EURASIP Journal on Applied Signal Processing 2004:5, 591–604

PhantomNet: Exploring Optimal Multicellular Multiple Antenna Systems

Syed A. Jafar
Electrical Engineering and Computer Science, University of California, Irvine, Irvine, CA 92697-2625, USA
Email: syed@uci.edu

Gerard J. Foschini
Bell Laboratories, Lucent Technologies, 791 Holmdel-Keyport Road, Holmdel, NJ 07733, USA
Email: gjf@lucent.com

Andrea J. Goldsmith
Wireless Systems Laboratory, Stanford University, Stanford, CA 94305-9505, USA
Email: andrea@ee.stanford.edu

Received 20 December 2002; Revised 11 August 2003

We present a network framework for evaluating the theoretical performance limits of wireless data communication. We address the problem of providing the best possible service to new users joining the system without affecting existing users. Since, interference-wise, new users are required to be invisible to existing users, the network is dubbed PhantomNet. The novelty is the generality obtained in this context. Namely, we can deal with multiple users, multiple antennas, and multiple cells on both the uplink and the downlink. The solution for the uplink is effectively the same as for a single cell system since all the base stations (BSs) simply amount to one composite BS with centralized processing. The optimum strategy, following directly from known results, is successive decoding (SD), where the new user is decoded before the existing users so that the new users' signal can be subtracted out to meet its invisibility requirement. Only the BS needs to modify its decoding scheme in the handling of new users, since existing users continue to transmit their data exactly as they did before the new arrivals. The downlink, even with the BSs operating as one composite BS, is more problematic. With multiple antennas at each BS site, the optimal coding scheme and the capacity region for this channel are unsolved problems. SD and dirty paper (DP) are two schemes previously reported to achieve capacity in special cases. For PhantomNet, we show that DP coding at the BS is equal to or better than SD. The new user is encoded before the existing users so that the interference caused by his signal to existing users is known to the transmitter. Thus the BS modifies its encoding scheme to accommodate new users so that existing users continue to operate as before: they achieve the same rates as before and they decode their signal in precisely the same way as before. The solutions for the uplink and the downlink are particularly interesting in the way they exhibit a remarkable simplicity and an unmistakable, near-perfect, up-down symmetry.

Keywords and phrases: channel capacity, dirty paper coding, duality, broadcast channel, successive decoding, multiple-input multiple-output systems.

1. INTRODUCTION

The rapid growth of cellular networks and the anticipation of ever increasing demand for higher data rates have expanded the scope of wireless research from single user, and single cell, and single antenna systems to multiuser multicellular systems employing multiple antennas. A traditional way of handling the multiantenna, multiuser, and multicellular system has been to reduce it to a single antenna, single user, and single cell system by orthogonally splitting the channel among the users in time/frequency/code/space, employing the base station antennas for sectoring/beamforming, and treating cochannel interference from other cells as noise. Moreover, since early wireless networks have been designed primarily for voice traffic, rate adaptation was not considered. This constrained approach may be simpler, but quite often it leads to suboptimal strategies. In order to estimate the absolute performance limits of these multidimensional systems, we need to explicitly account for the presence of multiple users, multiple antennas, and multiple cells on both the uplink and the downlink.

In this paper, where wireless data communication is highlighted, the focus is on finding the best transmit strategy. Due to the presence of a multiplicity of contending users, the best transmit strategy is not as straightforward as for a single-user system. Assigning limited communication resources to effect the best transmit strategy is particularly relevant for handling delay tolerant data traffic since helping some users typically amounts to slowing others. The best strategy, of course, depends on the priorities assigned to each user. Given the prioritization, say, for example, first-come-first-served (FCFS), we find here the optimum communication means under different criteria.

Although we will proceed with the FCFS prioritization in our presentation, our results hold for other means of prioritizing such as last-come-first-served, random ordering, or any scheme that predetermines an ordering among users.

We consider both the uplink and the downlink of a multiuser multicellular system using multiple antennas at both ends. We consider a system that evolves in time with new users entering the system and old users leaving the system. Using FCFS, our objective is to provide the best service possible to the new users as they enter the system, without penalizing the users already in the system. Thus each user in the system has a higher priority than the users that come after him. Subsequent users are served under the requirement that the previous ones are not affected: interference-wise, new users must be invisible to exiting users. Since for both the uplink and the downlink only earlier entrants interfere while later entrants are invisible, the network is dubbed PhantomNet. The strategies that affect this invisibility will be seen to be successive decoding (SD) for the uplink (a form of multiuser detection) and dirty paper (DP) coding for the downlink. In our network context, these strategies are particularly interesting both because of their simplicity as well as the unmistakable symmetry evident between uplink-downlink operation. Just how resources like base stations, bandwidth, spatial modes, and power are used is not preordained. Rather, under the FCFS regime, the network can self-organize the deployment of these communication resources.

The FCFS model assigns lower priority to new users. However, as previous users complete their transmission, the user moves up on the priority scale. So users that stay in the system longer tend to experience a better average service. In other words, shorter messages experience a lower average rate, while longer messages experience a higher average rate. It is therefore reasonable to expect that the FCFS scheduling algorithm would make the time required to transmit to different users' messages more equal.[1]

Our scope here is limited to the presentation of theoretical findings. These findings provide a tractable framework in which performance of multicellular, multiuser, and multiantenna wireless networks can be numerically evaluated through simulation. Information theoretic optimization is at the core of our approach. Simulation results with DP coding presented in [1] complement this work.

2. SYSTEM MODEL

Although we are ultimately interested in a multicellular system, for simplicity, we start with a single base station. Multiple base stations will be addressed in Section 7.

2.1. Uplink

The uplink is characterized by the following equation:

$$Y = \sum_{i=1}^{K} H_i X_i + N, \tag{1}$$

where Y is the received vector at the base station, K is the number of users currently active in the system, H_i is the flat-fading matrix channel of user i, and N is the additive white Gaussian noise (AWGN) vector at the base station.

Without loss of generality, we assume that the users are indexed by the order in which they arrive. So user 1 is the first user in the system, while user K is the last user to join the system. The users are subject to transmit power constraints given by

$$\operatorname{trace}\left[E\left[X_i X_i^{\dagger}\right]\right] \le P_i, \quad 1 \le i \le K. \tag{2}$$

Note that there is no data coordination between users, so the X_i are independent.

2.2. Downlink

Finding the optimal transmit strategy for the downlink with multiple antennas is a hard problem. This is because the multiple antenna downlink channel is a nondegraded broadcast channel and its capacity region is a long standing unsolved problem in information theory [2]. The optimal coding strategy for the multiple antenna downlink is therefore unknown. The special cases of the AWGN broadcast channel where the optimal coding strategy is known include the degraded broadcast channel (single transmit antenna at the BS), and the recently solved sum rate capacity of multiple user vector broadcast channel with multiple transmit antennas at the BS and at each of the mobiles [3, 4, 5, 6, 7]. While SD achieves capacity in the first case, DP coding based on the results of [8] achieves capacity in the latter. DP coding can also be shown to achieve capacity for the degraded AWGN broadcast channel. Note that for all these cases where the capacity is known, it is achieved with SD or DP coding and with Gaussian codebooks. For this reason, in this paper, we will restrict our downlink transmit strategies to these

[1] If one chooses instead a last-come-first-served model, short messages would see higher average rates, and long messages would see lower average rates. Thus last-come-first-served scheduling would make the time required to transmit different users' messages more disparate. The average number of simultaneously active users would reflect the average interference seen by the users. Overall, the choice of the scheduling algorithm for a system will depend on such criteria.

two coding schemes and we will assume that Gaussian codebooks are used. These assumptions may not be restrictive at all in case the conjectures about the optimality of Gaussian codebooks on the downlink can be established [9, 10]. Thus, our downlink model is given by the following equation:

$$Y_i = H_i \sum_{j=1}^{K} X_j + N_i, \tag{3}$$

where Y_i, X_i, H_i, and N_i are the output vector, the input vector, the channel matrix, and the AWGN vector for user i. For both SD and DP coding strategies, the input vectors corresponding to different users are independent. As in the uplink model described earlier, the downlink model also assumes that the users are indexed by the order in which they arrive. Further, the power in each user's input vector is given by

$$\operatorname{trace}\left[E\left[X_i X_i^{\dagger}\right]\right] \leq P_i, \quad 1 \leq i \leq K. \tag{4}$$

We would also like to point out that a "ranked known interference" scheme based on the results of [3] was used in [11] to minimize the delay in a multiuser multicellular system with multiple antennas at the base station and a single receive antenna at each mobile. While the scheme itself is suboptimal and limited in scope to a single receive antenna at each mobile, it is another example of a simple way to perform resource allocation on the downlink. The results of [11] are interesting and complement this work.

Unlike the uplink where users have individual power constraints, on the downlink, it is possible to redistribute transmit powers across users without changing the total transmitted power from the base station. Thus the downlink is typically characterized by a sum power constraint.

For both the uplink and the downlink, the channel is assumed to experience slow and flat fading. Note that, with a sufficiently refined partition of the frequency band, a frequency-selective fading channel can be viewed as a number of parallel spectrally disjoint noninterfering essentially flat subchannels. It follows that, for any desired accuracy, the resulting channel matrix is equivalent to a block-diagonal flat-fading channel matrix. Hence the flat channel analysis presented here extends to frequency-selective fading in a straightforward manner. We assume that the channel matrices are perfectly known to the BS. The users are assumed to know their own channel and the spatial covariance structure of the sum of the noise and the relevant interference seen at the receiver.

Lastly, since the notion of *substreams* comes up in later sections, we elaborate what we mean by it. Note that a user's input vector X_i may further be composed of several *independent* vectors $X_{i1}, X_{i2}, \ldots$. This amounts to splitting the total rate for that user among several *substreams*. For a single user, it can be shown that rate splitting does not decrease capacity. For a single-antenna multiple access AWGN channel, rate splitting allows all points in the capacity region to be achieved without time-sharing [12]. For our purpose, splitting a users' power into substreams allows the substreams from different users to be interleaved in any manner with respect to the encoding/decoding order.

3. PROBLEM DEFINITION

Based on the FCFS model, our primary objective is to accommodate new users only to the extent that the users that are already active in the system are not affected. While this constitutes the general idea, to be precise, we need to distinguish between the following two cases.

Existing users are unaffected (preserving rates)

This would mean that the existing users continue to have the same rates as before. However, this leaves open the possibility that the existing users may adjust their transmit strategy on the uplink or their receive strategy on the downlink in some way to accommodate the new user. For example, on the downlink, it is conceivable that if superposition coding was used, then the existing users may need to decode and subtract out the new users signal before detecting their own signal. If this allows the existing users to achieve the same rates as before, we say that the existing users are not affected, or the rates are preserved.

Existing users are strictly unaffected (making the accommodation of new users invisible)

We could be more strict in our problem statement. We could demand that the new users be accommodated in such a way that not only do the existing users continue to achieve the same rates as before but also they are completely oblivious to the presence of new users. That is, the existing users' transmitters/receivers on the uplink/downlink continue to process the input data stream/received signal exactly as before to generate the transmitted signal/output data stream. Thus the only changes needed to accommodate the new user are made at the base stations. To distinguish this case from the previous one, we say that the existing users are *strictly* unaffected, or the new users are invisible.

Within each of the cases mentioned above, there are several, more or less equally significant, problems that one can pose. We list these problems in Sections 3.1 and 3.2 for the uplink and the downlink, respectively. We will see later that all the uplink problems really amount to the same problem—basically the same solution procedure covers all of the uplink variations. Among the downlink problems, we will encounter some substantive differences.

3.1. Uplink

On the uplink, the user's transmit power is the limiting factor. So, for the uplink, the first set of problems UP1a and UP1b (uplink problems 1a and 1b) that we wish to solve are as follows.

UP1a (preserving rates). Allocate the maximum possible rate to user K (new user) with transmit power P_K such that the existing users' rates are not affected. Note that this allows the existing users to modify their transmit strategy to accommodate the new user so long as their rates are unaffected.

UP1b (making the new user invisible). Allocate the maximum possible rate to user K (new user) with transmit power P_K such that the existing users are *strictly* unaffected. Note that now, we require that the new user be invisible to the existing users, that is, the existing users must not modify their transmit strategy or their rates. Thus, the existing users are, in effect, oblivious to the presence of the new user.

We also briefly address the alternate problem where users have certain rate requirements and wish to achieve those rates with the minimum possible transmit power as follows.

UP2a (preserving powers). Determine the minimum possible transmit power for a new user K with rate requirement R_K such that the existing users' transmit powers are not affected.

UP2b (making the new user invisible). Determine the minimum possible transmit power for a new user K with rate requirement R_K such that the existing users are strictly unaffected.

3.2. *Downlink*

On the downlink, each base station distributes the total transmit power among the users it serves. Thus, unlike the uplink where each user has an individual power constraint, the downlink is characterized by a sum power constraint instead. The coding schemes we consider for the downlink are SD and DP. A brief description of these schemes is presented later. In particular, we wish to determine the following.

DP1. Is DP or SD a better scheme for the downlink in general?

For FCFS scheduling, the corresponding problems on the downlink would be as follows.

DP2a (preserving rates). Determine the maximum possible rate for user K subject to a total transmit power $P_1 + P_2 + \cdots + P_K$ such that existing users' rates are not affected.

DP2b (making the new user invisible). Determine the maximum possible rate for a user K subject to a total transmit power $P_1 + P_2 + \cdots + P_K$ such that existing users are strictly not affected.

Note that in problems DP2a and DP2b, the BS adds a power P_K to the total power to accommodate a new user (user K) into the system. The powers $P_1, P_2, \ldots, P_K$ determine how the rates are allocated to the users and need not be the actual transmitted powers in each user's input signal.

Note that as the channel changes, the users' rates/powers may change. So for each channel realization, we solve the FCFS scheduling problems listed above. The assumption that the channel varies slowly is important in this respect.

4. MIMO CAPACITY REVIEW

Before proceeding with the solutions to the problem defined in Section 3, we briefly visit the MIMO capacity expression. Consider the MIMO channel

$$Y = HX + \sum_{i=1}^{I} H_i X_i + N. \tag{5}$$

Here, X is the desired signal and $X_1, X_2, \ldots, X_I$ represent I independent interference signals. All input signals are assumed to be Gaussian with input covariance matrices $Q, Q_1^\star, Q_2^\star, \ldots, Q_I^\star$, respectively. Recall that the input covariance matrices identify the optimal spatial eigenmodes and the optimal power allocation across those eigenmodes. The input covariance matrices of the interfering signals $Q_i^\star$ are already fixed. We are interested in the optimal input covariance matrix $Q^\star$ for the desired signal X subject to total power constraint $\operatorname{trace}(Q) \le P$. The H matrices represent the channels. The noise is assumed to be AWGN with covariance matrix normalized to identity. Note that this could apply to either the downlink or the uplink.

Since the interference is independent of the signal, the capacity of this channel is

$$\begin{aligned} C &= \max_Q I(X;Y) \\ &= \max_Q h(Y) - h(Y \mid X) \\ &= \max_Q h\left(HX + \sum_{i=1}^{I} H_i X_i + N\right) - h\left(HX + \sum_{i=1}^{I} H_i X_i + N \middle| X\right) \\ &= \max_Q h\left(HX + \sum_{i=1}^{I} H_i X_i + N\right) - h\left(\sum_{i=1}^{I} H_i X_i + N\right) \\ &= \max_Q \log \left| I + HQH^\dagger + \sum_{i=1}^{I} H_i Q_i^\star H_i^\dagger \right| \\ &\quad - \log \left| I + \sum_{i=1}^{I} H_i Q_i^\star H_i^\dagger \right| \\ &= \max_Q \log \left| I + \left(I + \sum_{i=1}^{I} H_i Q_i^\star H_i^\dagger \right)^{-1} HQH^\dagger \right|. \end{aligned} \tag{6}$$

Thus the capacity of this channel can be expressed as $C = \log |I + (I + \sum_{i=1}^{I} H_i Q_i^\star H_i^\dagger)^{-1} HQ^\star H^\dagger|$. The optimal $Q^\star$ is determined as follows.

Since $\log |I + AB| = \log |I + BA|$, we can also express the capacity as

$$C = \max_Q \log \left| I + \left(I + \sum_{i=1}^{I} H_i Q_i^\star H_i^\dagger \right)^{-1/2} \times HQ^\star H^\dagger \left(I + \sum_{i=1}^{I} H_i Q_i^\star H_i^\dagger \right)^{-1/2^\dagger} \right| \tag{7}$$

$$= \max_Q \log \left| I + \tilde{H} Q \tilde{H}^\dagger \right|, \tag{8}$$

where

$$\tilde{H} = \left(I + \sum_{i=1}^{I} H_i Q_i^{\star} H_i^{\dagger}\right)^{-1/2} H. \tag{9}$$

But (8) is the familiar MIMO capacity expression for a single user with channel $\tilde{H}$ in the presence of AWGN and without interference. The optimal input covariance matrix Q is obtained by the well-known waterfilling algorithm over the eigenmodes of $\tilde{H}$ [13].

Thus, in summary, the capacity for the channel (5) is given by

$$C = \log \left| I + \left(I + \sum_{i=1}^{I} H_i Q_i^{\star} H_i^{\dagger}\right)^{-1} H Q^{\star} H^{\dagger} \right|, \tag{10}$$

where $Q^{\star}$ is the optimal input covariance matrix obtained by waterfilling over the *effective* channel (9). Similar expressions appear quite frequently in later sections. To avoid repetition, instances of the same expressions presented later may be less descriptive. We advise the reader to refer back to this section and the references for details.

5. UPLINK SOLUTION

The uplink presents a relatively simple problem since the capacity region and the optimal coding strategy are known even with multiple antennas at the BS and the mobiles [14]. The desired solution is easily seen to be the well-recognized points on the capacity region corresponding to SD of users in a particular order. However, for the sake of completeness, and to strike a parallel with the downlink solutions presented later, we provide the solution and a self-contained proof as follows.

The solution to the first uplink problem UP1a (preserving rates) is given by the following theorem.

Theorem 1. *The optimal set of rates $R_i^{\star}$ on the uplink is*

$$R_i^{\star} = \log \left| I + \left(I + \sum_{j=1}^{i-1} H_j Q_j^{\star} H_j^{\dagger}\right)^{-1} H_i Q_i^{\star} H_i^{\dagger} \right|, \tag{11}$$

where $Q_i^{\star}$ is the optimal input covariance matrix obtained by waterfilling over the eigenmodes of the effective channel matrix $(I + \sum_{j=1}^{i-1} H_j Q_j^{\star} H_j^{\dagger})^{-1/2} H_i$ subject to the power constraint $\operatorname{trace}(Q_i) = P_i$.

In other words, an optimal strategy for the uplink is to use SD (multiuser detection with successive interference cancellation) at the base station in the inverse order of the user's indices. The new user gets decoded first and his signal is subtracted out so that the existing users do not see him as interference. The highest rate that the new user can support without affecting existing users is simply given by the single-user waterfilling solution treating the existing users' signal as colored Gaussian noise.

Proof. We start with user 1. Ignoring the rest of the users, the highest rate he can support with power P_1 is

$$R_1^{\star} = \max_{p_1(\cdot)} I(X_1; H_1 X_1 + N), \tag{12}$$

where the maximization is over all distributions $p_1(X_1)$ that satisfy the power constraint (2). The optimal $p_1^{\star}(\cdot)$ is the well known zero-mean vector Gaussian distribution with covariance matrix $Q_1^{\star}$ determined by waterfilling over the eigenmodes of H_1. Let $X_1^{\star} \sim p_1^{\star}$. Note that the users' channels H_i are known and therefore H_1 is not a random variable in (12).

Now for the user 2, ignoring all but the user 1, from the multiple access capacity region, we have

$$R_1 + R_2 \leq \max_{p_1(\cdot), p_2(\cdot)} I(X_1, X_2; H_1 X_1 + H_2 X_2 + N). \tag{13}$$

But R_1 and p_1 are already determined by the user 1. So we have

$$R_2^{\star} = \max_{p_2(\cdot)} I(X_1^{\star}, X_2; H_1 X_1^{\star} + H_2 X_2 + N) - R_1^{\star}, \tag{14}$$

$$\begin{aligned} R_2^{\star} = {} & \max_{p_2(\cdot)} I(X_1^{\star}, X_2; H_1 X_1^{\star} + H_2 X_2 + N) \\ & - I(X_1^{\star}; H_1 X_1^{\star} + N), \end{aligned} \tag{15}$$

$$\begin{aligned} R_2^{\star} = {} & \max_{p_2(\cdot)} I(X_2; H_1 X_1^{\star} + H_2 X_2 + N) \\ & + I(X_1^{\star}; H_1 X_1^{\star} + H_2 X_2 + N | X_2) \\ & - I(X_1^{\star}; H_1 X_1^{\star} + N), \end{aligned} \tag{16}$$

$$\begin{aligned} R_2^{\star} = {} & \max_{p_2(\cdot)} I(X_2; H_1 X_1^{\star} + H_2 X_2 + N) \\ & + I(X_1^{\star}; H_1 X_1^{\star} + N) - I(X_1^{\star}; H_1 X_1^{\star} + N), \end{aligned} \tag{17}$$

$$R_2^{\star} = \max_{p_2(\cdot)} I(X_2; H_1 X_1^{\star} + H_2 X_2 + N), \tag{18}$$

where (16) follows from the chain rule of mutual information and (17) follows from the independence of $X_1^{\star}$ and X_2. Note that this corresponds to decoding user 2 while treating user 1 as noise. Thus, at the base station, user 2 is decoded first and his signal is subtracted to obtain a clean channel for user 1. The optimal input distribution for user 2 is the waterfill distribution over the eigenmodes of $(I + H_1 Q_1^{\star} H_1^{\dagger})^{-1/2} H_2$.

Proceeding in this fashion, we obtain the result of Theorem 1. □

It is interesting to note the simplicity of the solution. Note that the SD scheme requires only the BS to make some changes in the way it decodes the received signal. Specifically, the BS needs to decode the new user and subtract his signal before proceeding to decode the existing users' signals. However, the existing users themselves do not need to do anything different because of the new user. Thus the new user is completely invisible to existing users. Thus, we conclude that on the uplink, an optimal strategy that leaves the existing users' rates unaffected also leaves the existing users unaffected. In particular an optimal solution to UP1a (preserving rates) is also the optimal solution to UP1b (making the new user invisible).

The second pair of uplink problems UP2a (preserving powers, while using minimum additional power to meet a new user's rate) and UP2b (making the new user invisible, while meeting his rate with minimum additional power) are also very similar to UP1a and UP1b. Clearly for the user 1, the required transmit power is the one that achieves a capacity equal to his required rate R_1 with optimal waterfilling over his channel. In order for user 1's transmit power to be unaffected by user 2, the BS must decode user 2 before user 1. This also ensures that user 1 is not affected by user 2. Therefore, user 2 must see user 1 as noise. The required transmit power for user 2 is the one that achieves a capacity equal to his required rate R_2 with optimal waterfilling over his channel in the presence of colored noise due to the interference from user 1's signal. Thus, except that we know the rates and we need to solve for the transmit powers, the solution is the same as given by Theorem 1. Again UP2a and UP2b have the same solution.

6. DOWNLINK

6.1. Successive decoding and dirty paper

We begin this section with a brief summary of the key features of the SD and DP schemes. The details can be found in references.

SD is the well-known strategy, where several substreams are encoded directly on the channel input alphabet and *independent* of each other. Figure 1 shows an SD encoder. If a user has access to all codebooks, then he can decode any substream that is encoded at a rate lower than the capacity of his channel for that substream's input covariance matrix and treat other simultaneously transmitted codewords as noise. This allows him to reconstruct the transmitted codeword for the decoded substream and subtract its effect from the received signal, thus obtaining a cleaner channel for detecting other substreams.

With this strategy, a user may need to decode several codewords carrying other users' data and subtract their effect before he achieves a channel good enough to decode the codeword carrying his own data. Notice from Figure 1 that each encoder operates independent of all the other encoders.

Now, without loss of generality, we can assume that the substreams are encoded in some order, one after the other. This means that while choosing the codeword $\mathcal{C}_i^n$ for the ith substream, the transmitter has precise, noncausal information about the interference caused by all the $i-1$ substreams that have already been encoded. This brings us into the realm of DP coding. Figure 2 shows a DP encoder. Notice that unlike the SD scheme illustrated in Figure 1, where each encoder operates independent of the rest, in the DP scheme, there is a definite order such that the output of each encoder depends not only on the input substream data but also on the outputs of the encoders before it. This is possible because the encoders are collocated at the base station which allows them to cooperate perfectly.

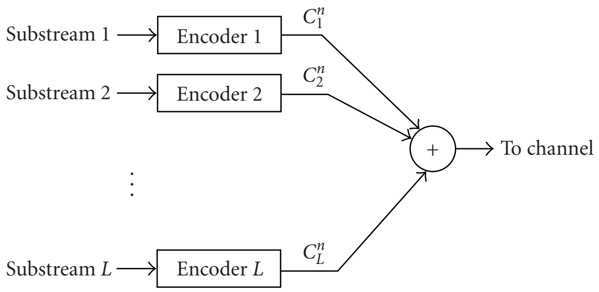


FIGURE 1: Encoding of L substreams in a successive decoding scheme.

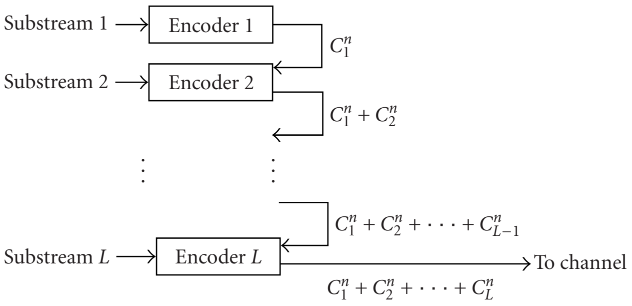


FIGURE 2: Encoding of L substreams in a dirty paper scheme.

The most powerful aspect of the DP scheme comes from the interesting work of Costa [8]. This paper presented the following result.

Costa's dirty paper result

Consider the scalar channel

$$Y_i = X_i + S_i + N_i, \tag{19}$$

where at each instant $i \in \mathbb{Z}^+$, Y_i is the output symbol, N_i is AWGN with power P_N, X_i is the input symbol constrained so that $E[X_i^2] \leq P_X$, and S_i is the interference symbol generated according to a Gaussian distribution. Now suppose the entire realization of the interference sequence $S_1, S_2, \ldots$ is known to the transmitter noncausally, that is, before the beginning of the transmission. This information is not available at the receiver. Then the capacity of the channel is given by

$$C = \log\left(1 + \frac{P_X}{P_N}\right), \tag{20}$$

irrespective of the power in the interference signal. In other words, if the interference is known to the transmitter beforehand, the capacity is the same as if the interference was not present. The capacity-achieving input distribution is $X \sim \mathcal{N}(0, P_X)$. Further, the channel input X and the interference S are independent.

Costa's result assumed a Gaussian distribution for the interference. The coding scheme described in [8] requires a knowledge of the distribution of the interference for designing the codebooks. Thus, if the statistics of the interference changed from one codeword to another, the receiver would have to be informed and it would have to switch to a different codebook. Thus, with Costa's scheme, even though the capacity of a channel with interference known only to the transmitter would be the same as without it, the receiver would have to be informed about any change in the interference statistics so it can use the correct codebook.

Recent work by Erez et al. [15] showed that lattice strategies can be used to extend the Costa's result to *arbitrarily* varying interference. Their scheme is able to handle arbitrarily varying interference by communicating modulo a fundamental lattice cell and using dithering techniques. It is this lattice strategy that we imply by the term DP coding in this paper. For a detailed exposition of the scheme and the required background, see [15, 16, 17, 18].

Although Costa's work in [8] and the recent work of Erez et al. in [15] assume a scalar channel, the extension to the complex matrix channel is straightforward. A MIMO system with the channel matrix H known to both the transmitter and the receiver can be transformed into several parallel noninterfering *scalar* channels by a singular value decomposition [19] of the channel. Thus, it is easily verified that Costa's result carries through to the MIMO system with arbitrary interference and we have the following.

Extension to complex MIMO systems with arbitrarily varying interference

Consider the MIMO channel

$$Y_i = HX_i + S_i + N_i, \tag{21}$$

where H is the channel matrix known to both the transmitter and the receiver and at each instant $i \in Z^+$, Y_i is the output vector, N_i is AWGN vector with covariance matrix Q_N, X_i is the input vector constrained so that $Q_X = \operatorname{trace}(E[X_i X_i^\dagger]) \leq P_X$, and S_i is an arbitrarily varying interference vector. All symbols are complex. Now suppose the entire realization of the interference sequence $S_1, S_2, \ldots$ is known to the transmitter non-causally. Then the capacity of the channel is given by

$$C = \max_{Q_X : \operatorname{trace}(Q_X) \leq P_X} \log \frac{|HQ_X H^\dagger + Q_N|}{|Q_N|}, \tag{22}$$

irrespective of the power in the interference signal. In other words, if the interference is known to the transmitter beforehand, the capacity is the same as if the interference was not present. It is worth mentioning that this does assume that both the transmitter and receiver have access to a common source of randomness to allow the dithering operation. The capacity-achieving input distribution is $X \sim \mathcal{N}(\mathbf{0}, Q_X)$. Further, the channel input X and the interference S are independent.

Unlike Costa's scheme, the DP scheme works for arbitrarily varying interference. Therefore, no knowledge of interference statistics is required at the receiver. Thus, even if the interference statistics change from one codeword to another, the receiver continues to operate exactly the same way. This property in particular is crucial for our FCFS scheduling problem.

An important feature of the DP scheme is that the capacity-achieving codes are not the channel input symbols $\mathcal{C}_i^n$ but the functions used to map the data and the transmitter side information to the channel input alphabet. Since the coding is not performed on the channel input alphabet itself, even if one decodes the data carried by a substream, it is not possible to subtract the effect of the transmitted symbols of the substream and obtain a cleaner channel. For example, refer to Figure 2. Decoding the ith substream does not allow a user to reconstruct the transmitted symbols $\mathcal{C}_i^n$ and therefore the user cannot subtract out $\mathcal{C}_i^n$ to obtain a cleaner channel.

In Figure 2, before encoding substream i, the transmitter knows the interference from substreams $1, 2, \ldots, i-1$. Thus the capacity achieved by substream i is the same as if substreams $1, 2, \ldots, i-1$ were not present. The interference from substreams $i+1, i+2, \ldots, L$ is not known and so it must be treated as noise.

To highlight the distinction between SD and DP, consider the following example of a broadcast system with two encoded substreams: substream 1 and substream 2. With SD, especially on a nondegraded broadcast channel, it is possible that one user can decode and cancel substream 2 before decoding substream 1, and at the same time another user with a different channel can decode and cancel substream 1 before decoding substream 2. Thus the decoding order may vary from user to user. On the other hand, with DP, there is a *fixed encoding order* such that the substreams encoded later achieve the same capacity as if the substreams encoded before them were not present. Moreover, the substreams encoded earlier can achieve a capacity no higher than that achievable by treating all substreams encoded after them as noise. In a nutshell, in SD, the encoding order is irrelevant and the optimal decoding order may vary from one user to another. In DP, there is no notion of decoding order. Instead, there is only one encoding order, where each substream has a unique position relative to every other substream. For each receiver, this unique order decides which substreams have to be treated as noise and which substreams do not impact the capacity of its own substream.

6.2. Solution to DP1 (DP versus SD)

The first problem we address on the downlink is to determine whether SD or DP is a better scheme in general. Before stating the solution, we see why it is not trivial. Consider two substreams intended for two different users. With DP, one of the users (the one encoded second) can achieve the same capacity as if the other user was not present. However, the other user (who was encoded first) must treat this user as noise and his capacity is reduced. With SD on the other hand, depending on the users' channels and the input covariance matrices, several situations are possible. It could be that the channels are such that each user can decode the other user's substream and subtract it before decoding his own substream. This seems to be better than DP. However, it could also happen that the channels are such that neither user can decode the other user's substream. In that case, SD would be worse than DP. Since it is the downlink, one can also optimize the transmit power across users while keeping the same total transmit power. Further, the rate regions may not be convex. In such a case, we can make the rate region convex by including rate vectors achievable with time-sharing. With all these possibilities, the question as to whether SD or DP is the better strategy on the downlink does not seem to have an obvious answer.

With the following theorem, we show that DP is the better downlink strategy in general.

Theorem 2. *Subject to a sum power constraint, the set of rate vectors achievable with SD and time-sharing is also achievable with DP and time-sharing.*

In other words, the convex hull of the achievable rate region with SD is completely contained within the convex hull of the achievable rate region with DP.

Proof. We prove this by showing that the boundary of the achievable rate region with SD and time division is contained within the boundary of the achievable rate region with DP and time-sharing. Note that in either scheme, the points in the interior can always be attained by throwing away some codewords.

The boundary points of the rate region are obtained by maximizing

$$\sum_{i=1}^{K} \mu_i R_i \tag{23}$$

for all $\vec{\mu}$ such that $\vec{\mu} \geq \vec{0}$ and $\sum_{i=1}^{K} \mu_i = 1$.

Let $\mathcal{R}^{\text{SD}}$ and $\mathcal{R}^{\text{DP}}$ denote the sets of rate vectors achievable with SD and DP, respectively. Note that in order to prove the result of Theorem 2, it suffices to prove that for all $\vec{\mu}$,

$$\max_{\vec{R}\in\mathcal{R}^{\text{DP}}} \sum_{i=1}^{K} \mu_i R_i \geq \max_{\vec{R}\in\mathcal{R}^{\text{SD}}} \sum_{i=1}^{K} \mu_i R_i. \tag{24}$$

In order to prove (24), we assume without loss of generality that the users' priorities are arranged as $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. We start with the SD scheme and show that DP can achieve at least the same value of $\vec{\mu}\cdot\vec{R}$. Let $\vec{R}^{\text{SD}}$ be the rate vector that maximizes $\vec{\mu}\cdot\vec{R}$ with SD. Without loss of generality, we can assume that $\vec{R}^{\text{SD}}$ does not use time-sharing. This is because simple linear programming tells us that a rate vector corresponding to time-sharing between several different rate vectors is a convex combination of those rate vectors and therefore cannot achieve a higher value of $\vec{\mu}\cdot\vec{R}^{\text{SD}}$ than the best of those rate vectors.

Let the total number of substreams being transmitted be L. Further, and again without loss of generality, we label the substreams from 1 to L such that if $i < j$ and substream i carries data for user $u(i)$ and substream j carries data for user $u(j)$, then $\mu_{u(i)} \geq \mu_{u(j)}$. That is, the substreams are arranged in decreasing order of the priority of the user whose data they are carrying. For multiple substreams carrying the same user's data, we label them in the order in which they are decoded by that user.

Now note that no user can decode a substream carrying data for a user with a lower priority. This is easily proved by contradiction as follows. Suppose that user A can decode a substream that carries user B's data at a rate r. Now if user A has a higher priority than user B, that is, if $\mu_A > \mu_B$, then we can increase $\vec{\mu}\cdot\vec{R}^{\text{SD}}$ by simply having the substream carry user A's data instead of user B's data at the same rate, r so that,

$$\vec{\mu}\cdot\vec{R}(\text{new}) = \vec{\mu}\cdot\vec{R}^{\text{SD}} - \mu_B r + \mu_A r > \vec{\mu}\cdot\vec{R}^{\text{SD}}. \tag{25}$$

But this is a contradiction since we assumed that the rate vector $\vec{R}^{\text{SD}}$ maximizes $\vec{\mu}\cdot\vec{R}$ over all rate vectors $\vec{R}$ achievable with SD and without time-sharing.

In light of this observation, it is clear that while decoding substream l, the intended user must treat substreams $l+1$ to L as noise. The substreams 1 to $l-1$ may or may not be treated as noise depending upon whether it is possible to decode and subtract those substreams or not. So with SD, the rate achieved on the lth substream is no greater (could be smaller) than r_l, where r_l is the achievable rate when the substreams $l+1$ to L are treated as noise while substreams 1 to $l-1$ are not present. Next, we show that DP can achieve r_l on each of these substreams.

Suppose we use DP to encode the L substreams in the order in which they are labeled. Then the lth substream sees substreams $l+1$ to L as noise since these substreams are encoded *after* substream l and therefore the interference caused by them is not known. However, since substreams 1 to $l-1$ have already been encoded, they present known interference to substream l and therefore do not affect the data rate that substream l is capable of supporting. Thus DP allows substream l a rate r_l that is at least as large as the maximum allowed rate for that substream in the optimum SD rate vector that maximizes $\vec{\mu}\cdot\vec{R}$. This proves (24) and completes the proof of Theorem 2. □

We can also easily extend this theorem to show that the achievable rate region of the pure DP scheme includes the achievable rate region of not only the pure SD scheme but also any hybrid scheme where some users use SD while others use DP. Lastly, we need time-sharing for this result because the achievable rate region for SD and DP without time-sharing may not be convex.

6.3. Downlink solutions for DP2a (preserving rates) and DP2b (making the accommodation of new users invisible)

In DP2a, we are only requiring rate conservation in dealing with the Kth user. This leaves open the possibility that, in meeting the earlier rates, if the earlier users are handled in a different way than before, we can actually achieve a strictly greater rate for the Kth user. Indeed, in some instances, a greater rate is possible. This DP2a problem is exceptional in that we encounter the most difficult of the optimization problems in this paper and a solution is only presented for a special case. In the general case, *based on the conjecture in [9]*, a solution can, in theory, be obtained by solving a number of convex programming problems to obtain the achievable rate region with DP coding [20]. However, the complexity of this is exponential in the number of users.

In problem DP2b, we insist that earlier users be treated exactly as before. Later users must be invisible (phantoms) to earlier ones. It turns out that, with this added constraint, we can obtain a complete solution. Moreover, as we will see in Section 7, a solution is possible for the full multiple base station setup.

6.3.1. Solution to DP2a (preserving rates)

Next, we address the problem of assigning the maximum rate to new user K subject to total power $P_1 + P_2 + \cdots + P_K$ such that the existing users' *rates* are not affected. So we wish to allocate the maximum possible rates to each user such that

(i) user 1 gets $R_1^\star$, the maximum rate possible with power P_1 as if no other user was present,

(ii) user 2 gets $R_2^\star$, the maximum rate possible with *total* power $P_1 + P_2$ such that user 1 still gets $R_1^\star$ and as if users $3, \ldots, K$ were not present,

(iii) user K gets $R_K^\star$, the maximum rate possible with total power $P_1 + P_2 + \cdots + P_K$ such that users 1 through $K-1$ still get rates $R_1^\star$ through $R_{K-1}^\star$.

While the overall optimization seems hard for the general multiple antenna broadcast system, limiting the number of transmit antennas at the base station to one does lead to a simple solution. A single transmit antenna at the base station makes the channel degraded and the optimality of Gaussian inputs is established from Bergman's proof in [21]. Note that although Bergman's proof is for scalar broadcast channels, that is, broadcast channels with a single transmit antenna at the base station and a single receive antenna at each user, the vector broadcast channel with a single antenna at the base station and multiple receive antennas at each user is easily seen to be equivalent to the scalar broadcast channel [22]. Thus, in this case, the capacity region is well known and we do not need the conjecture of [9]. Next, we present this solution to gain some insight.

With a single transmit antenna at the base station, the downlink is a degraded broadcast channel. Even with multiple receive antennas, each user can perform spatial matched filtering to yield a *scalar* AWGN channel for himself [22]. For this channel, the broadcast capacity is well known and either SD or DP can be used to achieve any point in the capacity region. In particular, all the rate points can be achieved with SD/DP with the *same* encoding/decoding order [23]. The user with the weakest channel is decoded/encoded first so that he sees everyone else as noise. The decoding/encoding proceeds in the order of the users' channel strengths so that weaker users who cannot decode the stronger users are forced to treat their signal as noise while the stronger users can decode the weaker users' data, and are therefore unaffected by the presence of weaker users. Thus, in this case, the encoding/decoding order is decided by the users' channels and not by the order of users' arrivals or their relative priorities.

For each channel state, we calculate the optimal rates and powers in an iterative fashion as follows. We start with only user 1 in the system with total power P_1 and find $R_1^\star$. Then we incrementally add users to the system, in the order $2, 3, \ldots, K$, each time finding the optimal rates for the set of users in the system with total power given by the sums of the powers of those users. The ith user is added as follows.

(1) Arrange the users in the order of their channel strengths.

(2) The users with a stronger channel than user i are not affected. That is, they continue to use the same power and rates as before.

(3) The users with a weaker channel than user i have to treat user i as noise. So the additional power P_i available to the system is distributed among user i and the weaker users so that the weaker users can sustain the same rates as before.

The optimal distribution of the additional power among the new user and the weaker users requires only a one-dimensional optimization and is easily obtained. Proceeding in this fashion, after the Kth user has been added, we obtain the optimal rate and power allocation for all the users in the system. Note that this is the optimal allocation because the rate vector obtained in this fashion lies on the boundary of the capacity region.

While this solution does not affect the existing users' rates, it does affect the existing users in that they may have to decode the new user before decoding their own signals if SD is used. If DP is used, then the existing users may have to see the new user as spatially colored noise. They are still able to achieve the same rates as before because they have a higher power. Thus, the solution does not allow the existing users to continue operating as before.

Next, we present a solution that gives the new user K the maximum rate possible with total transmit power $P_1 + P_2 + \cdots + P_K$ without affecting existing users.

6.3.2. Solution to DP2b (making the accommodation of new users invisible)

Theorem 3. *The optimal set of rates $R_i^\star$ on the downlink such that existing users are oblivious to the presence of the new users is given by*

$$R_i^\star = \log \left| I + \left(I + \sum_{j=1}^{i-1} H_i Q_j^\star H_i^\dagger \right)^{-1} H_i Q_i^\star H_i^\dagger \right|, \qquad (26)$$

where $Q_i^\star$ is the optimal input covariance matrix obtained by waterfilling over the eigenmodes of the effective channel matrix $(I + \sum_{j=1}^{i-1} H_i Q_j^\star H_i^\dagger)^{-1/2} H_i$ subject to the power constraint $\operatorname{trace}(Q_i) = P_i$.

In other words, an optimal strategy for the downlink that does not allow new users to affect existing users is to use DP encoding at the base station in the inverse order of the user's indices. The new user gets encoded first so his signal is a known interference and the existing users' rates do not get affected. The highest rate that the new user can support without affecting existing users is simply given by the single-user waterfilling solution treating the existing users' signal as colored Gaussian noise. A simple example to illustrate the optimal downlink scheme is presented after the proof.

Proof. DP's ability to handle arbitrarily varying interference makes it the obvious choice in this case. Using SD would require existing users to decode the new user, thus acknowledging the new user's presence. However, since DP is able to handle arbitrary interference, it does not matter if the interference known to the ith user's encoder comes from users $i, i+1, \dots, K-1$ or from users $i, i+1, \dots, K$. The rate and decoding strategy for user i depend only on the interference from users $1, 2, \dots, i-1$ that came before him and whose signals must be treated as noise for user i.

Note that time-sharing and rate-splitting are not required. This is easily seen as follows. With only user 1 in the system, time-sharing between different rates at different powers would decrease his overall rate since capacity is strictly concave in transmit power (Jensen's inequality). Rate splitting is not needed either. Thus user 1 does not use time-sharing when he is the only user in the system. Since user 1 is oblivious to the presence of new users, the BS cannot use time-sharing or split user 1's data into substreams and rearrange the encoding order of these substreams when new users appear. The same logic applies to all users.

Thus, no time-sharing or rate-splitting is required and the optimal DP vector is the one where users are encoded in the inverse order of their indices. □

To better illustrate the downlink strategy, we present a detailed example for a system with 3 users. The base station follows the following sequence of steps *in this order*.

(1) Determine the rate $R_1^\star$ and the input covariance matrix $Q_1^\star$ for user 1 according to equation (26). Note that these are simply the single-user capacity of user 1's channel and the waterfilling distribution that achieves that capacity when no other user is present.

(2) Determine the rate $R_2^\star$ and the input covariance matrix $Q_2^\star$ for user 2 according to equation (26). These are the single-user capacity and the waterfilling distribution that achieves that capacity for user 2's channel treating the interference from user 1 at the output of user 2's channel as colored Gaussian noise.

(3) Determine the rate $R_3^\star$ and the input covariance matrix $Q_3^\star$ for user 3 according to equation (26). These are the single-user capacity for user 3's channel and the waterfilling distribution that achieves that capacity treating the interference from users 1 and 2 as colored Gaussian noise.

(4) Encode user 3's data. That is, generate $\mathcal{C}_3^n$.

(5) Using the knowledge of the interference caused by $\mathcal{C}_3^n$ at the output of user 2's channel, encode user 2's data. That is, generate $\mathcal{C}_2^n$. Thus, user 3 presents known interference to user 2 and does not affect user 2's capacity.

(6) Using the knowledge of the interference caused by $\mathcal{C}_3^n + \mathcal{C}_3^n$ at the output of user 1's channel, encode user 1's data. That is, generate $\mathcal{C}_1^n$. Thus, users 2 and 3 present known interference to user 1 and do not affect user 1's capacity.

Note that in order to determine the users' optimal rates and input distributions, we need to proceed in the order $1, 2, \dots, K$. However, after that the actual codes are generated in the order $K, K-1, \dots, 1$.

The solution for the downlink is interesting for its simplicity and also for its striking symmetry with the uplink solution.

7. MULTIPLE BASE STATIONS

In this section, we incorporate multiple base stations to model a multicell environment. We assume that all the base stations are connected through a high-speed reliable network. It allows perfect coordination and information exchange between base stations. *Cooperation between base stations has also been considered previously for the uplink by Wyner in [24] and for the downlink by Shamai and Zaidel in [25].*

7.1. Uplink

On the uplink, the received signal at the bth base station is characterized by the following equation:

$$Y^{[b]} = \sum_{i=1}^{K} H_i^{[b]} X_i + N^{[b]}, \tag{27}$$

where $Y^{[b]}$ is the received vector at the bth base station, K is the number of users currently active in the system, $H_i^{[b]}$ is the flat-fading $B_b \times U_i$ matrix channel between user i and base station b, B_b and U_i are the numbers of antennas at the bth base station and the ith user, respectively, and N^b is the AWGN vector at the bth base station.

However, since we allow perfect coordination and information exchange between base stations, note that we can treat all the base stations together as one big base station with all the antennas. The equivalent description of the received signal at this base station is given by (1).

$$Y = \sum_{i=1}^{K} H_i X_i + N. \tag{28}$$

Here Y, H_i, and N are obtained by stacking up on top of each other the corresponding $Y^{[b]}$, $H_i^{[b]}$, and $N^{[b]}$ for all the base stations. But this brings us back to the single-cell model. Thus, for the uplink, the optimal solutions for the single cell simply carry through to the multicell environment.

7.2. Downlink

We extend the downlink solution to DP2b (existing users oblivious to the presence of new users) with multiple cells.

The downlink with B base stations is described as

$$Y_i = \sum_{b=1}^{B} H_i^{[b]} \sum_{j=1}^{K} X_j^{[b]} + N_i, \quad 1 \le i \le K,\ 1 \le b \le B, \tag{29}$$

where Y_i is the output vector, $X_i^{[b]}$ and $H_i[b]$ are the input vector and the channel matrix from base station b, and N_i is the AWGN vector for user i. Further, the additional power for each new user is limited per base station so that

$$\operatorname{trace}\left[E\left[X_i^{[b]} X_i^{[b]\dagger}\right]\right] \le P_i^{[b]}, \quad 1 \le i \le K,\ 1 \le b \le B. \tag{30}$$

Note that a system where each user is assigned to only one base station is included as a special case by setting the appropriate power constraints to zero.

Again, since we allow perfect coordination between base stations, we can represent the B base stations as one big base station. Defining

$$H_i = \left[H_i^{[1]} \ H_i^{[2]} \ \cdots \ H_i^{[B]}\right], \quad 1 \le i \le K, \tag{31}$$

and X_i as the vector obtained by stacking all the $X_i^{[b]}$ into a single column, we obtain an equivalent representation for the downlink as (3). Now this looks similar to the single-cell downlink model we had earlier. However, note that the components of the input vector X_i come from different base stations. There is a different input power constraint on each base station. Thus, the solution presented earlier does not apply in the exact same form. *However, a natural extension of the single-cell downlink solution to multiple base stations is obtained as follows.*

Although rate splitting is not necessary, recall that it does not reduce capacity. We explain the multicell extension of the single-cell downlink solution in terms of rate splitting for clarity. Specifically, we split each user's rate into B substreams. The idea is to perform the waterfill in B stages. At each stage, we waterfill until a base station meets its power constraint. Then we null out the antenna gains from that base station so that no more power is allocated to it and proceed with the waterfill. This gives us B layers or B substreams that can be encoded using DP. Consider the ith user. As shown in Theorem 3, this user sees the interference from users $1, 2, \ldots, i-1$ as colored noise and is unaffected by the interference from users $i+1, i+2, \ldots, K$. Therefore, the maximum rate he can achieve is given by

$$R_i^{\star} = \max_{Q_i} \log \left| I + \left(I + \sum_{j=1}^{i-1} H_j Q_j^{\star} H_j^{\dagger}\right)^{-1} H_i Q_i H_i^{\dagger} \right|, \tag{32}$$

where the maximization is over all input covariance matrices that satisfy the power constraints per base station. We split the user's rate into B substreams to be encoded in the order $B, B-1, \ldots, 1$ using DP encoding. So the Bth substream sees all the other substreams as noise, while the first substream's rate is unaffected by substreams $B, B-1, \ldots, 2$. Let the rates on these substreams be $R_i^{[b]\star}$, and the corresponding input covariance matrices be $Q_i^{[b]\star}$. Then we have

$$\begin{aligned} R_i^{\star} &= R_i^{[1]\star} + R_i^{[2]\star} + \cdots + R_i^{[B]\star}, \\ Q_i^{\star} &= Q_i^{[1]\star} + Q_i^{[2]\star} + \cdots + Q_i^{[B]\star}. \end{aligned} \tag{33}$$

The optimal $Q_i^{\star}$ is obtained as follows.

(1) Perform a singular value decomposition of the effective composite channel $(I + \sum_{j=1}^{i-1} H_j Q_j^{\star} H_j^{\dagger})^{-1/2} H_i$ as

$$\left(I + \sum_{j=1}^{i-1} H_j Q_j^{\star} H_j^{\dagger}\right)^{-1/2} H_i = F_i \Lambda_i M_i. \tag{34}$$

Start water-pouring over the eigenmodes of this channel. Continue adding power until one of the base stations meets its power constraint for the ith user $P_i^{[b]}$. Without loss of generality, we assume base station 1 runs out of power for user i. This corresponds to the first rate split, that is, call this the first substream for user i. The input covariance matrix obtained in this way is $Q_i^{[1]\star}$. Among the B substreams corresponding to user i, this substream will be encoded last, so it is unaffected by the interference from the remaining $B-1$ substreams. The rate on this substream is

$$R_i^{[1]\star} = \log \left| I + \left(I + \sum_{j=1}^{i-1} H_j Q_j^{\star} H_j^{\dagger}\right)^{-1} H_i Q_i^{[1]\star} H_i^{\dagger} \right|. \tag{35}$$

(2) Since base station 1 already used up its power for user i, we null out the contribution from $H_i^{[1]}$ to the compound channel matrix by setting it to zero. Define a new composite channel

$$H_i^{[-1]} = \left[\mathbf{0} \ H_i^{[2]} \ H_i^{[3]} \ \cdots \ H_i^{[B]}\right]. \tag{36}$$

Again, perform a singular value decomposition on the new composite effective channel

$$\begin{aligned} &\left(I + \sum_{j=1}^{i-1} H_j Q_j^{\star} H_j^{\dagger} + H_i Q_i^{[1]\star} H_i^{\dagger}\right)^{-1/2} H_i^{[-1]} \\ &\quad = F_i^{[-1]} \Lambda_i^{[-1]} M_i^{[-1]}. \end{aligned} \tag{37}$$

Note that this treats the interference from the first substream as noise. Again, start water-pouring over the eigenmodes of this new channel until another base station meets its power constraint. Without loss of generality, we call this base station 2. This gives us the input covariance matrix $Q_i^{[2]\star}$ on the second substream. The rate for the second substream is

$$\begin{aligned} R_i^{[2]\star} = \log \Bigg| I &+ \left(I + \sum_{j=1}^{i-1} H_j Q_j^{\star} H_j^{\dagger} + H_i Q_i^{[1]\star} H_i^{\dagger}\right)^{-1} \\ &\times H_i^{[-1]} Q_i^{[2]\star} H_i^{[-1]\dagger} \Bigg|. \end{aligned} \tag{38}$$

Proceeding in this fashion, we obtain the input covariance matrices on all the substreams and the corresponding rates as well. Combining the substreams, we get the overall rate and input covariance matrix for each user from equations (33).

Thus we find that multiple base stations only affect the downlink solution to the extent that the waterfilling algorithm needs some modification in order to accommodate the different power constraints per base station. Otherwise, the solution does not change. In particular, we still use DP coding, and the ordering of users is the same as before. Also note that while we used rate splitting to derive the optimal input covariance matrix, it is not necessary to split the rates into substreams. The same overall input covariance matrix can be used without rate-splitting to achieve the same capacity.

8. CONCLUSIONS AND DISCUSSION

We addressed the problem of providing best possible rates to new users as they enter a wireless data network, without penalizing the existing users. We have dubbed the network a PhantomNet. This is because of the design theme, that when a user enters, all subsequent entrants must, to him, be phantoms, that is, interference-wise, they must be invisible. For both the uplink and the downlink, only earlier entrants can interfere with an entering user. PhantomNet operation involves treating all bases as a single composite base, so that the actual bases simply serve as multiple antenna sites which are networked, say with fibers, to and from a single central processor.

For the uplink we found that, to achieve the phantom requirement, we could make a straightforward application of the well-established SD strategy where the new user is *decoded* before the existing users. For the downlink, achieving the invisibility requirement is more problematic. The optimal downlink strategy is to use DP coding, where the new user is *encoded* before the existing users. This makes use of the fact that the bases have knowledge of all signals that are to be transmitted. This enables simultaneous communication to the users despite arbitrarily varying interference by signalling modulo a fundamental lattice cell and using dithering techniques.

The striking feature of the uplink and the downlink strategies is their simplicity, and even more than that, their similarity. In both cases, the new users are forced to see the existing users as noise while the existing users are not affected by the presence of the users who joined the system after them. That is, they can continue to operate exactly as before. The only changes need to be made at the BS. For the uplink, the base station is the decoder and thus the solution hinges on the optimal decoding order, whereas for the downlink, the base station is the encoder and the solution is based on the encoding order. Note that as users leave the system, the same structure is maintained. As a user exits, it does not affect the rates of the users who joined the system before him. It does help the users who joined the system after him since they no longer have to face interference from his signal.

With multiple cells, we found that the uplink was effectively the same as a single-cell system since all the base stations are treated as one composite base station. Thus the single-cell strategy extends to multiple cells without loss of optimality. In contrast to the uplink, while the downlink is also viewed as a single virtual base station, there is a refinement since each of the actual base stations has a separate total power constraint. Consequently, the multiple cell downlink solution is different in that the distinct total transmit power constraints require a multistage waterfilling solution in determining the optimal input covariance matrix for each user. At each stage, waterfilling is performed until each base station meets its total power constraint. Those base stations that have already met their power constraints are not considered in the successive waterfilling stages.

While we drew heavily on published results, the novelty of our finding is the generality achieved in our setting: multiple base stations and multiple users with multiple antennas accommodated at both the transmit and receive sites. We also proved a general result that extends beyond our framework. We showed that the achievable rate region with SD and time-sharing is contained within the achievable rate region with DP coding and time-sharing.

We stress that PhantomNet uses information theoretic means for self-organizing the allocation of communication resources. There is allowance of extreme flexibility in allocating resources to a user. For example, which bases, which antennas at the bases, (which sectors) and which frequency bands are made available to a user need not be imposed over the network area. Instead, resource allocations can be left to develop, dynamically as needed (on the fly), in a fine-grained manner as expressed by the information theory formulas that we have presented. Dynamic choices would emerge as users come and go. Whenever and wherever and to what extent such amorphous allocations result in a superior network compared to imprinting a rigid regular structure from the outset is a topic for future study. Through constraints, one is free to impose structure when it looks advisable. A simulation testbed could be used to study PhantomNet operation to learn which beneficial features should first be moved into practice. Such a testbed could also be used to quantify the value of more antennas, more sectorization, and so forth.

ACKNOWLEDGMENT

It is a pleasure to thank Angel Lozano, Harish Viswanathan, Howard C. Huang, and Sriram Vishwanath for helpful discussions.

REFERENCES

[1] H. Viswanathan, S. Venkatesan, and H. Huang, "Downlink capacity evaluation of cellular networks with known interference cancellation," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 802–811, 2003.

[2] T. M. Cover, "Comments on broadcast channels," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2524–2530, 1998.

[3] G. Caire and S. Shamai, "On achievable rates in a multi-antenna Gaussian broadcast channel," in *Proc. IEEE International Symposium on Information Theory (ISIT '01)*, Washington, DC, USA, June 2001.

[4] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[5] S. Vishwanath, N. Jindal, and A. Goldsmith, "On the capacity of multiple input multiple output broadcast channels," in *Proc. IEEE International Conference on Communications (ICC '02)*, pp. 1444–1450, New York, NY, USA, April 2002.

[6] W. Yu and J. Cioffi, "Trellis precoding for the broadcast channel," in *Proc. IEEE Global Telecommunication Conference (Globecom '01)*, pp. 1338–1344, San Antonio, Tex, USA, November 2001.

[7] P. Viswanath and D. N. Tse, "Sum capacity of the multiple antenna Gaussian broadcast channel and uplink-downlink duality," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1912–1921, 2003.

[8] M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, 1983.

[9] S. Vishwanath, G. Kramer, S. Shamai, S. Jafar, and A. Goldsmith, "Capacity bounds for Gaussian vector broadcast channels," in *Proc. of DIMACS Workshop on Signal Processing for Wireless Transmission*, Rutgers University, Piscataway, NJ, USA, October 2002.

[10] P. Viswanath and D. Tse, "On the capacity of the multiple antenna broadcast channel," in *Proc. DIMACS Workshop on Signal Processing for Wireless Transmission*, Rutgers University, Piscataway, NJ, USA, October 2002.

[11] H. Viswanathan and K. Kumaran, "Rate scheduling in multiple antenna downlink wireless systems," Technical Memorandum 10009626-010720-01TM, Bell Laboratories, 2001.

[12] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple access channel," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 364–375, 1995.

[13] F. R. Farrokhi, G. J. Foschini, A. Lozano, and R. A. Valenzuela, "Link-optimal space-time processing with multiple transmit and receive antennas," *IEEE Communications Letters*, vol. 5, no. 3, pp. 85–87, 2001.

[14] W. Yu, W. Rhee, S. Boyd, and J. Cioffi, "Iterative water-filling for vector multiple access channels," in *Proc. IEEE International Symposium on Information Theory (ISIT '01)*, p. 332, Washington, DC, USA, June 2001.

[15] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice-strategies for cancelling known interference," in *Proc. International Symposium on Information Theory and Its Applications (ISITA '00)*, Honolulu, Hawaii, USA, November 2000.

[16] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Transactions on Information Theory*, vol. 42, no. 4, pp. 1152–1159, 1996.

[17] R. Urbanke and B. Rimoldi, "Lattice codes can achieve capacity on the AWGN channel," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 273–278, 1998.

[18] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[19] J. Kim and J. Cioffi, "Spatial multiuser access with antenna diversity using singular value decomposition," in *Proc. IEEE International Conference on Communications (ICC '00)*, pp. 1253–1257, New Orleans, La, USA, June 2000.

[20] S. Vishwanath, N. Jindal, and A. Goldsmith, "On the capacity of multiple input multiple output broadcast channels," in *Proc. IEEE International Conference on Communications (ICC '02)*, New York, NY, USA, April 2002.

[21] P. P. Bergmans, "A simple converse for braodcast channels with addtiive white Gaussian noise," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 279–280, 1974.

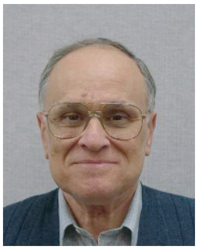
[22] S. Jafar and A. Goldsmith, "On the capacity of vector Gaussian MAC and BC with feedback," in *Proc. of 41st Annual Allerton Conference on Communication, Control and Computing*, Monticello, Ill, USA, October 2003.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 1991.

[24] A. D. Wyner, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1713–1727, 1994.

[25] S. Shamai and B. Zaidel, "Enhancing the cellular downlink capacity via co-processing at the transmitting end," in *Proc. of 53rd Vehicular Technology Conference*, vol. 3, pp. 1745–1749, Rhodes, Greece, Spring 2001.

Syed A. Jafar received his B. Tech. degree in electrical engineering from the Indian Institute of Technology (IIT), Delhi, India, in 1997, his M.S. degree in electrical engineering from California Institute of Technology (Caltech), Pasadena, USA, in 1999, and his Ph.D. degree in electrical engineering from Stanford University in 2003. He was a Senior Engineer at Qualcomm Inc., San Diego, in 2003. He did a summer internship in the Wireless Communications Group, Lucent Bell Laboratories, Holmdel, NJ, in 2001, and has two pending patents resulting from that work. He was an Engineer in the satellite networks division, Hughes Software Systems, India, from 1997 to 1998. In January 2004, he joined the Department of Electrical Engineering and Computer Science at the University of California, Irvine, where he is currently an Assistant Professor. His research interests include spread-spectrum systems, multiple antenna systems, and multiuser information theory.



Gerard J. Foschini received his B.S.E.E. degree from New Jersey Institute of Technology (NJTT) in 1961, his M.E.E. degree from New York University (NYU) in 1963, and his Ph.D. degree in mathematics from Stevens Institute of Technology in 1967. Mr. Foschini has been at Bell Laboratories for 41 years. He is a Bell Labs Fellow and he holds the position of Distinguished Member of Staff and Distinguished Inventor. He has conducted data communications research on many kinds of systems, most recently wireless communications and optical communications systems. Gerard has done extensive research on point-to-point systems as well as on networks. He has authored one hundred papers and holds seven patents. He has won the IEEE Eric E. Sumner Award, the Thomas Alva Edison Patent Award, and the Bell Labs President's Gold Award. He has taught at both Princeton and Rutgers Universities.

Andrea J. Goldsmith received her B.S., M.S., and Ph.D. degrees in electrical engineering from U.C. Berkeley in 1986, 1991, and 1994, respectively. She was an Assistant Professor in the Department of Electrical Engineering at Caltech from 1994 to 1999. In 1999, she joined the Electrical Engineering Department at Stanford University, where she is currently an Associate Professor. Her industry experience includes affiliation with Maxim Technologies from 1986 to 1990, where she worked on packet radio and satellite communication systems, and with AT&T Bell Laboratories from 1991 to 1992, where she worked on microcell modeling and channel estimation. Her research includes works in capacity of wireless channels and networks, wireless information and communication theory, multiantenna systems, joint source and channel coding, cross-layer wireless network design, communications for distributed control, and adaptive resource allocation for cellular systems and ad-hoc wireless networks. Dr. Goldsmith is a Terman Faculty Fellow at Stanford and a recipient of the Alfred P. Sloan Fellowship, the National Academy of Engineering Gilbreth Lectureship, a National Science Foundation Career Development Award, the Office of Naval Research Young Investigator Award, a National Semiconductor Faculty Development Award, an Okawa Foundation Award, and the David Griep Memorial Prize from U.C. Berkeley. She was an Editor for the IEEE Transactions on Communications from 1995 to 2002, and has been an Editor for the IEEE Wireless Communications Magazine since 1995. She is also an elected member of Stanford's faculty senate and the board of governors for the IEEE Information Theory Society.