Rate Scheduling in Multiple Antenna Downlink Wireless Systems

Harish Viswanathan and Krishnan Kumaran
Lucent Technologies Bell Laboratories
Murray Hill
NJ 07974
{harishv, kumaran}@lucent.com

Abstract

We consider scheduling strategies for multi-antenna and multi-beam cellular wireless systems for high speed packet data services on the downlink. We establish a fundamental connection between the stability region of the queuing system and the set of feasible transmission rates which provides the basis for the scheduling algorithm proposed in this paper. Transmission using adaptive steerable beams and fixed sector beams are considered and average delay versus throughput results are obtained through simulations for the proposed scheduling algorithm in each case. While in single antenna systems multi-user diversity gains are achieved by the scheduling algorithms that transmit to a single user in each scheduling interval, our results show that with multiple antennas, transmitting to a carefully chosen subset of users has superior performance. The multi-antenna scheduling problem is closely related to the problem of coordinated scheduling for transmission through multiple base stations, where a user can receive signals from several base stations simultaneously. We consider the special case when three single-antenna base stations are allowed to cooperate and transmit to the users in the triangular region between the base stations and propose scheduling strategies that demonstrate significant gains.

Key Words: scheduling, packet data, multiple antenna, downlink, interference avoidance

1 Introduction

High speed packet data services are expected to be one of the main applications of future wireless communication systems. One of the main characteristics that make packet scheduling over mobile wireless networks distinct from scheduling over wireline networks is the fact that the wireless channel is time-varying due to multipath fading. In delay tolerant packet data systems it is thus possible, with the aid of channel condition feedback from the users, to schedule transmission to users when their fading conditions are favorable thereby achieving multi-user diversity [1],[6], [11]. Most of these scheduling algorithms that try to achieve multi-user diversity gains are based on scheduling rules that are designed to transmit to a single "best" user during each scheduling interval. Under the assumption that only a single antenna is available at the base station it can be shown that transmission to a single user in each scheduling interval is a good strategy [1], [11]. However, with multiple transmit antennas at the base station it is not clear that transmission to a single user is optimum. We study scheduling strategies for wireless systems in which the base station is equipped with multiple antennas or multiple beams where transmission to more than one user in each scheduling interval can be superior to transmitting to only a single user. We present algorithms for choosing a good subset of users to transmit to in each scheduling interval and study their delay-throughput performance.

The performance of any scheduling algorithm critically depends on the transmission rates achieved in each scheduling interval which in turn depend on the coding, modulation and beamforming techniques employed. Under reasonable assumptions that the wireless channel is static during the entire duration of the scheduling interval and that there are sufficient number of symbol durations (high bandwidth communication) in each scheduling interval, the feasible set of rate vectors or the *rate region* that is achievable is well defined. We consider scheduling strategies for rates achievable through adaptive beam steering in the case when the vector channel is available at the base station and also fixed beamforming for the case when complete channel knowledge is not available at the transmitter. Most of the literature on multiple antenna signal processing is focused on the physical layer techniques and enhancements. In this paper we focus on using the well known beamforming techniques to schedule transmissions to different users for delay tolerant packet data applications. Starting from a key result in [7] we establish a tight connection between *stability* of the multi-user queuing system and the achievable rate region in Proposition 1 that forms the basis for the

scheduling algorithms we propose. We then apply this result to the rate region achievable through beamforming techniques.

Note that for the transmission strategies based on beamforming do not jointly encode the information of the different users transmitted to simultaneously in each scheduling interval. On the other hand, the optimum transmission strategy may involve joint coding and might lead to a larger set of feasible rate vectors. Downlink transmission from a single base station to many users falls under the class of broadcast channels in multi-user information theory. Since the multi-antenna broadcast channel does not fall into the category of degraded broadcast channels, determining the rate region is known to be a difficult problem [3], [4]. Nevertheless, we show the optimality of the simple beamforming technique with separate coding under some special conditions which then motivates the proposed scheduling strategy.

The multi-antenna scheduling problem is closely related to the problem of coordinated scheduling through multiple base stations, where a user can receive signals from several base stations simultaneously. We consider the special case of coordinated scheduling among three base stations, each with a single transmit antenna transmitting to users in a triangular region between the base stations, using the same underlying scheduling principle developed for the multi-antenna single cell scheduling problem. We demonstrate significant gains from coordinated scheduling through a simulation study of the proposed scheduling algorithms.

The paper is organized as follows. The multi-antenna downlink queuing problem is described in Section 2. The connection between the rate region and stability of the queuing system is established in Proposition 1 in Section 3. In Section 4 we present some results that indicate that in the case of a system with large number of users, the interference avoidance based beamforming strategy will be optimum. The rate scheduling algorithms are described in Section 5 and the delay-throughput performance simulation results are presented in Section 6. In Section 7 we treat the multi-cell scheduling problem. We conclude with a summary of results in Section 8.

2 Problem Statement

Consider the scenario where a base station of a cellular system is equipped with multiple transmit antennas and is serving a set of K users with single receive antennas. Packets arrive at the base

station for all the users in that cell according to some stationary arrival process at some specified average arrival rate λ. Packets are buffered in the base station until the scheduling rule assigns resources for transmission. We seek to design efficient scheduling strategies that respond to queue occupancies while effectively exploiting the delay tolerance of data and channel variations of the wireless medium.

For the purposes of our comparative study of scheduling algorithms, we make several simplifying assumptions, but we expect our conclusions to hold qualitatively even without these. Firstly, the buffer is assumed to be arbitrarily large to accommodate any number of packets, and the main performance measure we use is average delay, which is related to average buffer content. We also assume that the buffers can be served at arbitrary rates, while in practice the rates would come from a discrete set. For these simplifications to be useful, we must require that the data arrives as large numbers of small, finely granularized units, i.e. packets that are small compared to the transmission capacity in a single scheduling interval. We thus consider a fluid flow model of the queuing system in which the arrival and departures are viewed as continuous, infinitesimally divisible, fluid. Perhaps our most stringent assumption, from a wireless system point of view, is perfect knowledge of the channel condition of every user at each scheduling interval. In practice, this can only be approximated through feedback, and in some cases, can be partially inferred from knowledge of *uplink* (mobile-to-base) channel characteristics. Further, the transmission rates depend on the choice of the coding strategy and the power allocation for the various users, and we adopt specific models for these. Given these simplifications, the goal of the scheduling algorithm is to minimize the average delay across all users for any given arrival rate. In our simulations, we assume that the arrival rates are the same for all users, but our approach may be extended to inhomogeneous user populations.

We now proceed to describe the channel model we use in our study. The received sampled signal at the the k^{th} user's receiver for the n^{th} symbol period y_k^n is given by

$$y_k^n = \mathbf{h}_k^\dagger \mathbf{x}^n + v_k^n \tag{1}$$

where $\mathbf{x}^n$ is $M-$dimensional complex vector corresponding to the n^{th} transmitted symbol, $\mathbf{h}_k$ is the $M-$dimensional vector channel between the M antennas at the base station and user k, and v_k^n represents the additive complex white Gaussian noise and the interference signals from neighboring base stations (assumed to be Gaussian) at the k^{th} user with total variance N_k. The received signal

at all K receivers in matrix-vector notation is denoted by

$$\mathbf{y}^n = \mathbf{H}^\dagger \mathbf{x}^n + \mathbf{v}^n \tag{2}$$

where $\mathbf{y}^n = (y_1^n, \cdots, y_K^n)^t$, $\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_K]$ and $\mathbf{v}^n = (v_1^n, \cdots, v_K^n)^t$. Throughout this paper we assume that the channel matrix $\mathbf{H}$ is available at the base station. We assume that the channel remains static for the duration of the scheduling interval and is independent for different scheduling intervals as is typically assumed in block fading channel models. In practice, the time-variations depend on the vehicle speed. Furthermore, we assume that the number of symbols in each scheduling interval is sufficiently large so that we can assume that practical coding schemes can achieve data rates close to the Shannon capacity.

Based on the channel knowledge and the size of the queues of all the users at the start of the scheduling interval, the scheduling rule picks a subset of users to transmit to, and the power allocation across the users for the given transmission technique (coding and beamforming). A fixed total power level P is available at the base station to be shared among all the users. When transmitting to multiple users, at any given user the interference from signals intended for the other users is treated as additive white Gaussian noise for simplicity. Similarly, the signal from all other base stations are treated as additive Gaussian noise as well. Furthermore, in any scheduling interval if the signal-to-interference-and-noise ratio (SINR) at any user is γ then a rate of $R = \log(1+\gamma)$ bits/symbol corresponding to the AWGN channel Shannon capacity is assumed to be achievable whenever the information for that user is encoded separately.

3 Rate Region and Stability

Consider the problem of optimal scheduling of resources among a set of K users when packet arrivals can be queued at the server. A very general and useful scheduling result regarding the maximum arrival rates at which all the users' queues are *stable* as a function of the set of feasible rate vectors $\mathcal{R}$ for simultaneous transmission to all users was presented in [7]. Furthermore, an algorithm for choosing the rate vectors over time that guarantees the stability of the queues for that maximum arrival rate was also presented in [7]. In that work, it was assumed that the feasible set of rate vectors $\mathcal{R}$ was deterministic and time-invariant. However, for the multi-antenna scheduling for mobile cellular wireless under consideration the rate region is time-varying and is governed by

the channel conditions. Hence we consider the extension of the result in [7] to the case when the feasible set of rate vectors or the rate region is governed by an underlying random process $S(n)$ referred to as the state, i.e., the feasible rate region during the scheduling period n is given by $\mathcal{R}^n = \mathcal{R}(S(n))$. The following proposition gives the equivalent of the optimal scheduling result in [7] to the stochastic feasible rate region case.

In what follows the set of rate vectors $\mathbf{R} \in \mathcal{R}(s)$ is indexed by a parameter i from some index set which, with some abuse of notation, is also denoted by $\mathcal{R}(s)$.

Proposition 1 *Given a feasible set of rate vectors for simultaneous transmission to multiple users during each scheduling interval, indexed by $n \in \mathcal{Z}_+$, referred to as the* rate region $\mathcal{R}^n = \mathcal{R}(S(n))$ *where $S(n)$ is the underlying ergodic* state *process governing the feasible rate region, the vector of arrival rates $\mathbf{\Lambda}$ is dominated by*

$$\mathbf{\Lambda} \leq \sum_{s \in \mathcal{S}} \pi(s) \sum_{i \in \mathcal{R}(s)} \phi_{s,i} \mathbf{R}_i(s) \tag{3}$$

where $\pi(s)$ is the steady state probability of the ergodic state process $S(n)$ and $\phi_{s,i}$ is some stochastic matrix satisfying $\sum_i \phi_{s,i} = 1 \ \ \forall s \in \mathcal{S}$. Furthermore, choosing the rate vector $\mathbf{R} \in \mathcal{R}^n$ that solves the following optimization problem guarantees stability *in the queuing sense (keeps all user queue lengths bounded when feasible) for all vector of arrival rates strictly dominated by $\mathbf{\Lambda}$.*

$$\max_{\mathbf{R} \in \mathcal{R}^n} \mathbf{Q}(n) \cdot \mathbf{R} \tag{4}$$

Here $\mathbf{Q}(n)$ is the vector of queue lengths for the users at the start of the scheduling interval n.

Proof: By definition, if $\mathbf{\Lambda}$ is a vector of arrival rates for which the queue is stable, then there exists a scheduling rule that picks the feasible rate vector $\mathbf{R}(n) = \mathbf{R}_i(s) \in \mathcal{R}(S(n))$ such that $\mathbf{\Lambda} \leq E[\mathbf{R}(n)]$. Now set

$$\phi_{s,i} = E\left[I(\mathbf{R}(n) = \mathbf{R}_i(s)) | S(n) = s\right],$$

where $I(\cdot)$ is the indicator function on the set of indices $\mathcal{Z}_+$. It follows that

$$\begin{aligned}
\mathbf{\Lambda} &\leq E[\mathbf{R}(n)] \\
&= E\left[E\left[\mathbf{R}(n) | S(n) = s\right]\right] \\
&= \sum_s \pi(s) E\left[\sum_i \mathbf{R}_i(s) I(\mathbf{R}(n) = \mathbf{R}(s,i)) | S(n) = s\right] \\
&= \sum_s \pi(s) \sum_i \phi_{s,i} \mathbf{R}_i(s).
\end{aligned}$$

Thus, inequality (3) is proved.

To prove the second part of the Proposition we follow the approach in Theorem 6.1 in [7]. Denote by $\mathbf{R}(n)$ the sequence of rate vectors obtained as the solution to the optimization (4) and by $\lambda(n)$ the arrival at time n. Then, since the optimal value in (4) is larger than the value obtained using any linear combination of achievable rate vectors we have,

$$\mathbf{Q}(n) \cdot \mathbf{R}(n) \geq \sum_i \phi_{S(n),i} \mathbf{Q}(n) \cdot \mathbf{R}_i(S(n)) \tag{5}$$

Hence

$$\begin{aligned}
\|\mathbf{Q}(n+1)\|^2 &= \|\mathbf{Q}(n) + \lambda(n) - \mathbf{R}(n)\|^2 \\
&= \|\mathbf{Q}(n)\|^2 + \|\lambda(n)\|^2 + \|\mathbf{R}(n)\|^2 + 2\mathbf{Q}(n) \cdot \lambda(n) - \\
&\quad 2\,\mathbf{Q}(n) \cdot \mathbf{R}(n) - 2\,\mathbf{R}(n) \cdot \lambda(n) \\
&\leq \|\mathbf{Q}(n)\|^2 + \|\lambda(n)\|^2 + \|\mathbf{R}(n)\|^2 + 2\,\mathbf{Q}(n) \cdot \left(\lambda(n) - \sum_i \phi_{S(n),i} \mathbf{R}_i(S(n))\right)
\end{aligned}$$

where the last inequality follows from (5). Taking expectations under the assumption that future arrivals are independent of the current queue size we have

$$\begin{aligned}
E[\|\mathbf{Q}(n+1)\|^2] &\leq E[\|\mathbf{Q}(n)\|^2] + E[\|\lambda(n)\|^2] + E[\|\mathbf{R}(n)\|^2] + && (6) \\
&\quad 2E[\mathbf{Q}(n)] \cdot \left(\mathbf{\Lambda} - \sum_i \sum_s \pi(s) \phi_{s,i} \mathbf{R}_i(s)\right) && (7) \\
&\leq E[\|\mathbf{Q}(n)\|^2] + 2E[\mathbf{Q}(n)] \cdot \left(\mathbf{\Lambda} - \sum_i \sum_s \pi(s) \phi_{s,i} \mathbf{R}_i(s)\right) + \lambda + r && (8)
\end{aligned}$$

where we have set $r \stackrel{\triangle}{=} \max_{\mathbf{R} \in \mathcal{R}} \|\mathbf{R}\|^2$ (assuming that the feasible rate vectors are bounded) and $\lambda \stackrel{\triangle}{=} E[\|\lambda(n)\|^2]$ for a stationary arrival process. Hence, because of (3) whenever the queue size becomes sufficiently large there is a strictly negative drift that makes the queue smaller implying that the queue size remains bounded. Thus stability is established.

The objective function in (4) was chosen to establish the fact that the queues will remain stable. It should be noted that there are other objective functions that are functions of $Q(n)$ that can also guarantee stability. The different objective functions will have different throughput-delay characteristics. For example, in the "exp" rule [8] the objective function is chosen to be $\sum_{k=1}^{K} \exp(a_k Q_k(n)) R_k)$ for some constants $a_k > 0$ and was shown to achieve stability in the single antenna system in [9]. It is clear that this rule tries to equalize the queue sizes for the different users

rapidly and hence is useful for delay constrained applications. Proposition 1 gives the basis for all the scheduling algorithms proposed in this paper. The feasible set of rate vectors depends on the choice of coding, modulation and beamforming techniques employed. Note that the optimization also gives the solution to the optimal power allocation across the users since the maximization in (4) over all feasible rate vectors implies all possible power allocation across the users have to be considered.

4 Optimal Scheduling in a Large System

It is well known that for broadcast channels that do not fall into the category of degraded-broadcast channels, characterizing the rate region in a computable form is in general a hard problem [4]. The multi-antenna broadcast channel is not a degraded-broadcast channel and hence determining the information-theoretic optimal rate region is still an open problem [3] for the general case of M transmit antennas and K users. To motivate the scheduling strategy that relies on separate coding and beamforming that we propose in the next section, we demonstrate the optimality of the rate region achievable using this technique in some special cases.

Consider the special case when there are exactly M users with identical signal-to-noise ratios, i.e., $\|\mathbf{h}_k\|^2/N_k = c$ is the same for all k. Furthermore, assume that the users are all *mutually orthogonal* in the sense that $\mathbf{h}_i \cdot \mathbf{h}_j = 0$ for all $1 \leq i \neq j \leq M$. For this special case we have the following result.

Proposition 2 *The boundary of the rate region for the mutually orthogonal users with identical SNRs $c = \|\mathbf{h}_k\|^2/N_k$ is given*

$$R_k \;=\; \log\left(1 + cP_k\right), \quad 1 \leq k \leq K$$

with $\sum_{k=1}^{K} P_k = P$. Thus, given any set of users with equal SNRs, the total rate is maximized by transmitting to an orthogonal subset of users.

Proof: Since the $M \times M$ matrix $\mathbf{H}$ with k^{th} column equal to $\mathbf{h}_k/\sqrt{N_k}$, is a scaled version of a unitary matrix, the capacity region is equal to that of a system with the k^{th} user's channel given by $\tilde{\mathbf{h}}_k = \sqrt{c}[0, \cdots, 1, \cdots, 0]^t$ [3]. Hence it follows from Proposition 4 in the appendix that the rate

region is bounded by

$$R_k \quad \leq \quad \log\left(1 + cP_k\right)$$

with $\sum_{k=1}^{K} P_k \leq P$. It is easy to see that the above rate region is also achievable through independent Gaussian Codebooks for each user and "beamforming weights" matched to the channel.

The above result can be combined with the following result on rate scheduling to study the performance of the interference avoidance scheduling algorithm described in the next section where simultaneous transmissions are scheduled in any scheduling interval only to nearly orthogonal users. In particular, we argue below that for a system of users with identical SNRs the interference avoidance approach is optimal in the limit of large number of users K. The following result is an adaptation from [1], and generalizes the SSS (*static service split*) rule described therein.

Proposition 3 *For every scheduling interval indexed by t, let $\mathcal{R}_t$ denote the* rate region*, i.e. the set of feasible rate vectors $\mathbf{R}_t$. Then, there exist* static *user weights γ_i such that one or more rate vectors given by*

$$\mathbf{R}_t^* \stackrel{\triangle}{=} \arg\max_{\mathbf{R}\in\mathcal{R}_t} \sum_i \gamma_i R_i \tag{9}$$

achieve stability, and hence maximum throughput. When more than one, an appropriate form of time-sharing among the corresponding rate vectors within the time-slot is required.

From the isotropy of the fading and the equal long term rate requirement, it is easy to see that the static user weights γ_i will depend only on SNRs of the users. Since the SNRs are identical for all users, the rate vectors required in each scheduling interval to maximize long-term throughput will be such that $\sum_i R_i$ is maximized. It follows from Proposition 2 that whenever orthogonal users exist, the rate sum is maximized by transmitting to the orthogonal subset of users. In the more general case when the instantaneous SNRs are not identical, and each user undergoes fading independently with the same distribution, it is reasonable to expect in the limit of large number of users in the system that the rate vectors chosen during each scheduling interval will correspond to subsets of users that achieve the highest instantaneous SNRs, and among such users transmission to an orthogonal subset will be optimum.

5 Scheduling Algorithms

The first algorithm we consider is based on an interference avoidance strategy in which users are coded independently and the signals are then transmitted with simple maximum SNR beamforming [12] with the subset of users transmitted to in each scheduling interval being chosen judiciously to avoid mutual interference. This technique is described in Sub-section 5.1. We then consider a scheme based on a different antenna technology in which fixed, nearly non-overlapping sector beams are designed a-priori and simultaneous transmission to multiple users only in non-overlapping region of the beams is allowed. This technique is described in detail in Sub-section 5.3. The scheduling rule for all three algorithms is based on the rate scheduling result of Proposition 1 in Section 3. These algorithms differ in the achievable rate regions used in (4). We present results comparing the three different algorithms in the numerical results section.

5.1 Scheduling with Maximum SNR Beamforming

Each user's transmitted signal is obtained through a standard Gaussian codebook corresponding to some known signal-to-noise ratio and then transmitted over the multiple antennas using an appropriately chosen beamforming vector. This strategy is motivated by Propositions 2 and 3 in the previous section where we showed that when a large number of users are present in the system and it becomes possible to find orthogonal subsets of users near the highest instantaneous SNR, then it is optimal to transmit to such orthogonal subsets. When users are neither completely orthogonal nor identical, we expect that joint rate regions of subsets of "nearly" orthogonal users are convex, while those of subsets "highly" non-orthogonal users are concave. With this motivation in mind, and observing the simple form of the solution in the completely orthogonal users case (Proposition 2) we seek to formulate a simple mathematical problem that exploits the convexity of the rate-region for near-orthogonal user sets.

We apply Proposition 1 to our multiple transmit, single receive antenna scheduling problem. We first conjecture that, given the queue lengths of individual users, the optimizing rate vector would consist of non-zero entries only for users that make up a near-orthogonal set. This conjecture, motivated by arguments presented earlier, allows us to rewrite the optimization problem (4) as

$$\max_{\mathbf{R}:R_i,R_j>0 \text{ iff } \left|\frac{\mathbf{h_i}^\dagger \mathbf{h}_j}{|\mathbf{h_i}|\,|\mathbf{h_j}|}\right| \leq \epsilon} \mathbf{Q}(n)\cdot\mathbf{R} \tag{10}$$

where ϵ is a parameter much smaller than unity and $|\mathbf{h_i}| \stackrel{\triangle}{=} \sqrt{\mathbf{h_i}^\dagger \mathbf{h_i}}$. Hereafter, we focus on a specific scheduling interval and drop the interval argument n. We approximate "near-orthogonality" as practically perfect orthogonality for small ϵ. Under these assumptions, it is clear that $\mathbf{w_i} = c_i \mathbf{h}_i$, i.e., the beamforming vectors mimic the gain vectors up to constants of proportionality to be determined. Given the subset of users for which $R_i > 0$, the rate region can be derived in a straightforward manner using $R_i = \log(1 + \frac{c_i^2 |\mathbf{h}_i^\dagger \mathbf{h}_j|^2}{N_i})$ in (8) with $\sum_j c_j^2 |\mathbf{h}_j^\dagger \mathbf{h}_j| \leq P$. It remains to determine the optimum subset for which $R_i > 0$.

Define binary variables $x_i \in \{0,1\}$, with the value 1 (resp. 0) indicating whether $R_i > 0$ (resp $R_i = 0$). Note that, given an orthogonal set of users, and defining

$$\alpha_i \stackrel{\triangle}{=} \frac{N_i}{P\mathbf{h_j}^\dagger \mathbf{h_j}},$$

the rate vector that maximizes $\mathbf{Q}\cdot\mathbf{R}$ for given $\mathbf{Q}$ is of the form

$$R_i = \log\left(\frac{\theta Q_i}{\alpha_i}\right)$$

for some choice of the Lagrange multiplier θ. Further, define the pre-computed binary parameters

$$e_{ij} = I\left[\left|\frac{\mathbf{h_i}^\dagger \mathbf{h}_j}{|\mathbf{h_i}|\,|\mathbf{h_j}|}\right| > \epsilon\right] \tag{11}$$

that specify whether or not users i and j can both belong to the transmitting set. In terms of these, the optimization problem (10) reduces to

$$\max_{\theta,\mathbf{x}} \sum_i x_i Q_i \log\left(\frac{\theta Q_i}{\alpha_i}\right) \tag{12}$$

subject to

$$x_i + x_j \leq 2 - e_{ij} \quad \forall\, i \neq j \tag{13}$$

$$\sum_i x_i(\theta Q_i - \alpha_i) \leq 1. \tag{14}$$

It is further easy to see that the total power constraint, is necessarily binding at optimum since otherwise θ can be increased to increase the objective. The solution to (10)-(12) can be obtained by enumeration for small number of users, but would require heuristics like LP-relaxation and rounding for larger number of users.

The choice of ϵ in equation (11) depends on the peak SNR and the maximum number of simultaneous users that is allowed. It can be chosen so that the interference seen from all other users is below the noise floor of the receiver. Note that when $\mathbf{w}_i = \frac{\mathbf{h}_i}{|\mathbf{h}_i|}$ as specified by the interference avoidance algorithm, the total interference is upper bounded by $\sum_{j \neq i} \frac{|\mathbf{h}_i^\dagger \mathbf{h}_j|^2}{|\mathbf{h}_j|^2} P$ where the summation is over all users transmitted to simultaneously in the same scheduling interval. In order for the interference to be small compared to the noise, it is thus sufficient to choose ϵ small compared to $\sqrt{\frac{1}{M \mathrm{SNR}_{\max}}}$ where $\mathrm{SNR}_{\max}$ is the maximum SNR achieved by any user in the system and a maximum of M simultaneous users is possible with M antennas.

5.2 Interference Avoidance under Limited Feedback

It is possible to essentially use the algorithm described above in the case when only the channel gains from the different transmit antennas to the user is fed back to the base station. In this case, the transmit weights $\mathbf{w}_i$ for the i^{th} user when it is selected for transmission, are chosen such that all but one of the antenna weights is zero corresponding to transmitting the i^{th} user's signal from a single antenna. The $e_{i,j}$ in this case would be based on "orthogonality" of channel gains:

$$e_{ij} = I\left[\frac{\sum_{m=1}^{M} |h_i(m)||h_j(m)|}{\|\mathbf{h}_i\|\|\mathbf{h}_j\|} > \epsilon\right].$$

Thus the phase information is not used at the transmitter to determine the optimal schedule.

5.3 Scheduling with Fixed Sector Beams

In low angle spread environments such as macro-cells, it is possible to obtain the angle of arrival information from the uplink signal. It is then possible to achieve interference avoidance using a fixed beam forming network and using the angle of arrival information to pick subset of users that are nearly orthogonal. Note that the beams can be overlapping since users in the overlap region can be scheduled one user per scheduling interval.

The scheduling algorithm is as in Section 5.1 except that the e_{ij} will now be based on whether users i, j belong to the overlapping region of two beams. Denote the subset of fixed beams that illuminate user i by $\mathcal{B}_i$ and define the pre-computed binary parameters that depend only on the user location

$$e_{ij} = \begin{cases} 1 & \text{if } \mathcal{B}_i \cap \mathcal{B}_j \neq \phi \\ 0 & \text{otherwise} \end{cases}$$

that specify whether or not users i and j can both belong to the transmitting set. We then solve a problem similar to (10)-(12) to choose the user sets for transmission. Assuming that the beam shapes are ideal in that they have a constant gain within the beam and complete suppression outside the beam-width, the α_i are defined according to $\alpha_i = \frac{N_i}{|h_i|^2 G}$ where G is the pre-determined beamforming gain that depends on the beam width and h_i is the scalar multipath fading gain between the base and user i.

6 Simulation Results

We studied the performance of the different scheduling algorithms through detailed system simulations. We considered a hexagonal cellular architecture with the central cell surrounded by 18 cells corresponding to two rings of interfering base stations. Each cell is divided into three 120 degree sectors with ideal sector beams that have zero side-lobes outside the sector. Users are distributed uniformly in one sector of the central cell. All other cell/sectors are assumed to be fully loaded and hence all interfering base stations are assumed to be transmitting at full power. Log normal shadow fading with standard deviation of 8 dB is simulated for each user from each of the base stations. Flat fading from each antenna to each user is simulated under the block fading channel model where the channel is constant during each scheduling interval and is independent across scheduling intervals. The total noise power N_k at each receiver is given by the sum of the receiver AWGN power and the interference signal power received from all other base stations except the base station transmitting the information to the receiver.

Fluid flow model of queuing as described in Section 2 is simulated. At the start of each scheduling interval the number of bits entering the queue for each user is determined by generating an exponential random variable with mean corresponding to the arrival rate. The number of bits of

information that can be transmitted to each user in a given scheduling interval is determined based on the scheduling algorithm being simulated and the queues are emptied accordingly. To avoid issues with frame-fill efficiency at low loads, for which individual queues can be empty, we implemented time sharing within each scheduling interval when necessary. As mentioned in Section 2 the channel is assumed to be constant in each scheduling interval and independent across scheduling intervals and users. For each arrival rate, the average queue size for all the users is calculated over the 40000 scheduling intervals simulated and then the average delay is obtained using Little's formula [2]. The simulation is repeated for 100 different user locations in the central cell sector and the average delay results are further averaged over user locations as well.

Figure 1 shows the performance for the single-user algorithm where in each slot the base station transmits to only one user. The user to transmit to is obtained as a solution to the optimization problem in (12) where ϵ is set equal to 0 so that at most one user can transmit in a given slot. The optimum beamforming weights in this case is indeed given by (11). It is clear from the figure that there is significant gain in going from 2 to 4 and then to 8 transmit antennas.

Figure 2 shows the performance comparison between transmitting to multiple users using the algorithms in Section 5 with $\epsilon = 0.1$ compared to transmitting to only one user using the *maximum weight* algorithm of Proposition 1 for 4 and 8 transmit antennas. In the figure, the legend "max weight-adaptive beams" is the scheduling strategy of Sub-section 5.1 while the legend "max weight - 4 45^o beams" is the scheduling strategy in Sub-section 5.3 with four fixed sector beams each 45 degree wide resulting in a 20 degree overlap between beams. The fixed beams are assumed to have a constant gain of 6 dB inside the beam and a -20 dB sidelobe outside the beam. Round robin and proportional fair algorithms [9] with beamforming to a single user in each scheduling interval are also included for comparison. The superior performance of the fixed beams is to be expected because the fixed beams are assumed to have very small sidelobes by using large antennas and thus allows for multiple users to be scheduled simultaneously in the same scheduling interval more often than for the adaptive beam case. The figure also shows that with 8 transmit antennas the simple interference avoidance based maximum SNR adaptive beams algorithm of Sub-section 5.1 has significant gain over transmitting to only one user.

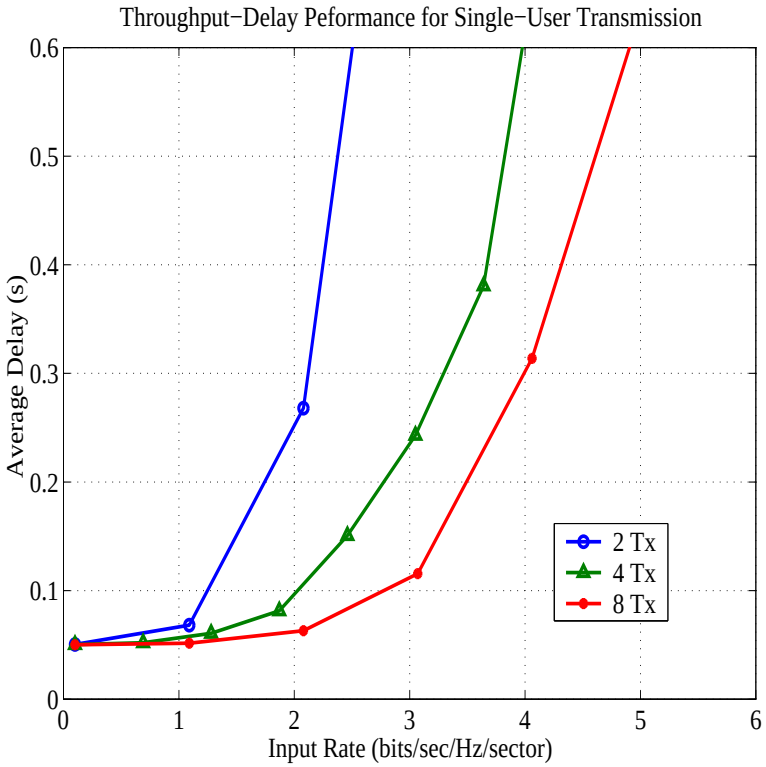

Figure 1: Scheduling Performance for different number of antennas

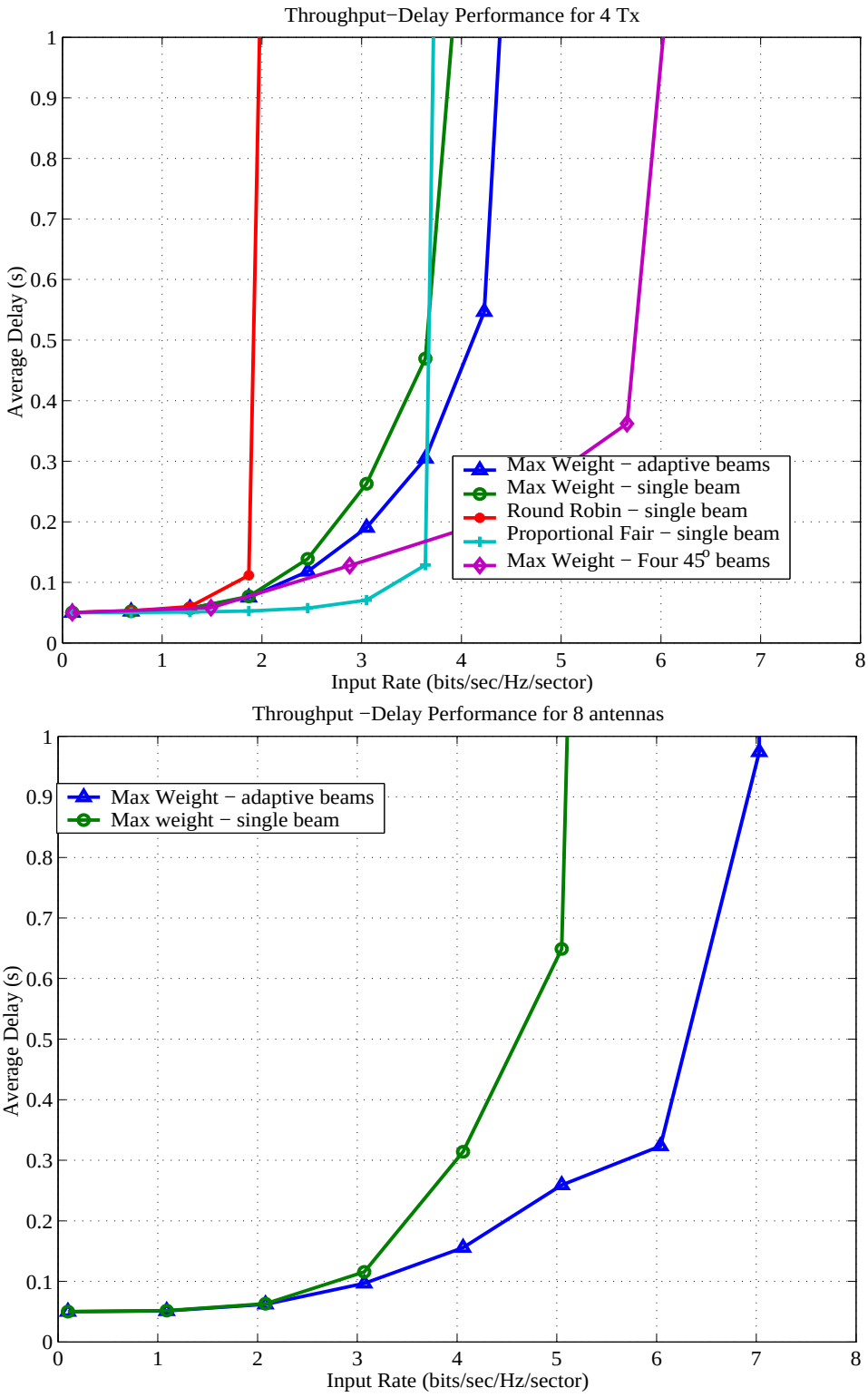

Figure 2: Scheduling performance for different scheduling strategies

7 Multi-Base Problem

In a cellular system with universal frequency reuse as in second and third generation cellular CDMA systems, users at the edges of cells suffer from significant interference from signals transmitted from base stations in neighboring cells. This issue is addressed in spread spectrum systems like CDMA systems through spreading [13]. The process of spreading the signal over the entire available bandwidth with corresponding despreading at the receiver results in suppression of the out-of-cell interference. This technique allows the neighboring base stations to operate independently and transmit simultaneously over the entire bandwidth thereby avoiding extensive frequency planning during the system deployment. On the other hand, when the base stations are allowed to cooperate by appropriately scheduling transmissions from the base stations it is possible to avoid interference from base stations in the neighborhood of the user. Interference avoidance can be superior in performance to interference mitigation through spreading.

When considering coordination among multiple base stations it is likely that in order to reap most of the benefits it is sufficient to restrict coordination to a few base stations that are closest to the user location. Furthermore, practical implementation constraints will force to limit coordination to such small sets. Specifically, we consider a restricted level of cooperation where only the three nearest base stations cooperate as shown in the Figure 3. Users inside the triangular region formed by the lines joining the three base stations A,B and C are allowed to be served simultaneously by one or more of the three base stations. All other base stations are assumed to be transmitting with full power and will be viewed as interfering base stations for the users in this region served by base stations A, B and C. Assuming a 60 degree sectorization, each base can schedule users within the cell but outside the triangular region independently of the users in the triangular region and hence those users need not be considered. Thus the coordination scheme involving three 60 degree sectors each from one of three adjacent cells can be extended throughout the network. To evaluate the potential gains from such a coordination scheme it is sufficient to focus on a single such coordination region. Since we are primarily interested in determining the potential gains available through inter-cell coordination in a packet data system, we assume that the channel gains h_k^i from each of the base stations to the users during any scheduling interval are available to each of the three base stations. It is easy to see that, under this model for inter-cell coordination, simultaneous transmission from the three base stations can simply be viewed within the framework of the multiple

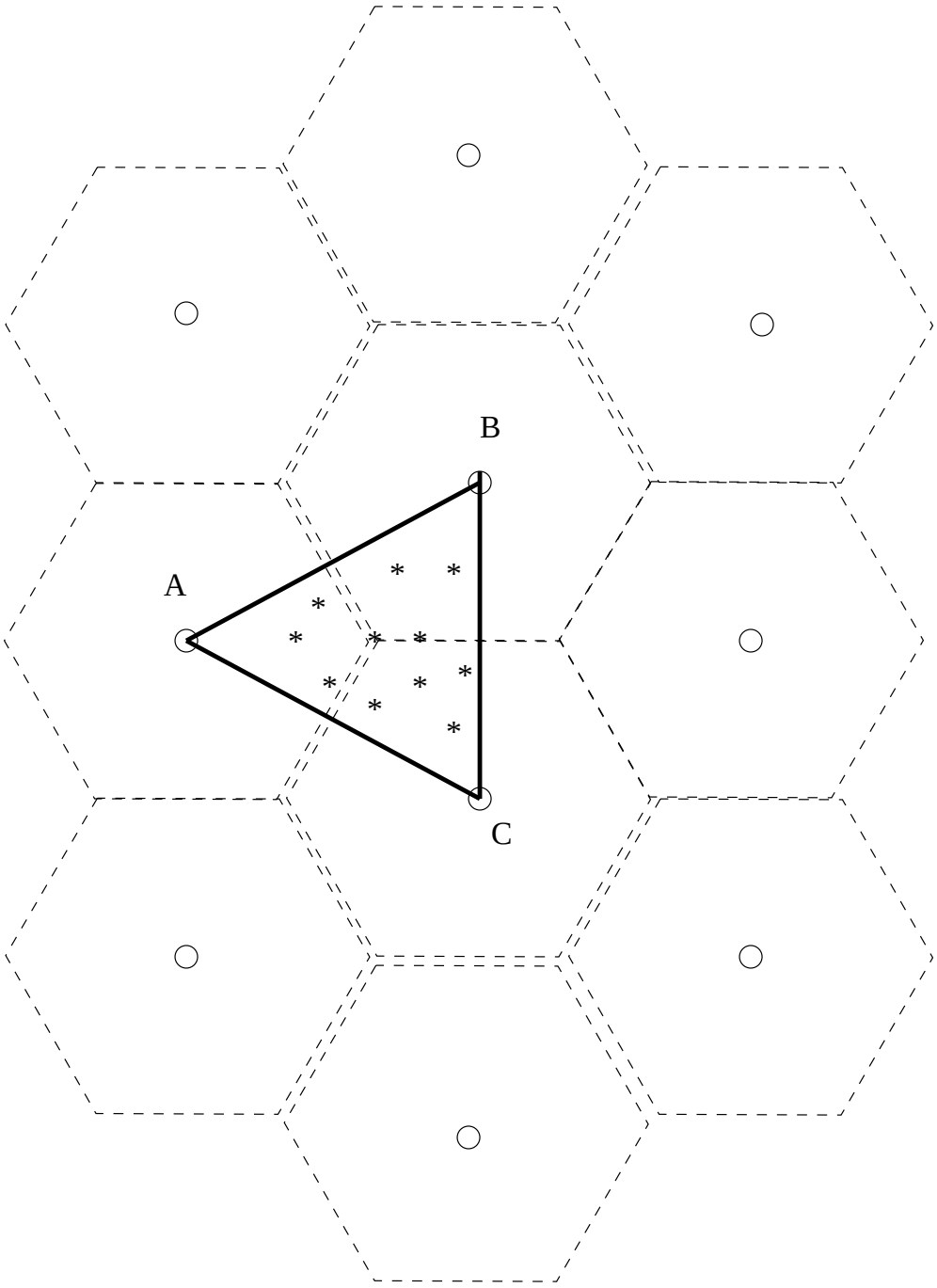

Figure 3: Inter-cell Coordination Scenario

antenna transmission from a single base station as considered in the previous sections. However, now the "antennas" are spatially distributed and hence will have different path loss from each transmit "antenna" to the users. Furthermore, the power constraint is a total power of P per "antenna" as opposed to the sum power constraint in the multi-antenna model in Section 5. One other significant difference is the fact that it would not be possible to coordinate the three spatially distributed transmitters for beam steering because in practice it will not be possible to synchronize the phase of the signals from the three transmitters to the required degree of accuracy to form beams. Furthermore, small differences in the propagation delay from each of the transmitters to the user will destroy the phase relationship in the signals when they arrive at the user. Thus the coding and beamforming strategies of Section 4 cannot be applied immediately to the inter-cell coordination based scheduling problem. Nevertheless, scheduling algorithms can still be derived from Proposition 1 with different transmission strategies as in the case of the algorithms in Section 5.

We present two different scheduling algorithms for the specific case of coordination in Figure 3. For simplicity we assume that each of the base stations has only a single antenna. In the first algorithm called *multi-user inter-cell*, the base stations jointly schedule to one or more users in the triangular region depending on the channel gains and the queue sizes of all the users. The second algorithm called *single-user inter-cell* is based on soft Han-doff in CDMA [13] for voice systems. In soft handoff, signals carrying the same information is transmitted from multiple base stations and are combined at the user prior to demodulation and decoding. Maximal ratio combining at the user is assumed for our simulation results. The last algorithm serves as the base-line for comparison and involves no inter-cell coordination. Here each base station transmits to a separate user in each scheduling interval.

7.1 Algorithms

Formal descriptions of the scheduling algorithms are given below. Denote by $\mathcal{T}$ the set of users in the service area of base stations A,B and C indexed by $1, 2$ and 3, respectively. The noise $\mathbf{v}_k$ in (2) now includes the receiver noise and the signals from all other base stations except base stations A, B and C. Let $h_k^i(t)$ be the channel gain from the single-antenna base station i to user k during scheduling interval t. As before, let $Q_k(t)$ be the queue size at the start of the scheduling interval t for user k. In what follows the scheduling interval is fixed and hence t is suppressed in $Q_k(t), h_k^i(t)$.

1. **Multi-user Inter-cell**: For any given scheduling interval, a subset $\mathcal{S}$ of $\mathcal{T}$ of size at most 3 is chosen for transmission. Each base station transmits to exactly one of the members of the set $\mathcal{S}$ at the peak power level P, or does not transmit at all. The subset $\mathcal{S}$ is chosen according the following optimization derived from the general result in Proposition 1.

$$(\mathcal{S}^*, \pi^*) \quad = \quad \arg \max_{\mathcal{S} \subset \mathcal{T}, \pi} \sum_{i \in \mathcal{S}} Q_k \log \left(1 + \frac{|h_k^{\pi(k)}|^2 P}{\sum_{i \in \mathcal{S}, i \neq k} |h_i^{\pi(i)}|^2 P + N_k} \right) \tag{15}$$

where $\mathcal{S}$ is all possible subsets of $\mathcal{T}$ of size 1, 2 or 3 and π is all possible permutations of assigning members of $\mathcal{S}$ to the three base stations such that each base station is transmitting to at most one user. As before, in (15) above we have assumed that the interfering signals are Gaussian and that the transmission rate is given by Shannon capacity of an AWGN channel with SNR equal to the SINR at the user. In essence, (15) determines the optimum subset of users to transmit to simultaneously, under the assumption that each base station transmits to at most one user with full transmit power. Typically, when users close to the base stations are being served more than 1 user can be served simultaneously since the interference from a neighboring base station at a user close to the desired base station will be minimal. On the other hand, when a user in the center of service area has to be served, two of the three base stations may be turned off to avoid interference. The optimization in (15) picks the appropriate set based on both the channel gains and the queue sizes of the various users. Given the optimum set $\mathcal{S}^*$ and the optimum permutation π^* the transmission rate to each user $k \in \mathcal{S}^*$ is given by

$$R_k = \log \left(1 + \frac{|h_k^{\pi^*(k)}|^2 P}{\sum_{i \in \mathcal{S}^*, i \neq k} |h_i^{\pi^*(i)}|^2 P + N_k} \right)$$

Clearly the above algorithm can be improved further at the expense of additional complexity by optimizing the transmission power level of each base station and by considering multiple base stations transmitting to the same user when $|\mathcal{S}| < 3$. However, our goal is to demonstrate gains form inter-cell coordination using the above algorithm.

2. **Single-user Inter-cell**: This algorithm picks a single user to be served by all three base stations with maximal ratio combining at the user. The user to transmit to in any given

scheduling interval is obtained as a solution to the optimization problem

$$k^* = \arg\max_{k\in\mathcal{T}} Q_k \log\left(1+\frac{\sum_{i=1}^{3}|h_k^i|^2P}{N_k}\right). \quad (16)$$

3. **Intra-cell**: In this algorithm at the start of each scheduling interval each user is assigned to the base station from which it has the largest SINR. This can be achieved in practice by feedback from the users and requires no cooperation between the base stations. Once the set of users is partitioned into three subsets $\mathcal{B}_i, 1 \le i \le 3$, each base station schedules transmission to a single user in its subset obtained by the solution to $\arg\max_{k\in\mathcal{B}_i} Q_k R_k$ where the transmission rate R_k for $k \in \mathcal{B}_i$ is given by

$$R_k = \log\left(1+\frac{|h_k^i|^2P}{N_k + P\sum_{j=1 j\neq i}^{3}|h_k^j|^2P}\right).$$

Thus three users are always served in every scheduling interval, one by each base station. However, unlike the multi-user inter-cell algorithm the set of users are chosen independently without a requirement for cooperation between base stations.

Note that this algorithm allows the serving base station for each user to change from one scheduling interval to another as in *fast cell site selection* in third generation high speed data systems [5].

7.2 Simulation Results

Hexagonal cells with uniform distribution of users in each cell as shown in Figure 3 is simulated. All base stations excepting A,B and C are assumed to be transmitting at full power. 15 users are placed in random locations drawn from a uniform distribution in the triangular region for each instance of the simulation. Delay-throughput results are obtained based on the fluid flow queuing system described in Section 2 for each instance corresponding to a given placement of users. Simulation is performed for 50 different instances of locations and averaged to get the final delay-throughput results.

Figure 4 shows the performance comparison between the intra-cell only scheduling and inter-cell scheduling described above. Average delays up to 1 second and 10 seconds are illustrated in the two figures, respectively. There is significant gain from doing inter-cell scheduling as seen from the figures.

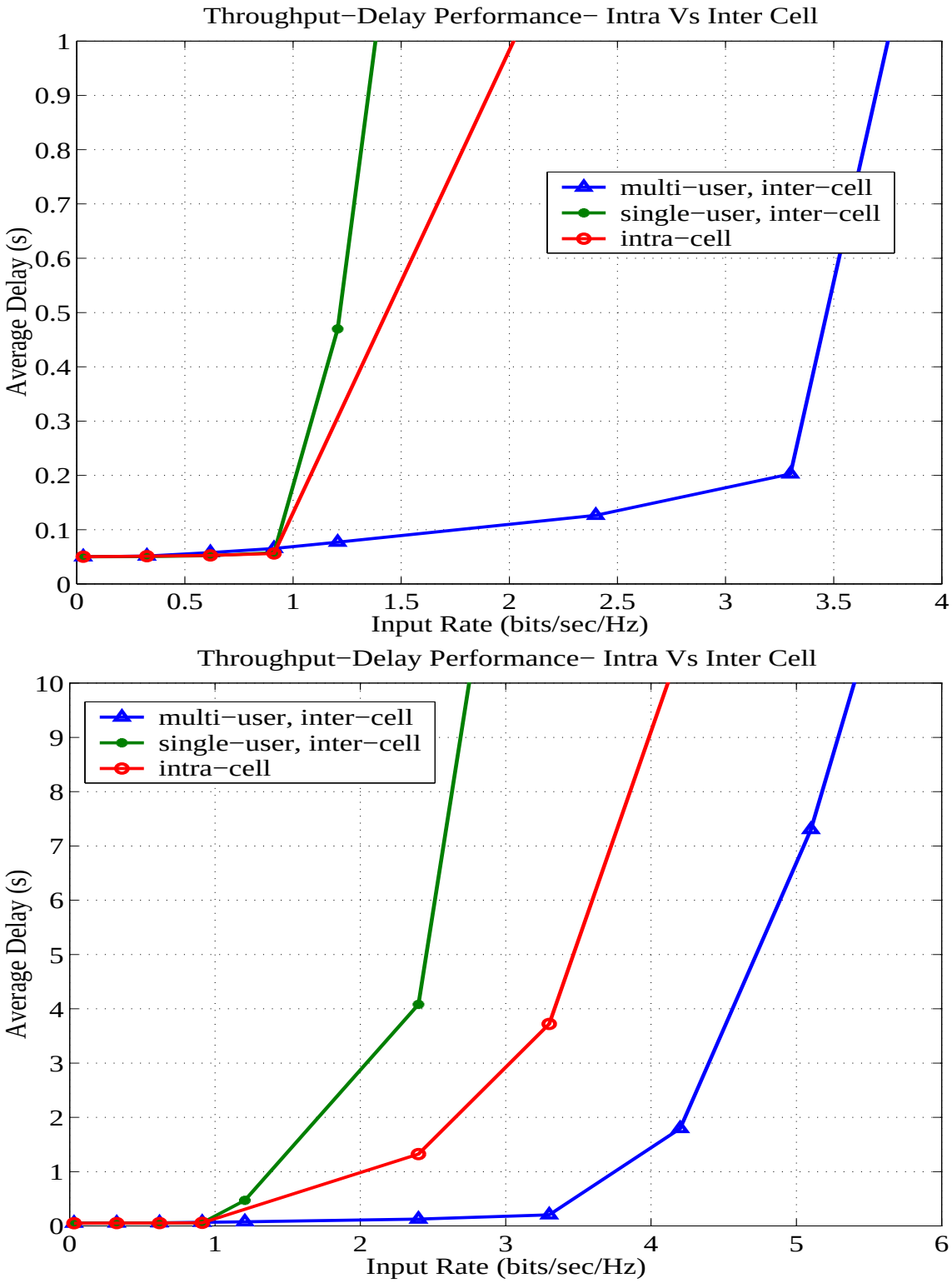

Figure 4: Performance for Intra-cell and Intercell Scheduling

8 Summary

Scheduling strategies for adaptive beamforming in multi-antenna downlink and for multiple fixed beam downlink were presented. The main theme for these scheduling algorithms is to transmit to multiple users in each scheduling interval with the the individual rates specified by the solution to an optimization problem. The maximum SNR beamforming based interference avoidance scheme showed modest gains over transmission to a single user at a time. The fixed beam scheme was based on a different antenna technology where antennas are designed to form small sidelobe beams allowing the possibility to transmit to multiple users simultaneously. This scheme has significant gains and requires only beam selection information apart from the SNR or rate feedback from the users as opposed to the vector channel knowledge required for the other scheme. The low complexity of implementation of this scheme also makes it attractive from an implementation standpoint. However, this scheme will not be applicable when the angle spread of propagation at the base station is large compared to the beam widths.

The multi-antenna scheduling idea was then applied to the multi-base scenario. Coordinated scheduling between three base stations in a cellular network was considered. Particular algorithms for joint scheduling were presented and the results were compared to the case of separate scheduling from the different base stations. Inter-cell scheduling showed significant gains of up to a factor of 2. However, these scheduling algorithms require that channel state information of every user is available to every base station base station in the region of interest. Similarly, all base stations should know the status of the queues for all the users.

All of the scheduling strategies considered in this paper were based on assuming separate coding of the information of the different users. When joint coding is also allowed, information-theoretic optimal transmission strategy and corresponding optimal rate region can also be combined with Proposition 1 to obtain optimum scheduling algorithms.

Appendix

Proposition 4 *The set of achievable rate vectors $\mathcal{R}(\mathbf{h}_1, \cdots, \mathbf{h}_K)$ for multi-antenna downlink Gaussian broadcast channel is bounded by the region $\mathcal{R}^*(\mathbf{h}_1, \cdots, \mathbf{h}_K)$ which is given by (for any set $\mathcal{A}$,*

$\mathcal{A}+\mathcal{B}=\{\mathbf{a}+\mathbf{b}:\mathbf{a}\in\mathcal{A},\mathbf{b}\in\mathcal{B}\ \}$)

$$\mathcal{R}^*(\mathbf{h}_1,\cdots,\mathbf{h}_K) \;=\; \max_{\sum_{i=1}^M P_i=P}\sum_{i=1}^M \mathcal{R}_i\left(P_i,\{\gamma_k^i\}_{k=1}^K\right)$$

where γ_k^m is the inverse of the effective noise power for the k^{th} user in the i^{th} parallel channel as given below

$$\gamma_k^i \;=\; \begin{cases} \frac{\|\mathbf{h}_k\|^2}{N_k} & \text{if } h^k(i)\neq 0 \\ 0 & \text{otherwise} \end{cases}$$

and $\mathcal{R}_i\left(P_i,\{\gamma_k^i\}\right)$ are the AWGN degraded broadcast channel capacity region with total power P_i and is given below

$$\mathcal{R}_i\left(P_i,\{\gamma_k^i\}_{k=1}^K\right)=\left\{\mathbf{R}:R_k\leq\log\left(1+\frac{\alpha_k\gamma_k^iP_i}{1+\sum_{j<k}\alpha_j\gamma_k^iP_i}\right)\right\},$$

assuming the users are ordered such that $\gamma_1^i\geq\gamma_2^i\cdots\geq\gamma_K^i$.

Proof: Consider the genie aided system where the received signal for n^{th} symbol period at user k is given by the vector $\mathbf{Y}_k^n=(Y_k^n(1),\cdots,Y_k^n(M))$, with the i^{th} component given by

$$Y_k^n(i) \;=\; \begin{cases} h^k(i)x^n(i)+V_k^n(i) & \text{if } h^k(i)\neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $V_k^n(i)$ is additive white Gaussian noise with variance $N_k(i)=N_k\frac{|h_k(i)|^2}{\|\mathbf{h}_k\|^2}$ and is independent across the different components i and users k. Note that the noise variances are chosen to satisfy the condition $\sum_{i=1}^M N_k(i)=N_k$. Consider the following signal derived from the received signal in the genie-aided system

$$\begin{aligned}\tilde{Y}_n^k &= \sum_{i=1}^M Y_n^k \\ &= \sum_{i=1}^M h_k(i)x^n(i)+\sum_{i=1}^M V_k^n(i). \end{aligned}\tag{17}$$

Comparing the received signal in equation (1) at k^{th} receiver in the actual system and the above derived received signal for the genie-aided system in equation (17), we note that the transmitted signal component of the two signals are identical and the AWGN variances are equal. Hence any encoding and decoding rules applied to the actual system can be applied to the genie-aided system on the derived signal to obtain the same performance. Hence any rate that is achievable in the actual system is also achievable in the genie-aided system. Thus the rate region of the actual system is contained within the rate region of the genie aided system. However, since the genie-aided system has access to the vector signal $\mathbf{Y}_k^n$ its performance can be superior to the actual system. Thus the rate region of the genie-aided system is an outer bound to the actual rate region.

The genie-aided system is clearly a set of identical M parallel degraded Gaussian broadcast channels and hence its rate region is computable. Thus equation (17) now follows from the characterization of the degraded Gaussian broadcast channel rate region [10].

References

[1] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar. Providing quality of service over a shared wirless link . *IEEE Communications Magazine*, 39(2):150–154, Feb 2001.

[2] D. Bertsekas and R. Gallager. *Data Networking.* Prentice-Hall Inc., 1987.

[3] G. Caire and S. Shamai. Achievable Rates in Multi-antenna Broadcast. In *Proc. of Allerton Conference on Control and Communications*, 2000.

[4] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, 1991.

[5] P. Bender et al. CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users. *IEEE Communications Magazine*, 38(7):70–77, July 2000.

[6] A. Jalali, R. Padovani, and R. Pankaj. Data throughput of CDMA-HDR: A high efficiency, high data rate personal wireless system. In *Proceedings of the IEEE Vehicular Technology Conference*, May 2000.

[7] N. Kahale and P. E. Wright. Dynamic Global Packet Routing in Wireless Networks. In *Proc. of INFOCOM'97*, 1997.

[8] S. Shakkottai and A. Stolyar. A study of scheduling algorithms for a mixture of real and non-real time data in HDR. In *Presented at the 17th International Teletraffic Congress (ITC-17)*, September 2001.

[9] A. Stolyar and S. Shakkottai. Scheduling for multiple flows sharing a time-varying channel: the exponential rule. In *To appear in the Transalations of AMS, a volume in memory of F. Karpelvich.* American Mathematical Society, 2001.

[10] D. Tse. Optimal power allocation over parallel Gaussian broadcast channels. In *Proceedings of the International Symposium on Information Theory*, 1997.

[11] D. Tse. Forward-link multi-user diversity through rate adaptation and scheduling. In *Bell Labs Presentation*, 1999.

[12] E. Visotsky and U. Madhow. Optimum Beamforming Using Transmit Antenna Arrays. *Proc. of VTC'99*, 1999.

[13] A.J. Viterbi. *CDMA Principles of Spread Spectrum Communication.* Addision-Wesley Wireless Communication Series, 1995.