tries to schedule resource to a user whose instantaneous channel is good and the interference is weak. Thus, multiuser diversity naturally takes advantage of the time-varying interference to increase the spatial reuse of the network [21].

From this point of view, power and phase randomization at the base-station transmit antennas plays an additional role: it increases not only the amount of fluctuations of the received signal to the intended users *within the cells*, it also increases the fluctuations of the interference the base-station causes *in adjacent cells.* Hence, opportunistic beamforming has a dual benefit in an interference-limited cellular system. In fact, opportunistic beamforming performs *opportunistic nulling* simultaneously: while randomization of power and phase in the transmitted signals from the antennas allows near coherent beamforming to some user within the cell, it will create near nulls at some user in adjacent cells. This in effect allows *interference avoidance* for that user if it is currently being scheduled.

Let us focus on the slow, flat fading scenario to get some insight on the performance gain from opportunistic beamforming and nulling. Under power and phase randomization at all

base-stations, the received signal of a typical user being interfered by J adjacent base-station is given by

$$y(t) = h(t)x(t) + \sum_{j=1}^{J} g_j(t)u_j(t) + z(t).$$

Here, $x(t)$ is the signal of interest, $u_j(t)$ is the interference from the jth base-station and $w(t)$ is additive Gaussian noise. All base-stations have the same transmit power P and N transmit antennas and are performing power and phase randomization independently; $h(t)$ and $g_j(t)$'s are the *overall* channel gains from the base-stations:

$$h(t) := \sum_{n=1}^{N} \sqrt{\alpha_n(t)} e^{j\theta_n(t)} h_n, \qquad g_j(t) := \sum_{n=1}^{N} \sqrt{\beta_{nj}(t)} e^{j\phi_{nj}(t)} g_{nj}$$

where h_n and g_{nj} are the slow fading channel gains to the user from the nth transmit antenna of the base-station of interest and the interfering base-station j respectively. Averaging over the signal $x(t)$ and the interference $u_j(t)$'s, the (time-varying) SINR of the user can be computed to be:

$$\mathrm{SINR}(t) = \frac{P|h(t)|^2}{P\sum_{j=1}^{J}|g_j(t)|^2 + \sigma^2}.$$

The SINR varies because of both the variations of the overall gain from the base-station of interest as well as those from the interfering base-station. In a system with many other users, the proportional fair scheduler will serve this user while its SINR is at its peak $P\sum_{n=1}^{N}|h_n(t)|^2/\sigma^2$, i.e. when the received signal is the strongest and the interference is completely nulled out. Thus, the opportunistic nulling and beamforming technique has the potential to shift a user from a low SNR, interference-limited regime to a high SNR, noise-limited regime.

How close the performance of opportunistic beamforming and nulling in a finite-size system to this asymptotic limit depends on the probability that the received signal is near beamformed *and* all the interference is near null. In the interference-limited regime when $P/\sigma^2 \gg 1$, the performance depends mainly on the probability of the latter event. The probability of this latter event is larger when there are only one or two base-stations contributing most of the interference, as is typically the case. In contrast, when there is interference from many base-stations, interference averaging occurs and the probability that the *total* interference is near null is much smaller. Interference averaging, which is good for CDMA networks, is actually unfavorable for the opportunistic scheme described here, since it reduces the likelihood of the peaks of the SINR.

In a typical cell, there will be a distribution of users, some closer to the base-station and some closer to the cell boundaries. Users close to the base-station are at high SNR and are noise-limited; the contribution of the inter-cell interference is relatively small. These users benefit mainly from opportunistic beamforming (diversity gain plus 3dB power gain if there are 2 transmit antennas.) Users close to the cell boundaries, on the other hand, are at low SNR and are interference-limited; the average interference power can be much larger than the background noise. These users benefit both from opportunistic beamforming and from opportunity nulling of inter-cell interference. Thus, the cell-edge users benefit more

in this system than users in the interior. This is rather desirable from a system fairness point-of-view, as the cell-edge users tend to have poorer service. This feature is particularly important for a system without soft handoff (which is difficult to implement in a packet data scheduling system). To maximize the opportunistic nulling benefits, the transmit power at the base-station should be set as large as possible, subject to regulatory and hardware constraints.

7 System and Implementation Issues

We have introduced the opportunistic beamforming scheme to induce an artificial fading environment as motivated by taking a system design view. In this section, we continue this view further and delineate the impact of introducing this scheme in a complete wireless data system.

We begin with the variation of powers and phases. Observe that the data phase and power variations can all be achieved in baseband and the only extra hardware requirement is that each antenna have its own RF (radio frequency) card with individual power amplifiers. The important constraint on this variation is that it be slow enough so that the loops track the channel (in case pilot symbols are introduced and the mobiles have coherent demodulation in the downlink) and the feedback SNR measurements are stable. A 2π phase change in about 20 to 50 ms is realistic currently. In Theorem 1 we have seen analytically the best stationary distribution of power and phase variation. The phase distribution is uniform over $[0, 2\pi]$ and the power fraction α_1 has marginal distribution the same as $\frac{|h_1|^2}{\sum_{n=1}^{N}|h_n|^2}$ where $h_1, \ldots, h_N$ are i.i.d. Rayleigh distributed. Figure 13 plots the density of α_1 for the case of two and three antennas.

The range of phase and power variation could be quantized and each quantized "state" be visited in some deterministic (and continuous, i.e., not visiting very different states within a small interval of time) fashion. If the power variation is over the complete range (i.e., α_i's vary from 0 to 1), then the total power rating of each of the power amplifiers should be equal to that dictated by the total link budget. In this case, the system requirement is that the number of Class AB power amplifiers with the same linear region of operation has been multiplied by N. Even though base stations are usually powered by an AC supply, the very poor power efficiency of the amplifiers is a serious issue as are the cost and size (including the heat sink). One way to ameliorate this issue is to ensure that the power variation is never entirely over the range $[0, 1]$, but instead is over a smaller range, say $[0.3, 0.7]$. This way the power rating of the amplifiers is reduced.

There is another important reason to restrict the power variations. Varying the powers and phases over the entire range means that any user (in a slow fading state, suppose) will be beamformed at some point in time and completely nulled out at another point in time and the channel varies smoothly in between. Since the user is being scheduled only when the channel is at its peak, the performance of the system is described by the peak channel. However, observe that control channels (standard overheads to maintain proper

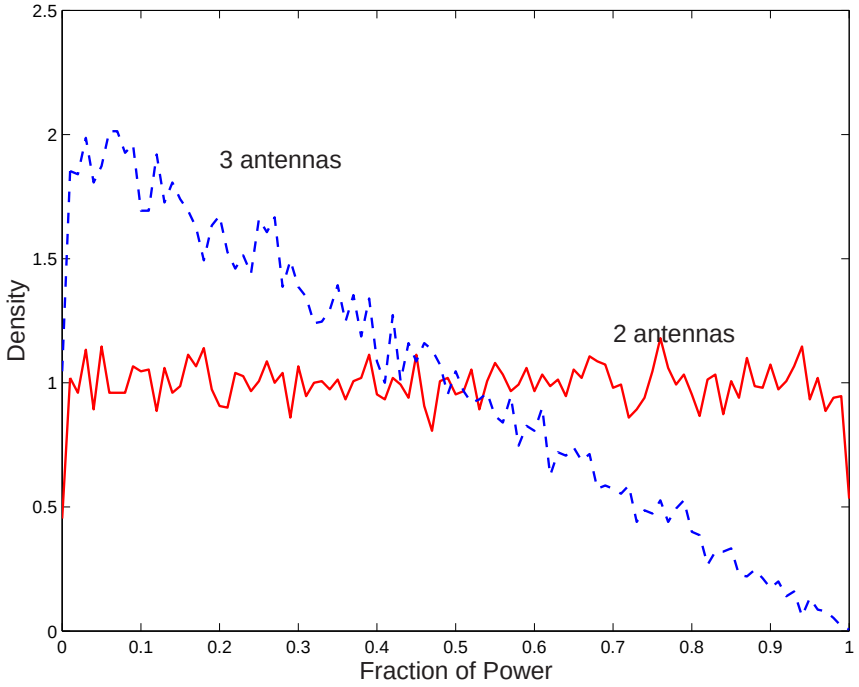

Figure 13: Plot of the marginal density of α_1 for two and three antennas. Observe that for two antennas, the density is almost uniform.

functioning of the wireless system) also follow this channel variation. Unlike the data which have the delay flexibility to allow dynamic scheduling, control channels are "circuit-switched" and have very tight latency requirements, i.e., they are being transmitted continuously and the users are demodulating them all the time. From the perspective of these signals, it is preferable that the channel remained nonfading; a requirement that is contradictory to our scheduler-oriented observation that we would prefer the channel to have fast and large variations. For example, when the channel is completely nulled out for a user, the loop that tracks the channel and feeds back SNR might go off track. Allowing the powers to vary only over a limited range is one way to try to ensure that the channel does not get completely nulled out (and the channel is also not completely beamformed in this case).

These issues suggest the following design perspective: separate very-low latency signals (such as control signals) from flexible latency data. One way to achieve this separation is to split the bandwidth into two parts. One part is made as flat as possible (by spreading over this part of the bandwidth and possibly space time codes on multiple transmit antennas) and used to transmit flows with very low latency requirements. The performance metric here is to make the channel as reliable as possible (equivalently keeping the probability of outage low) for some fixed data rate. The second part uses opportunistic beamforming to induce large and fast channel fluctuations and a scheduler to harness the multiuser diversity gains. The performance metric on this part is to maximize the multiuser diversity gain.

In traditional cellular wireless systems, the cell is sectorized to allow better focusing of the power transmitted from the antennas and also to reduce the interference seen by mobile

users from transmissions of the same base station but intended for users in different sectors. This technique is particularly gainful in scenarios when the base station is located at a fairly large height and thus there is limited scattering around the base station. In contrast, in systems with far denser deployment of base stations (a strategy that can be expected to be a good one for wireless systems aiming to provide mobile, broadband data services), it is unreasonable to stipulate that the base stations be located high above the ground so that the local scattering (around the base station) is minimal. In an urban environment, there is substantial local scattering around a base station and the gains of sectorization are minimal; users in a sector also see interference from the same base station (due to the local scattering) intended for another sector. The opportunistic beamforming scheme can be thought of as sweeping a random beam (the beam is logical in the diversity antenna case and physical in the correlated antenna case) and scheduling transmissions to users when they are beamformed. Thus the gains of sectorization are automatically realized. We conclude that the opportunistic beamforming technique is particularly suited to harness sectorization gains even in low height base stations with plenty of local scattering. In a cellular system, the opportunistic beamforming scheme also obtains the gains of nulling, a gain traditionally obtained by coordinated transmissions from neighboring base stations in a full frequency reuse system or by appropriately designing the frequency reuse pattern.

We saw in Section 4 that the opportunistic beamforming technique of using an array of multiple transmit antennas has approximately N fold improvement in received SNR at a user in a slow fading environment as compared to the best space time code. With an array of N *receive* antennas at each mobile (and say a single transmit antenna at the base station), the received SNR of any user gets an N fold improvement as compared to a single receive antenna (this gain is realized by receiver beamforming; an operation easy to accomplish since the mobile has full channel information at each of the antenna elements). Thus, the gains of opportunistic beamforming are about the same order as that of replacing the transmit antenna array of the base station with a set of receive antenna array at *each* of the mobiles. Thus, for a system designer, the opportunistic beamforming technique provides a compelling case for implementation, particularly in view of the constraints of space and cost of installing multiple antennas on *each* mobile. Further, this technique neither needs any extra processing on part of any mobile receiver, nor any updates to the air-link interface standards. We emphasize that the receiver can be totally ignorant of the use or non use of this technique. Thus it does not have to be "designed in" (by appropriate inclusions in the air interface standard and the receiver design) and can be added any time. This is one of the important and biggest benefits of this technique from an overall system design point of view.

8 Conclusion

Multiuser diversity has the potential to provide very significant performance boosts to wireless systems. The only system requirement is tight feedback of the channel quality by the users to the base station. The gain comes from scheduling a user's transmissions at times when the user's channel SNR is near its peak. In practice, the gain is often limited by the slow rate of channel variations and/or stringent latency requirements, resulting in smaller

time scale over which the users have to be scheduled. In this paper, we have introduced a technique, *opportunistic beamforming and nulling*, to *induce* fast and large fluctuations in the SNR of the users. This technique amplifies the possible multiuser diversity gain while satisfying reasonable latency requirements.

The opportunistic beamforming technique utilizes the transmit antenna array to change the overall channel seen by the users. We have made the case in this paper that the gains from the opportunistic beamforming and nulling technique are significant, especially in the context of the fact that this "dumb technique" has no overhead in terms of inserting pilot or training symbols on *each* of the transmit antennas or complicating any existing receiver (the state of the art space time codes require that the receivers be appropriately configured). In summary, this technique can be introduced in existing systems with minimal (and only base station specific) changes: an array of transmit antennas and radio frequency hardware circuits to drive each of them, combined with an appropriate scheduler.

A new design principle for wireless networks is emerging through the lens of multiuser diversity. Traditionally, much of the design techniques in wireless systems centered on making the individual point-to-point links as close to AWGN channels as possible, with a reliable channel quality that is constant over time. These techniques include diversity techniques such as multipath combining and time interleaving that attempt to keep the channel fading constant in time, as well as interference management techniques such as interference averaging by means of spreading. Indeed, all these techniques are used in spread-spectrum systems like IS-95 and CDMA-2000. However, if one shifts the point of view of the wireless system as a set of point-to-point links to the view as a system with multiple users sharing the same resources (spectrum and time), then quite a different design objective suggests itself. Indeed, the results in this paper suggest that one should instead try to make the channel fluctuations as large as possible so that the scheduler can "ride the peaks", i.e., each user is scheduled when it has a very strong channel. This is accomplished by varying the strengths of *both* the signal and the interference that a user receives.

Several open problems suggest themselves through this work. One relates to the variation of power and phases in a manner that causally depends on the feedback of overall channel quality from the users (as opposed to a random variation analyzed in this paper). Another topic is that of a general scheduler subject to rate and delay requirements of the users. This paper has focused mainly on the proportional fair scheduler and [12] is a recent work on schedulers that give more diverse service guarantees.

9 Acknowledgements

The authors would like to record their gratitude to the inquisitive questions raised during talks based on the material in this paper. The first author acknowledges useful comments from, and thanks, Robert Calderbank, Bruce Hajek, Raymond Knopp, Vahid Tarokh, Jack Salz, Shlomo Shamai and Andrew Viterbi.

A Proof of Theorem 1

First we review a basic theoretical property of the proportional fair scheduler.

Let $R_k(t)$ be the data rate that user k's channel can support at time t. Suppose $\{R_k(t)\}, k = 1, \ldots, K$, are jointly stationary and ergodic processes. A scheduler selects at each time which user to transmit data to and the decision at time t depends (causally) on $R_k(s)$, $s \leq t$, $k = 1, \ldots K$. Under any scheduler, define the throughput T_k achieved by user k as the limit of the long-term average data rate transmitted to user k, if exists, and otherwise the corresponding limit infimum. The proportional fair scheduler has the following optimality property.

Lemma 4 *[21] Under the proportional fair algorithm with averaging time-scale $t_c = \infty$, the long term average throughput of each user exists almost surely, and the algorithm maximizes*

$$\sum_{k=1}^{K} \log T_k$$

almost surely among the class of all schedulers.

Let the (discrete) slow fading states be $\mathbf{h}_j = (h_{j1}, \ldots, h_{jN}), j = 1, \ldots, M$, and p_j be the probability that a user is in slow fading state h_j. Let us denoted the rate of transmission for the beamformed SNR value of state $\mathbf{h}_j$ (equal to $\left(\sum_{n=1}^{N} \mid h_{jN} \mid^2\right)/\sigma^2$) by R_j^{bf}. Denote the beamforming state of power and phase variation corresponding to the slow fading state $\mathbf{h}_j$ by $(\alpha^{(j)}, \theta^{(j)})$. By hypothesis of the theorem, we have assumed that the joint stationary distribution of the power and phase variation process in time has probability p_j on the state $(\alpha^{(j)}, \theta^{(j)})$, and by ergodicity this is also the long term fraction of time the process spends in that state.

Fix the number of users K and denote the fraction of users in fading state $\mathbf{h}_j$ by $\psi_j^{(K)}$ for $j = 1 \ldots M$. A user is said to be in class j if it is in the slow fading state $\mathbf{h}_j$. We have for all K that $\sum_{j=1}^{M} \psi_j^{(K)} = 1$ and almost surely that

$$\lim_{K \to \infty} \psi_j^{(K)} = p_j, \quad j = 1 \ldots M \ . \tag{11}$$

We first consider the following simple algorithm. At the time the phase and power variation process is in state j, schedule a user in class j. Furthermore, within a class, schedule users equal number of times. Since a user in class j is scheduled only when the power and phase variation process is in the beamforming configuration with respect to it's fading state, the user is scheduled at it's peak rate. We can compute the average throughput seen by a user k_j in class j to be

$$\frac{p_j R_j^{\text{bf}}}{\psi_j^{(K)} K} \ . \tag{12}$$

Now consider the proportional fair algorithm. By symmetry, the algorithm schedules users in the same fading state equally in time and let us denote the fraction of time it schedules users in fading state j by $\beta_j^{(K)}$. We have for all K that

$$\sum_{j=1}^{M} \beta_j^{(K)} = 1 \ . \tag{13}$$

Denoting the throughput of user k under the proportional fair algorithm by $T_k^{(K)}$ we have the following simple upper bound by observing that a user cannot be scheduled at rate more than his peak rate:

$$\sum_{k=1}^{K} \log T_k^{(K)} \leq \sum_{j=1}^{M} \psi_j^{(K)} K \log \left(\frac{\beta_j^{(K)} R_j^{\text{bf}}}{\psi_j^{(K)} K} \right) \ . \tag{14}$$

Appealing to Lemma 4, we arrive at the following lower bound using (12):

$$\sum_{k=1}^{K} \log T_k^{(K)} \geq \sum_{j=1}^{M} \psi_j^{(K)} K \log \left(\frac{p_j R_j^{\text{bf}}}{\psi_j^{(K)} K} \right) \ . \tag{15}$$

Combining (14) and (15), we get

$$\sum_{j=1}^{M} \psi_j^{(K)} \log \left(\frac{\beta_j^{(K)}}{p_j} \right) \geq 0, \ \ \forall K \ .$$

Using (11), we arrive at

$$\limsup_{K \to \infty} \sum_{j=1}^{M} p_j \log \frac{p_j}{\beta_j^{(K)}} \leq 0.$$

But

$$\sum_{j=1}^{M} p_j \log \frac{p_j}{\beta_j^{(K)}}$$

is the divergence between two probability vectors (nonnegative and equals to zero if and only if the two probability vectors are the same. Hence, $\forall j = 1 \ldots M$,

$$\lim_{K \to \infty} \beta_j^{(K)} = p_j \ .$$

The fraction of time any user in class j is scheduled scaled by the number of users is

$$K \frac{\beta_j^{(K)}}{\psi_j^{(K)} K} \to 1 \ \text{ as } K \to \infty \ , \tag{16}$$

i.e. asymptotically the proportional fair algorithm gives equal time to all users. Thus the throughput of any user k in class j under the proportional fair algorithm has the property

$$\liminf_{K \to \infty} K T_k^{(K)} \leq R_j^{\text{bf}} \ .$$

Combining this with the lower bound on the performance of the proportional fair algorithm in (15) and using (11) we arrive at the claim of the theorem. □

Sketch of Proof of Theorem 3:

We can now use the notation and the technique of the proof of Theorem 1 to prove the result in Theorem 3. The first extension is that the set of slow fading states has cardinality M^L where, as before, M denotes the number of slow fading states (the same, by hypothesis of the theorem) in any of the narrow bands. There are up to M distinct power and phase variation values, each corresponding to at least one of the beamforming coefficients of the M slow fading states.

Consider the following extension of the simple scheduling algorithm that was used above to lower bound the performance of the proportional fair algorithm. For each power and phase variation value, denoted with some abuse of notation, by $\mathcal{A}_j$ corresponding to the beamforming coefficients of, say, fading state j, schedule a user with the L fading states such that the maximum of the L beamforming SNR values is exactly that corresponding to the fading state j. Furthermore, schedule equal time among users with the above property. The key step is the identification of the fraction of users that will be scheduled at any power and phase variation value $\mathcal{A}_j$ to be exactly the conditional probability that the maximum beamforming SNR of the L narrow band channels is the beamforming SNR corresponding to the narrow band fading state $\mathbf{h}_j$.

An argument similar to the one above can now be made: under the proportional fair algorithm, each user is scheduled for approximately equal amount of time and since there are L users scheduled (one in each narrow band) at each time, the fraction of time any user is scheduled scaled by the number of users tends to L for large number of users. In making this observation, we used the hypothesis that the joint distribution of the L fading states is exchangeable; i.e., the probability that the maximum of the beamforming SNRs of the L fading states is any particular narrow band channel l is equal (to $1/L$). Since it is clear that a user cannot be scheduled at rate larger than the maximum of the beamforming rates corresponding to the fading states in each of the L narrow bands, the upper bound to the proportional algorithm takes a form similar to that of (16) with the quantity on the right side scaled by a factor of L. The result now follows. □

B Information Theoretic Capacity and Opportunistic Beamforming

In the comparison between opportunistic beamforming and space time codes in Section 4. we retained the TDMA strategy of transmitting to only one user (the user is decided by the scheduler) at any time slot. Given this strategy, the sum of the throughputs of the users with any use of the multiple antennas at the transmitter grows like

$$\log \mathrm{SNR} + o(\log \mathrm{SNR})$$

for high SNR. This TDMA strategy was motivated from an information theoretic result on the single transmit antenna downlink model. It is interesting to consider the information

theoretic capacity of the multiuser downlink communication problem with multiple antennas at the transmitter at high SNR. The appropriate channel model is that of a broadcast channel that is not degraded and the information theoretic capacity is not known (although some partial results have been obtained recently [3]). Focusing on the slow fading model, the following proposition characterizes the *sum capacity*, the sum of the throughputs of the users, at high SNR. Here we assume that the receivers track all the channels and the transmitter has full side information of the channels (both amplitude and phase).

Proposition 5 *The sum capacity at high SNR allows the following expansion:*

$$\min(K, N) \log \mathrm{SNR} + o(\log \mathrm{SNR}) \ . \tag{17}$$

The proof is furnished at the end of this section. We infer that the TDMA strategy loses *degrees of freedom*, equal to $\min(K, N)$, that is promised by information theory. In this section, we suggest a modification to the TDMA strategy which combined with opportunistic beamforming achieves all the degrees of freedom.

The conceptual idea is to have *multiple beams* at the same time. Separate pilot symbols are introduced on each of the beams and users feedback the SNR of each beam. Transmissions are scheduled to as many users as there are beams at each time slot. If there are enough users in the system, the user who is beamformed with respect to a specific beam and orthogonal to the other beams is scheduled on the specific beam. Suppose $K \geq N$ (if $K < N$ then we use only K antennas). Let $Q = [Q_{ln}]$ represent an $N \times N$ orthonormal matrix. The signal sent out of the antenna n at time t is

$$\sum_{l=1}^{N} x_l(t) Q_{ln}(t) \ .$$

Here $x_1, \ldots, x_N$ are the N independent data streams (in the case of coherent downlink transmission, these signals include pilot symbols as well). The data stream l has power and phase at antenna n set to $Q_{ln}(t)$ at time t. The orthonormal matrix $Q(t)$ is varied in time so that the individual components do not change abruptly in time. The signal received by user k at time t is

$$y_k(t) = \sum_{n=1}^{N} \sum_{l=1}^{N} x_l(t) Q_{ln}(t) h_{nk}(t) + z_k(t) \ .$$

Let us consider the slow fading model where the channel coefficients are not varying over the time scale of communication and focus on user k. Consider the scenario when the power and phases are at the following values.

$$Q_{ln} = \frac{h_{nk}^*}{\left(\sum_{n=1}^{N} \mid h_{nk} \mid^2\right)^{\frac{1}{2}}}, \qquad n = 1, \ldots, N \ . \tag{18}$$

The received signal at user k in this scenario is

$$y_k(t) = \left(\sum_{n=1}^{N} \mid h_{nk} \mid^2\right)^{\frac{1}{2}} x_l(t) + z_k(t) \ .$$

Thus user k is beamformed to beam l and is simultaneously orthogonal to the other beams in this setting. If there are enough users in the system, for every beam l, some user will be beamformed (and simultaneously orthogonal to the other beams) and analogous to Theorem 1, user k gets throughput approximately equal to (under the proportional fair algorithm)

$$\frac{N}{K} f\left(\frac{P\sum_{n=1}^{N} \mid h_{nk} \mid^2}{N\sigma^2}\right) . \tag{19}$$

Here we assumed that the total power transmitted P is split equally among the N independent data streams. It also follows that the total throughput of the system is

$$N \log \mathrm{SNR} + o(\log \mathrm{SNR}) .$$

We can make a rough estimate of the number of users required to achieve the performance of (19). In the "single beam" case, the number of independent variables was $2\,(N-1)$ with $N-1$ independent power fraction variations and $N-1$ independent phase variations. In the scenario of $1 \leq l \leq N$ beams, we can calculate the number of independent variables to be $2lN - l^2 - l$ (the dimension of the corresponding Stiefel manifold). When all the N beams are active, there are order N^2 number of independent variables as compared to order N in the single beam case. Thus, the number of users required grows very rapidly with the number of antennas.

We should evaluate the extra requirement on the system to support multiple beams. First, in the case of coherent downlink transmission, multiple pilot symbols, one set for each beam, have to be inserted and thus the fraction of pilot symbol power increases. Second, the receivers now track N separate beams and feedback SNRs of each on each of the beams. On a practical note, the receivers could feedback only the *best* SNR and the beam which yields this SNR without much degradation in performance. Thus with almost the same amount of feedback as the single beam scheme (amplitude alone), the modified opportunistic beamforming scheme yields a total throughput in a system with large number of users equal to that of the information theoretic limit with full (amplitude and phase) feedback at high SNR.

B.1 Proof of Proposition 5

We begin by recalling some notation. Fix K, number of receivers and N, number of transmit antennas. The channel from the transmitter to user k is given by (as in (1)):

$$y_k(t) = \sum_{n=1}^{N} h_{nk} x_n(t) + z_k(t) \; k = 1, 2, \ldots, K .$$

The difference from (1) is that the transmitted signal $\mathbf{x} \stackrel{\text{def}}{=} (x_1, \ldots, x_N)$ is a vector. The power constraint is now on the average of $\mathbf{x}^\dagger \mathbf{x}$. We are focusing on the slow fading model where h_{nk} do not change in time and we assume that these are perfectly known by the transmitter and the receivers. We make the assumption, true with probability 1 when the

channel coefficients are drawn independently from a continuous distribution like Rayleigh, that the matrix H defined by $H_{nk} \stackrel{\text{def}}{=} h_{nk}$ has full rank. Let us denote the sum capacity, the maximum sum of rates at which the transmitter can jointly reliably communicate with the receivers, by C_{sum}. We are interested in characterizing this quantity with respect to $\text{SNR} \stackrel{\text{def}}{=} \frac{P}{\sigma^2}$ where σ^2 is the variance of the noise $z_k(t)$.

An upper bound to the sum capacity is obtained by using the general outer bound given by Sato [13].

$$\begin{aligned} C_{\text{sum}} &\leq \max_{\text{distribution on } \mathbf{x} \text{ subject to} \mathbb{E}[\mathbf{x}^\dagger \mathbf{x}] \leq P} I\left(\mathbf{x}; y_1, \ldots, y_K\right) , && (20) \\ &\leq \log\det\left(I + \text{SNR} H H^\dagger\right) , && (21) \\ &\leq \min(K, N) \log \text{SNR} + o(\log \text{SNR}) . && (22) \end{aligned}$$

The upper bound in (20) follows by allowing the receivers to cooperate and jointly decode, (21) was arrived at by using the fact that the independent Gaussian distribution with variances P/N achieve the maximum and we used the full rank property of H to arrive at (22).

A lower bound to the sum capacity is obtained by the general inner bound by Marton [11].

$$C_{\text{sum}} \geq \sum_{k=1}^{K} \left(I(u_k; y_k) - h(u_k)\right) + h(u_1, \ldots, u_K) , \quad (23)$$

where $u_1, \ldots, u_K$ are allowed to be arbitrary random variables and the choice of the joint distribution of $\mathbf{x}$ and the u_k's is left open. Since we are deriving a lower bound, we will use the following policy. If $K > N$ we will only transmit positive rates to N users and if $N > K$ we will only use K of the transmit antennas. Thus we will henceforth assume that $K = N$. By a QR factorization, we can also assume without loss of generality that H is upper triangular. Consider the following specific choice of u_k's and the input $\mathbf{x}$. We pick $x_1, \ldots, x_N$ to be i.i.d. Gaussian distribution with variance $\frac{P}{N}$ each. We let

$$u_k = x_k + \left(\frac{\mid h_{kk} \mid^2}{\mid h_{kk} \mid^2 + \frac{N}{\text{SNR}}}\right) \sum_{l=1}^{k-1} h_{lk} x_l, \quad k = 1 \ldots N .$$

Analogous to the calculation in [4], we can evaluate the right side of (23) with this choice of joint distribution to arrive at

$$C_{\text{sum}} \geq \sum_{k=1}^{N} \log\left(1 + \frac{\text{SNR}}{N} \mid h_{kk} \mid^2\right) .$$

Using the full rank property of H and comparing with the upper bound of (22), the proof is complete. □

References

[1] S. M. Alamouti, "A simple transmitter diveristy scheme for wireless communications", *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1451-1458, Oct. 1998.

[2] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana and A. Viterbi, "CDMA/HDR: a bandwidth efficient high speed wireless data service for nomadic users", *IEEE Communications Magazine*, vol. 38(7), pp. 70-78, July 2000.

[3] G. Caire and S. Shamai, "On achievable rates in a multi-antenna broadcast downlink", *38th Annual Allerton Conference on Commun., Control and Computing*, Monticello, IL, Oct. 4-6, 2000.

[4] M. H. M. Costa, "Writing on dirty paper", *IEEE Transactions on Information Theory*, pp. 439 - 441, May 1983. Trans.

[5] H. A. David, *Order Statistics*, Wiley, 1970 (1st Edition).

[6] J-C Guey et al., "Signal Designs for transmitter diversity wireless communication system over Rayleigh fading channels", Proc. VTC'96, pp. 136-140.

[7] T. Hattori and K. Hirade, "Multitransmitter simulcast digital signal transmission by using frequency offset strategy in land mobile radio telephone system," *IEEE Trans. Vehicul. Technology*, vol. VT-27, pp. 231-238, 1978.

[8] A. Hiroike, F. Adachi and N. Nakajima, "Combined effects of phase sweeping transmitter diversity and channel coding", *IEEE Trans. Vehicul. Technology*, vol. 41, pp. 170-176, May 1992.

[9] Wen-Yi Kuo and M. P. Fitz, "Design and analysis of transmitter diversity using intentional frequency offset for wireless communications", *IEEE Transactions on Vehicular Technology*, vol.46(4), p.871-881, Nov. 1997. modulation for multiple

[10] R. Knopp, and P. Humblet, "Information capacity and power control in single cell multiuser communications", Proc. IEEE ICC 95, Seattle, Wa., June 1995.

[11] K. Marton, "A coding theorem for the discrete memoryless broadcast channel", *IEEE Transactions on Information Theory*, pp. 306 - 311, May 1979.

[12] S. Shakkottai and A. L. Stolyar, "Scheduling Algorithms for a mixture of real-time and non-real-time data in HDR", preprint.

[13] H. Sato, "An outer bound to the capacity region of broadcast channels", *IEEE Trans. on Information Theory*, pp. 374 - 377, May 1978.

[14] V. Tarokh, N. Seshadri and A. R. Calderbank, "Space-Time codes for high data rate wireless communication: performance, criterion and code construction", *IEEE Trans. on Information Theory*, vol. 44(2), pp. 744-765, Mar. 1998.

[15] V. Tarokh, H. Jafarkhani and A. R. Calderbank, "Space-Time block coding for wireless communications: performance results", *IEEE Journal on Select. Areas in Comm.*, vol. 17(3), pp. 451-460, Mar. 1999.

[16] V. Tarokh, H. Jafarkhani and A. R. Calderbank, "Space-Time block codes from orthogonal designs", *IEEE Transactions on Information Theory*, vol. 48(5), pp. 1456-1467, July 1999.

[17] V. Tarokh and H. Jafarkhani, "A differential detection scheme for transmit diversity", *IEEE Journ. Select. Areas in Commun.*, vol. 18(7), July 2000.

[18] TIA/EIA IS-856, "CDMA 2000: High rate packet data air interface specification", Nov. 2000.

[19] TIA/EIA IS-2000, *CDMA 2000*, Mar. 2000.

[20] D. Tse, "Optimal Power Allocation over parallel Gaussian channels", *Proc. of ISIT*, 1997.

[21] D. Tse, "Multiuser Diveristy and Proportional Fair Scheduling", in preparation.