Optimal Transmission Scheduling with Base Station Antenna Array in Cellular Networks

Tianmin Ren, Richard J. La and Leandros Tassiulas
Department of Electrical & Computer Engineering
and Institute for Systems Research
University of Maryland, College Park, MD 20742, USA
e-mail: {rtm, hyongla, leandros}@isr.umd.edu

***Abstract*— We study the downlink scheduling problem in a cellular wireless network. The base stations are equipped with antenna arrays and can transmit to more than one mobile user at any time instant, provided the users are spatially separable. In previous work, an infinite traffic demand model is used to study the physical layer beamforming and power control algorithms that maximize the system throughput. In this paper we consider finite user traffic demands. A scheduling policy makes a decision based on both the queue lengths and the spatial separability of the users. The objective of the scheduling algorithm is to maintain the stability of the system. We derive an optimal scheduling policy that maintains the stability of the system if it is stable under any scheduling policy. However, this optimal scheduling policy is exponentially complex in the number of users which renders it impractical. We propose four heuristic scheduling algorithms that have polynomial complexity. The first two algorithms are for the special case of single cell systems, while the other two algorithms deal with multiple cell systems. Using a realistic multi-path wireless channel model, we evaluate the performance of the proposed algorithms through computer simulations. The results demonstrate the benefits of joint consideration of queue length and dynamic base station assignment.**

I. Introduction

Wireless communication has been experiencing rapid development during the past decade. The increasing demand for fast wireless access and high-speed wireless links to users has been the driving force for active research in the telecommunications area. At present, wireless communication is undergoing a transition from the traditional circuit switched voice services to packet switched data services. A variety of data applications are implemented or proposed to provide mobile users with ubiquitous access to information of any kind. The advent of wireless applications such as wireless multimedia, *e.g.,* video conferencing, is only the first sign of the projected demand for rapid and reliable wireless data access.

New network architectures and protocols are proposed to support data applications in wireless networks. A typical architecture in many of current wireless systems, especially cellular networks, provides the wireless access to mobile users through access points (APs) or base stations (BSs) that are connected to the core network, which is a wireline network. For instance, 3G protocols have been standardized and are being implemented to provide mobile users with wireless data access. The most challenging task in designing these communication systems is to guarantee the quality of service (QoS) requirement to various data applications on wireless channels with limited bandwidth and time varying characteristics. Different notions of QoS are available at different communication layers. QoS at the physical layer is expressed as an acceptable signal to interference and noise ratio (SINR) or corresponding bit error rate (BER) at the receiver. At the MAC layer QoS is usually expressed in terms of achievable bit rate or packet error rate (PER), while at the higher layers QoS can be perceived as a minimum throughput and/or maximum delay requirement. The ability of a network to satisfy the QoS requirements and enhance the system capacity depends on the interaction of several layers.

A wide spectrum of approaches have been proposed to reuse the communication resources in time, frequency and/or space domain, in order to provide the QoS guarantee to mobile users and improve the capacity of the wireless networks. Among these approaches, the application of antenna arrays, which explores the spatial diversity of mobile users, is considered a more promising one and the last frontier for future capacity improvement of wireless networks. This is due to the beamforming capability of the antenna arrays that can form the beam pattern directed to a desired user while nulling out the others. This helps greatly suppress co-channel interference, and spatially separable users can share the same channel with their QoS requirements satisfied.

Previous research on the application of antenna arrays in cellular networks can be categorized into two classes. The first class of research is on the physical layer; given a set of users, the problem is to design optimal algorithms to calculate the beamforming weights for each user. The problem is modeled as an optimization problem, where the objective is to minimize the total transmission power subject to the constraint that each user's SINR requirement is satisfied. Note that this problem may be infeasible, that is, there does not exist a set of beamforming weights such that each user's SINR value is larger than the required minimum threshold. In [1], iterative algorithms are proposed to minimize total transmitted power subject to the constraint that SINR of each user is satisfied for downlink transmissions in a single cell network. In [3], the problem of joint beamforming and base station assignment is considered, where each user can be served by any base station in the network. An algorithm that assigns each user to the optimal base station and computes the corresponding transmit beam pattern for each user is designed.

0-7803-8356-7/04/$20.00 (C) 2004 IEEE

The second class of research is on the MAC layer with consideration of physical layer user separability constraints. The goal of this approach is, given a set of users, to place as many users on the same channel as possible and compute the beamforming weights for each selected user subject to the SINR constraint. This helps maximize the throughput of the network. Here a channel can be a time slot in a TDMA system, a subcarrier in an OFDM system, or a code in a CDMA system. Algorithms aimed at maximizing total throughput are proposed in the literature [4]. These algorithms are based on the same idea of inserting users into a channel in a sequential manner, and vary in the criteria that determine the order in which users are inserted. This problem is extended to be combined with other multi-user access schemes such as TDMA, OFDM and CDMA in [5]. A common assumption in these studies is infinite packet backlog for any user, *i.e.,* there is always a packet to be served at the queue for each user. The major drawbacks of these works are the limitation of the focus on instant total throughput maximization and the lack of consideration of upper layer QoS requirements for each individual user. Thus, the assignment of users on each channel only reflects the feasibility at the physical layer, but not the current buffer occupancy or traffic demand of each user. This separation of physical layer algorithms and upper layer QoS requirements leads to the degradation of overall system performance, *e.g.,* user perception. Therefore higher layer QoS requirements need to be taken into consideration for the design of efficient MAC and physical layer algorithms. Moreover, the MAC layer scheduling policy and physical layer beamforming algorithms need to be considered jointly for QoS provisioning to users.

In this paper we study the problem of designing a scheduling algorithm for a central controller that handles multiple BSs, where each BS is equipped with an antenna array. Packets arrive at the central controller for transmission to mobile users. Buffer occupancy and traffic demand of each user are considered explicitly. In addition to feasibility of users sharing the same channel, the scheduling policies consider the current buffer occupancy and, thus, reflect the QoS requirement of each user in terms of throughput. We model this problem as a queueing system with multiple parallel servers. The SINR requirement constraints are imposed on the selection of users that can be served in each time slot. Instead of a policy that maximizes instant throughput, we seek an optimal scheduling policy that stabilizes the system if it is stable under some policy. In addition, under this optimal scheduling policy the user throughput requirements are satisfied and, hence, the long term total system throughput is maximized.

Similar queueing systems have been used to model other scenarios in [6], [7], [8], and was first proposed in [6] for a multihop radio network where the SINR requirement demands that two links can be active simultaneously only if they are separated by certain minimum required distance. The throughput region is defined as the set of arrival rate vectors for which the system is stable. The optimal scheduling policy that stabilizes the system whenever it is stable under some policy is identified. However, the complexity of an optimal scheduling policy increases exponentially with the number of users, and no practical sub-optimal scheduling policy is proposed in [6], [7], [8]. In this paper, we follow a similar approach as in [8], and propose scheduling policies of polynomial complexity that achieve sub-optimal performance for our problem.

This paper is organized as follows. In Section II we describe the problem of designing an efficient downlink scheduling algorithm with base station antenna arrays and derive the optimal scheduling policy based on feasible rate vectors. In Section III we propose heuristic algorithms to approximate the optimal scheduling policy with polynomial complexity. We evaluate and compare the performance of the proposed algorithms in Section IV. We conclude in Section V.

II. Optimal downlink scheduling with base station antenna arrays in cellular networks

A. System model

We consider a wireless network that consists of several base stations. Each base station is equipped with an antenna array so that several users can be served simultaneously. These base stations are coordinated by a single central controller. Mobile users in the network are able to receive data packets from any of these base stations. However, at any given time, a mobile user can receive data packet from at most one base station. The central controller maintains a separate queue for incoming data packets destined for each mobile user. We assume a time slotted system where the transmission time of each packet equals to one time slot if the lowest transmission rate is selected. In each time slot, the central controller collects the information regarding the wireless links of each user to different base stations. Based on this information and the number of backlogged packets of each user, the central controller assigns base stations to the users with respective transmission rates and calculate the beamforming weights that will be used by each assigned base station. The scheduling decision made by the central controller includes assignment of base stations to the users and the transmission rate of each user. The beamforming weights are calculated to support the scheduling decision.

The block diagram of the system under study is depicted in Fig. 1. User packets enter the scheduling module at the central controller, which determines the assignments of base stations and transmission rates. Beamforming and power adaptation are subsequently calculated for each BS for scheduled users. The transmitter of a BS can form at most M beams for scheduled users at the same time, where M is the number of antenna elements. We assume that there are M transceivers at each BS. A beam is formed by a dedicated transceiver and appropriate power is allocated to each scheduled user. Scheduling and beamforming are interdependent operations, and they also depend on queueing state and channel state information, which are assumed to be available at the central controller.

B. Problem statement

The network consists of I base stations shared by J mobile users. We denote the set of base stations by $\mathcal{I}$ and the set of users by $\mathcal{J} = \{1, \ldots, J\}$. There is a central controller that coordinates the operation of the I base stations. Each base station is equipped with an M-element antenna array. The users receive data packets from the base stations.

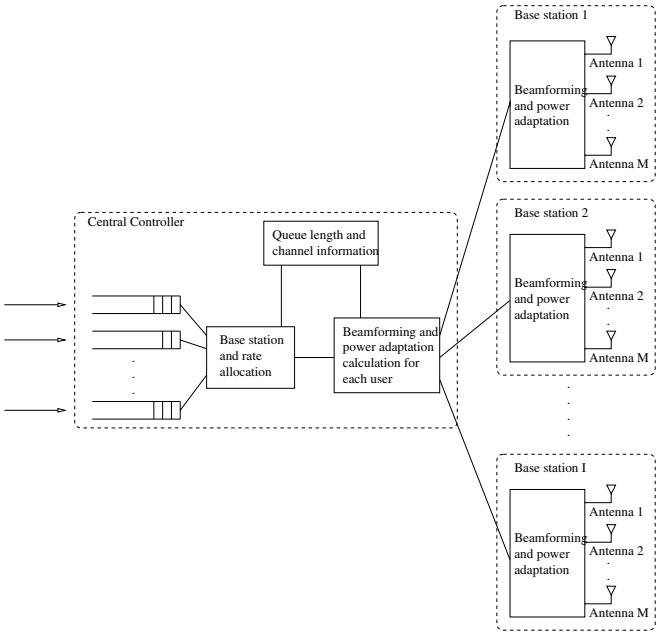

Fig. 1. The multiple cellular communication system

Several transmission rates are available for transmission to a user based on the channel conditions. The set of available transmission rates is denoted by $\mathcal{V}$. We assume each rate is a positive integer number. If rate $v \in \mathcal{V}$ is chosen, up to v packets can be transmitted in one time slot, depending on the number of packets waiting for transmission. We denote $|\mathcal{V}| = V$.

Packets arrive at the central controller for transmission to the users, which maintains a separate queue for each user. Let $a_j(t), j = 1, 2, \cdots, J$ and $t = 0, 1, \cdots$, denote the number of packets that arrive at queue j in time slot t. We assume that $a_j(t), t = 0, 1, 2, \ldots$, are i.i.d. random variables (rvs) with a finite second moment, *i.e.,* $E[a_j(t)^2] < \infty$. The number of backlogged packets for user j at queue j at the start of time slot t is given by $x_j(t)$. The average arrival rate of user j is denoted by $A_j = E[a_j(t)]$. We call $\mathbf{A} = (A_1, A_2, \cdots, A_J)^T$ an arrival vector.

We assume that the central controller has perfect channel information for each user with regard to every base station. In each time slot, the central controller (i) assigns the base stations to the users, (ii) computes transmission rates of each scheduled user, and (iii) calculates the beamforming weights of scheduled users for each base station. A scheduling decision can be expressed as an $I \times J$ matrix $\mathbf{R} = [r_{ij}]$ where the element $r_{ij} \in \mathcal{V}, i = 1, \cdots, I$, and $j = 1, \cdots, J$, is the transmission rate of base station i to user j. A rate matrix is feasible if and only if SINR requirement is satisfied for each user and each user receives packets from at most one base station.

The channel conditions change with time. Therefore, the feasibility of a rate matrix is also time varying. We model the channel process for all users as a Markov chain (MC) with a stationary distribution π. Each channel state represents the set of all feasible rate matrices, *i.e.,* a state of the MC is a set of all rate matrices that can be used for transmission given the channel condition. Let $\mathcal{S}$ be the channel state space. Our problem is to find the optimal scheduling policy that selects a feasible rate matrix in each time slot given the queue lengths and channel state, such that the system achieves maximum system throughput, while maintaining a stable system whenever possible under some policy. We define and characterize the throughput region in the following subsection.

C. Throughput region

Let us first define a stable arrival vector.

Definition 1: An arrival vector $\mathbf{A}$ is said to be stable if there exists a scheduling policy such that

$$\lim_{c\to\infty} \limsup_{t\to\infty} P(x_j(t) > c) = 0 \,, \quad \text{for all } j = 1, 2, \cdots, J \tag{1}$$

where $x_j(t)$ is the number of backlogged packets in queue j at the start of time slot t. If a scheduling algorithm satisfies (1), then we say that $\mathbf{A}$ is stable under the scheduling policy. The set of stable arrival vectors is called the throughput region, denoted by $\mathcal{A}$.

The following proposition characterizes the throughput region $\mathcal{A}$.

Proposition 1: For an arrival vector $\mathbf{A}$, the necessary and sufficient condition for $\mathbf{A}$ to be stable, *i.e.,* $\mathbf{A} \in \mathcal{A}$, is that there exists a scheduling policy that achieves

$$\mathbf{A} \le \mathbf{D} := \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} c_{\mathbf{SR}} \mathbf{R}^T \mathbf{1}_{I\times 1} \tag{2}$$

where $c_{\mathbf{SR}}, \mathbf{S} \in \mathcal{S}, \mathbf{R} \in \mathbf{S}$, are nonnegative numbers such that $\sum_{\mathbf{R}\in\mathbf{S}} c_{\mathbf{SR}} = 1$ for all $\mathbf{S} \in \mathcal{S}$.

Proof: See [8] for a proof. ■

D. Optimal scheduling policy

In this subsection we are interested in finding an optimal scheduling policy that satisfies (1) for each $\mathbf{A} \in \mathcal{A}$. In particular, we consider the following scheduling policy: given backlog vector $\mathbf{X(t)} = (x_1(t), \ldots, x_J(t))^T$ and system channel state $\mathbf{S(t)}$, the rate vector selected by the scheduling algorithm is given by

$$\mathbf{R(t)} = \arg \max_{\mathbf{R}\in\mathbf{S(t)}} \mathbf{X(t)}^{\mathbf{T}} (\mathbf{R}^{\mathbf{T}} \mathbf{1}_{I\times 1}) \tag{3}$$

where ties are broken arbitrarily.

The backlog process $\mathbf{X(t)}$ is a J-dimensional Markov process with countably infinite state space given that the scheduling policy is stationary, *i.e.,* the decisions made by the scheduling policy do not depend on time slot t, but only on $\mathbf{X(t)}$ and channel state $\mathbf{S(t)}$.

Define the following Lyapunov function

$$L(\mathbf{X}(t)) = \sum_{j=1}^{J} (x_j(t))^2 \ .$$

In order to prove the existence of a stationary distribution of $\mathbf{X(t)}$ and, hence, the stability of the system, we use the following theorem.

Theorem 1: ([9], [10]) For a given Lyapunov function $L(\mathbf{X}(\mathbf{t}))$, if there exists a compact region Σ of $\Re^J$ and a constant $\alpha > 0$ such that

1) $E[L(\mathbf{X}(\mathbf{t}+\mathbf{1}))|\mathbf{X}(\mathbf{t})] < \infty$ for all $\mathbf{X}(\mathbf{t}) \in \Re^J$
2) $E[L(\mathbf{X}(\mathbf{t}+\mathbf{1})) - \mathbf{L}(\mathbf{X}(\mathbf{t}))|\mathbf{X}(\mathbf{t})] \leq -\alpha$ whenever $\mathbf{X}(\mathbf{t}) \in \Sigma^C := \Re^J \setminus \Sigma$,

then a steady state distribution of the vector $\mathbf{X}(\mathbf{t})$ exists and, thus, the system is stable.

Essentially the theorem states that it suffices to show that there is a negative drift in the Lyapunov function when the backlogs are sufficiently large.

Now we state a proposition that establishes the optimality of scheduling policy given by (3).

Proposition 2: Suppose that $\mathbf{A} \in int(\mathcal{A})$, where $int(\mathcal{A})$ is the interior of the throughput region $\mathcal{A}$. Then, the system is stable under the scheduling policy given by (3).

Proof: The evolution of the backlog vector $\mathbf{X}(\mathbf{t})$ is given by the following recursive equation:

$$\mathbf{X}(\mathbf{t}+\mathbf{1}) = \max(\mathbf{X}(\mathbf{t}) + \mathbf{A}(\mathbf{t}) - \mathbf{D}(\mathbf{t}), 0)$$

It is clear that property 1 in Theorem 1 holds. Now we prove property 2 of the theorem.

$$\begin{aligned} x_j^2(t+1) &\leq (x_j(t) + a_j(t) - d_j(t))^2 \\ &\leq x_j(t)^2 - 2x_j(t)d_j(t) + 2x_j(t)a_j(t) \\ &\quad + d_j(t)^2 + a_j(t)^2 \end{aligned}$$

Using the above inequality

$$\begin{aligned} &E[L(\mathbf{X}(\mathbf{t}+\mathbf{1})) - L(\mathbf{X}(\mathbf{t}))|\mathbf{X}(\mathbf{t})] \\ &\leq \sum_{j=1}^{J} E[a_j(t)^2|\mathbf{X}(\mathbf{t})] + \sum_{j=1}^{J} E[d_j(t)^2|\mathbf{X}(\mathbf{t})] \\ &\quad -2\sum_{j=1}^{J} x_j(t)E[d_j(t) - a_j(t)|\mathbf{X}(\mathbf{t})] \\ &\leq B - 2\sum_{j=1}^{J} x_j(t)(E[d_j(t)|\mathbf{X}(\mathbf{t})] - A_j(t)) \end{aligned}$$

where $B := \sum_{j=1}^{J} E[a_j(t)^2] + J \cdot v_M^2$ since $\sum_{j=1}^{J} E[d_j(t)^2|\mathbf{X}(\mathbf{t})] \leq J \cdot v_M^2$, where $v_M = \max \mathcal{V}$ is the largest transmission rate available.

Since $\mathbf{A}$ lies in $int(\mathcal{A})$, we have

$$\begin{aligned} \sum_{j=1}^{J} x_j(t)A_j &\leq \sum_{j=1}^{J} x_j(t)D_j \\ &= \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} c_{\mathbf{SR}} \sum_{j=1}^{J} x_j(t)r_j(\mathbf{R}) \\ &\leq \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \max_{\mathbf{R}\in\mathbf{S}} \sum_{j=1}^{J} (x_j(t)r_j(\mathbf{R})) \\ &= \sum_{j=1}^{J} x_j(t)E[d_j(t)|\mathbf{X}(\mathbf{t})] \end{aligned}$$

where D_j is the j-th element of $\mathbf{D}$ with a scheduling policy that satisfies the inequality in (2), and $r_j(\mathbf{R})$ is the j-th element of vector $\mathbf{R}^T\mathbf{1}_{I\times 1}$.

Since $\mathbf{A} \in int(\mathcal{A})$, we can find a $J\times 1$ vector $\varepsilon = (\epsilon, \cdots, \epsilon)^T$ such that $\mathbf{A} + \varepsilon$ belongs to $int(\mathcal{A})$ and satisfies

$$\sum_{j=1}^{J} x_j(t)(A_j + \epsilon) \leq \sum_{j=1}^{J} x_j(t)E[d_j(t)|\mathbf{X}(\mathbf{t})] \,.$$

Therefore

$$\begin{aligned} &\sum_{j=1}^{J} x_j(t)(E[d_j(t)|\mathbf{X}(\mathbf{t})] - A_j) \\ &= \sum_{j=1}^{J} x_j(t)(E[d_j(t)|\mathbf{X}(\mathbf{t})] - (A_j + \varepsilon) + \varepsilon) \\ &\geq \epsilon \sum_{j=1}^{J} x_j(t) \end{aligned}$$

and

$$E[L(\mathbf{X}(\mathbf{t}+\mathbf{1})) - L(\mathbf{X}(\mathbf{t}))|\mathbf{X}(\mathbf{t})] \leq B - 2\epsilon \sum_{j=1}^{J} x_j(t) \,.$$

For any positive α, we can define a compact region

$$\Sigma_\alpha = \left\{ \mathbf{X}(\mathbf{t}) \in \Re^J \,\middle|\, \sum_{j=1}^{J} x_j(t) \leq \frac{B+\alpha}{2\epsilon} \right\} \,.$$

It is an easy exercise to show that whenever $\mathbf{X}(\mathbf{t}) \in \Re^J \setminus \Sigma_\alpha$, we have $E[L(\mathbf{X}(\mathbf{t}+\mathbf{1})) - L(\mathbf{X}(\mathbf{t}))|\mathbf{X}(\mathbf{t})] \leq -\alpha$, hence satisfying the second condition in Theorem 1. Therefore, by Theorem 1 the system is stable under the scheduling policy given by (3). ■

III. Heuristic downlink scheduling algorithms with base station antenna arrays

In the previous section we have derived an optimal scheduling policy. The next natural question is how to optimally assign the base stations and allocate rates to users given the queue lengths of the users in each time slot in practice. If the user channels are constant, the central controller can exhaustively search for all possible combinations offline, and select the one that maximizes (3) in each slot. However if the channels vary with time, this exhaustive search is not practical, if not impossible, even for a single cell system because the number of possible rate vectors is given by

$$c_1 = \sum_{j=1}^{J} \binom{J}{j} V^j \,,$$

which increases exponentially with the number of users. Therefore, one needs to design practical online scheduling policies with lower complexity that yield close to optimal performance.

In the previous sections, we have abstracted out the spatial separability of users using feasible rate matrices. This abstraction hides the physical channel characteristics. In the following subsection we present a physical channel model we adopt in the rest of this paper. Based on this model, we will propose joint scheduling, beamforming, and power control algorithms with polynomial complexity in the number of users, and study their performance in terms of average packet delay. We start with a single cell system followed by a multiple cell system.

A. Single cell system

In this subsection we consider simpler single cell systems to illustrate the basic intuition behind our proposed algorithms. We first describe the physical channel model that will allow us to capture channel conditions between a user and antenna arrays and compute the SINR values of the users given their power levels, beamforming vectors, and channel conditions.

1) Physical channel model and downlink beamforming algorithm: In making a scheduling decision, the first step that needs to be taken is to test if a set of users with their respective rates can be served at the same time. In order to decide whether a set of users can be served simultaneously or not, we need a physical channel model that will allow to check whether or not the required SINR constraint for a user is satisfied with the used beamforming vectors.

We first describe the adopted physical channel model [11] followed by the introduction of the downlink beamforming algorithm. The multi-path channel between user j and the m-th antenna element in the antenna array is given by

$$h_j^m(t) = \sum_{\ell=1}^{L} \beta_{j,\ell}\, \delta\left(t - \tau_{j,\ell} + \tau_{j,\ell}^m\right), \tag{4}$$

where L is the number of paths, $\beta_{j,\ell}$ is the complex gain of the ℓ-th path of user j, and $\tau_{j,\ell}$ is the delay for that path with respect to a reference antenna element. The gain $\beta_{j,\ell}$ is a complex random variable with variance $A_{j,\ell}$. The term $\tau_{j,\ell}^m = (d/c)(m-1)\cos\theta_{j,\ell}$ captures the delay to the m-th antenna, where d is the distance between two adjacent antenna elements, $\theta_{j,\ell}$ is the angle of the ℓ-th path of user j, and c is the electromagnetic wave propagation speed. In the rest of the paper we assume that the major limitation is co-channel interference rather than noise so that SINR can be approximated by SIR.

The received signal at the receiver of user j is

$$y_j(t) = \sum_{k\in\mathcal{U}} \sqrt{P_k} \sum_{m=1}^{M} u_k^m \sum_{\ell=1}^{L} \beta_{j,\ell}(\omega_0) e^{j\omega_0\tau_{j,\ell}^m} s_k(t-\tau_{j,\ell}),$$

where $\mathcal{U}$ is the set of users scheduled on the same channel, and u_k^m is the beamforming weight of user k by the m-th antenna element. Define the m-th element of the $M\times 1$ antenna steering vector $\mathbf{v}_0(\theta_{j,\ell})$ at direction $\theta_{j,\ell}$ and frequency ω_0 as $v_0^m(\theta_{k,\ell}) = e^{j\omega_0\tau_{j,\ell}^m}$. Then, the vector $\mathbf{a}_{0,j} = \sum_{\ell=1}^{L}\beta_{j,\ell}^*(\omega_0)\mathbf{v}_0^*(\theta_{j,\ell})$ is called the spatial signature of user j at ω_0 and captures spatial and multi-path properties of the user. If we omit subscript "0" from $\mathbf{v}$, the average interference caused by user k to user j is

$$\begin{aligned}
&E\left\{\left|\sqrt{P_k}\sum_{m=1}^{M} u_k^m \sum_{\ell=1}^{L}\beta_{j,\ell}(\omega_0)e^{j\omega_0\tau_{j,\ell}^m}s_k(t-\tau_{j,\ell})\right|^2\right\}\\
&= P_k\mathbf{u_k}^H\Big(\sum_{\ell_1=1}^{L}\sum_{\ell_2=1}^{L}\mathbf{v}(\theta_{j,\ell_1})v^H(\theta_{j,\ell_2})\\
&\quad\times E\{\beta_{j,\ell_1}(\omega_0)\beta_{j,\ell_2}^*(\omega_0)\}\\
&\quad\times E\{s_k(t-\tau_{j,\ell_1})s_k^*(t-\tau_{j,\ell_2})\}\Big)\mathbf{u_k}\\
&= P_k({\mathbf{u}_k}^H\mathcal{H}_j\mathbf{u}_k)
\end{aligned}$$

Observe that

$$E\left\{\beta_{j,\ell_1}(\omega_0)\beta_{j,\ell_2}^*(\omega_0)\right\} = \begin{cases} 0, & \text{if } \ell_1\neq\ell_2\\ A_{j,\ell}, & \text{if } \ell_1=\ell_2=\ell,\end{cases}$$

assuming that all paths are independent and signal power is normalized. Then we have,

$$\mathcal{H}_j = \sum_{\ell=1}^{L} A_{j,\ell}\,\mathbf{v}(\theta_{j,\ell})\mathbf{v}^H(\theta_{j,\ell})\,. \tag{5}$$

The matrix $\mathcal{H}_j$ is called spatial covariance matrix of user j, and in general it has rank$(\mathcal{H}_j) > 1$. The average SIR at the receiver of user j, denoted by W_j is given by

$$W_j = \frac{P_j\left(\mathbf{u}_j^H\mathcal{H}_j\mathbf{u}_j\right)}{\sum\limits_{\substack{k\in\mathcal{U}\\k\neq j}} P_k\left(\mathbf{u}_k^H\mathcal{H}_j\mathbf{u}_k\right)}\,. \tag{6}$$

Let $\mathbf{U}$ denote the ensemble of computed beamforming vectors for all users, *i.e.,* $\mathbf{U} = \{\mathbf{u}_j, j\in\mathcal{U}\}$. We define a $|\mathcal{U}|\times|\mathcal{U}|$ matrix $\mathbf{A}(\mathbf{U}) = [a_{ij}, i,j\in\mathcal{U}]$ where a_{ij} gives the co-channel interference caused by the j-th user to the i-th user, normalized by the signal power of user i, *i.e.,*

$$a_{ij} = \begin{cases} 1, & \text{if } i=j\\ \frac{T(v_i)\mathbf{u}_j^H\mathcal{H}_i\mathbf{u}_j}{\mathbf{u}_i^H\mathcal{H}_i\mathbf{u}_i}, & \text{otherwise.}\end{cases}$$

where $T(v_i)$ is the required SIR threshold that is a function of rate v_i.

Matrix $\mathbf{A}(\mathbf{U})$ is non-negative definite and irreducible. From the Perron-Frobenius theorem, the only eigenvector with strictly positive components is the one that corresponds to the maximum eigenvalue of $\mathbf{A}(\mathbf{U})$, denoted by $\lambda_{max}(\mathbf{A}(\mathbf{U}))$.

We introduce matrix $\mathbf{B}(\mathbf{U}) = [b_{ij}, i,j\in\mathcal{U}]$, whose elements are related to those of $\mathbf{A}(\mathbf{U})$ as follows:

$$b_{ij} = \begin{cases} a_{ij}, & \text{if } i\neq j\\ a_{ii}-1, & \text{if } i=j\end{cases}$$

Hence, $\mathbf{B}(\mathbf{U})$ is the interference matrix between users. A system in which all users achieve a common ratio γ_c of the SIR to some minimum target SIR that may depend on the selected transmission rate in the downlink, is described by the set of linear equations

$$\mathbf{B}(\mathbf{U})\cdot\mathbf{p} = \frac{1}{\gamma_c}\cdot\mathbf{p} \tag{7}$$

where $\mathbf{p}$ is the power vector. Thus, γ_c is a reciprocal eigenvalue of $\mathbf{B}(\mathbf{U})$ and is actually the relative SIR that is the ratio of real SIR to the required SIR threshold. If $\gamma_c \geq 1$ is satisfied, the given rate vector can be supported. Matrix $\mathbf{B}(\mathbf{U})$ has the same properties as $\mathbf{A}(\mathbf{U})$ with respect to the existence of an eigenvector $\mathbf{p}$ with positive components. Therefore, we have $1/\gamma_c = \lambda_{max}(\mathbf{B}(\mathbf{U}))$. If γ_c^* is the maximum possible common relative SIR, then

$$\gamma_c^* = \frac{1}{\min\limits_{\mathbf{U}} \lambda_{max}(\mathbf{B}(\mathbf{U}))} \tag{8}$$

Boche and Schubert proposed a downlink beamforming algorithm [2] where the power level and beamforming weights are computed iteratively. This algorithm is optimal in the sense that the convergence to the optimal beamforming weights and power levels are guaranteed if the problem is feasible. However the iterative process could be time consuming and computationally expensive. In this paper, we apply a simple algorithm that calculates the beamforming weights and power levels in one iteration. The pseudo-code of this algorithm is provided below.

ALGORITHM I

- **STEP 1**: Solve a set of N decoupled generalized eigenproblems.

$$\mathbf{u}_j = \arg \max_{\|\mathbf{u}_j\|=1} \frac{\mathbf{u}_j^H \mathcal{H}_j \mathbf{u}_j}{\mathbf{u}_j^H \mathcal{R}_j \mathbf{u}_j}, \quad \text{for all } j \in \mathcal{U}.$$

  where

$$\mathcal{R}_j = \sum_{\substack{k \in \mathcal{U} \\ k \neq j}} \mathcal{H}_k$$

- **STEP 2**: Solve the following eigenproblem

$$\mathbf{B}^T(\mathbf{U}) \cdot \mathbf{p} = \lambda_{max} \cdot \mathbf{p}$$

- **STEP 3**: If $\lambda_{max} \leq 1$, the rate vector is feasible.

Algorithm I is used in our proposed algorithms to check the feasibility of a rate vector in the following subsections.

2) Heuristic downlink scheduling algorithms: Eq. (3) suggests that in order to maintain the stability of the system, the users with large queue lengths should be given a higher service priority. However, these users may not be spatially separable and may not be served together. On the other hand, we may choose a set of compatible users such that the total throughput is maximized (see [4], [5] for examples of such scheduling algorithms). Therefore, there exists a tradeoff between serving the longest queues first and maximizing total instant system throughput. In order to balance the relative importance of serving the users with the largest queue lengths and improving the instant system throughput, we propose a new heuristic algorithm.

First, in order to give higher priorities to the users with the larger queue sizes we start with the user with the longest queue, and try to schedule the users sequentially in the decreasing order of their queue lengths. Each new user is allocated the highest possible rate such that the SIR requirement is satisfied with the new rate vector. However, when we insert users into the channel sequentially according to their queue lengths, it is possible that a user already scheduled for transmission prevents a number of other users from accessing the channel because the necessary spatial separability cannot be provided. Therefore, in order to improve the performance of the system further and maintain the linear complexity, we will consider several candidate rate vectors and select the one that maximizes (3). More specifically, we will consider P rate vectors out of all possible rate vectors. Clearly, this subset of candidate rate vectors should consist of the rate vectors that are more likely to maximize (3).

We explain how we generate this subset of candidate rate vectors to be considered. Suppose that we form a list of users by decreasing queue size. In order to generate the p-th rate vector, $p = 1, \ldots, P$, of the subset, we first move the p-th user in the list to the head of the list. Then, starting from the head of the list, go down the list sequentially and insert one user at a time using the largest rate that is allowed while maintaining the rates and required SIR values of the previously scheduled users. Note that in some cases, a user may need to be skipped because the user may not be compatible with other users already inserted. Once the P rate vectors are generated, out of these rate vectors we select the rate vector that maximizes (3). The pseudo-code of this algorithm is provided below.

ALGORITHM II

- **STEP 1**: Initialize $\mathcal{R} = \emptyset$.
- **STEP 2**: For $p = 1$ to P, do
  - **STEP 1.1**: Form a list $\mathcal{K}$ of users as follows: Insert the user with the p-th largest queue size at the head of the list, and insert the remaining users by decreasing queue size. Initialize the rate vector $\mathbf{r}^p = \underline{0}$.
  - **STEP 1.2**: Schedule the user at the head of the list, denoted by j^*, with the highest rate $r_{j^*}^p$, and add user j^* and its rate to $\mathbf{r}^p$. Remove user j^* from the list $\mathcal{K}$, *i.e.,* set $\mathcal{K} = \mathcal{K} \setminus \{j^*\}$.
  - **STEP 1.3**: If the number of scheduled users with a positive rate is strictly less than M and $\mathcal{K} \neq \emptyset$, go to STEP 1.2. Otherwise, stop and add $\mathbf{r}^p$ to $\mathcal{R}$.
- **STEP 3**: Among the rate vectors in $\mathcal{R}$, select $\mathbf{r_o}$

$$\mathbf{r_o} = \arg \max_{\mathbf{r} \in \mathcal{R}} \sum_{j=1}^{J} r_j x_j(t) \; . \tag{9}$$

The complexity of Algorithm II is given by

$$c_2 = PJV \; .$$

Intuitively, Algorithm II tries to serve the users with larger queue lengths, which is consistent with (3). However, these users may not be compatible with other users and could prevent a large number of other users with smaller queue lengths from accessing the channel, resulting in smaller system throughput.

However, in some cases it may be more desirable to schedule a large number of compatible users, potentially with smaller queue lengths, so as to maximize (3). Hence, this suggests that if the system attempts to maximize the instant system throughput, this may lead to smaller overall queue sizes, and hence in the long run the system may remain stable for a larger set of arrival vectors, potentially at the price of larger delay jitter. In [5], an algorithm that attempts to maximize the total system throughput is presented. The basic idea is to search through all the users and select the user that is most compatible with already scheduled users. However, only single transmission rate is allowed in [5]. Here we extend the algorithm to the multiple transmission rates case, and present the pseudo-code of the proposed algorithm. This algorithm has complexity $c_3 = (JV)^M$ due to the search process.

ALGORITHM III

- **STEP 1**: Initially let $\hat{\mathcal{J}}$ be the set of the users and $\mathcal{J}_S = \emptyset$ the set of scheduled users. Select the user with the largest queue
$$j^* = \arg\max_{j \in \mathcal{J}} x_j(t) \ .$$
Add user j^* to $\mathcal{J}_S$ with the highest rate $v_{j^*}(t)$ that can be accommodated, and remove user j^* from $\hat{\mathcal{J}}$.
- **STEP 2**: For each user $j \in \hat{\mathcal{J}}$ let v_j^* be the highest rate with which a user j can be accommodated in the channel simultaneously with the already scheduled users with their rates unchanged. Define $v^* = \max_{j \in \hat{\mathcal{J}}} v_j^*$.
  - If $v^* = 0$, STOP.
  - Else, let $\mathcal{J}' = \{j \in \hat{\mathcal{J}} \mid v_j^* = v^*\}$.
- **STEP 3**: For each $j \in \mathcal{J}'$ compute the minimum SIR value of the users in $\mathcal{J}_S \cup \{j\}$, denoted by SIR_{min}^j. Add the user $j^* \in \arg\max_{j \in \mathcal{J}'} SIR_{min}^j$ to $\mathcal{J}_S$ with rate v^*, and remove j^* from $\hat{\mathcal{J}}$. Ties are broken arbitrarily.
- **STEP 4**: If the number of scheduled users with a positive rate is strictly less than M and $\hat{\mathcal{J}} \neq \emptyset$, go to STEP 2. Otherwise, stop.

The basic intuition behind Algorithm III is that a user that maximizes the minimum SIR value of the scheduled users will be more likely to allow more users to be scheduled in the following iterations and, hence, increase the overall system throughput. As mentioned earlier in Algorithms II and III above, Algorithm I is used to test whether a rate vector is feasible or not.

B. Multiple cell case

When we have multiple base stations to serve a set of users, compared to the fixed assignment of users to the base stations, it is beneficial to dynamically assign users in order to better match the users to the base stations, especially when the channels are time-varying, and to load balance across the base stations. In this subsection we investigate the performance enhancement achieved by dynamic assignment of the users to the base stations.

We first extend Algorithm I to the multiple base station case. The set of base stations is given by $\mathcal{I}$, and the set of users scheduled on the same channel is denoted by $\mathcal{U}$. We assume that we are given the spatial covariance matrices $\mathcal{H}_j^i$, $i \in \mathcal{I}$ and $j \in \mathcal{U}$. The base station to which user j is assigned is denoted by i_j. We compute the beamforming weights and transmission power for each user in such a way that the common relative SIR for all users is maximized.

ALGORITHM IV

- **STEP 1**: Solve a set of $|\mathcal{U}|$ decoupled generalized eigenproblems.
$$\mathbf{u}_j = \arg\max_{\|\mathbf{u}_j\|=1} \frac{\mathbf{u}_j^H \mathcal{H}_j^{i_j} \mathbf{u}_j}{\mathbf{u}_j^H \mathcal{R}_j^{i_j} \mathbf{u}_j} \ , \quad \text{for all } j \in \mathcal{U}$$
where
$$\mathcal{R}_j^{i_j} = \sum_{\substack{k \in \mathcal{U} \\ k \neq j}} \mathcal{H}_k^{i_j}$$
- **STEP 2**: Solve the following eigenproblem
$$\mathbf{B}^T(\mathbf{U}) \cdot \mathbf{p} = \lambda_{max} \cdot \mathbf{p}$$
where
$$b_{jk} = \begin{cases} 0, & \text{if } j = k \\ \frac{T(v_j)\mathbf{u}_k^H \mathcal{H}_j^{i_k} \mathbf{u}_k}{\mathbf{u}_j^H \mathcal{H}_j^{i_j} \mathbf{u}_j}, & \text{otherwise} \end{cases}$$
where $T(v_j)$ is the required SIR threshold that is a function of transmission rate v_j.
- **STEP 3**: If $\lambda_{max} \leq 1$, the rate vector is feasible.

Algorithm IV will be used to check the feasibility of a rate vector in the following algorithms.

We propose two algorithms for downlink scheduling problem with multiple base stations. In Algorithm V, users are assigned to their respective closest base station, while the user assignment is dynamic in Algorithm VI.

ALGORITHM V

- **STEP 1**: Initialize $\mathcal{J}_S = \emptyset$ and $\hat{\mathcal{J}} = \mathcal{J}$. To each user $j \in \mathcal{J}$, assign a base station i_j closest to user j.
- **STEP 2**: Select user
$$j^* = \arg\max_{j \in \hat{\mathcal{J}}} x_j(t) \ .$$
- **STEP 3**: Schedule user j^* with the highest rate r_{j^*} that can be accommodated, add it to $\mathcal{J}_S$, and remove it from $\hat{\mathcal{J}}$.
- **STEP 4**: If $\hat{\mathcal{J}} \neq \emptyset$, go to STEP 2. Otherwise, stop.

Clearly, we can generate several candidate rate vectors following similar steps used in Algorithm II and select the rate vector that maximizes (3) if so desired.

Algorithm VI described below is different from Algorithm V in the way base stations are assigned to users. When a user is considered for scheduling, the best station is assigned to this user in Algorithm VI.

ALGORITHM VI

- **STEP 1**: Initialize $\mathcal{J}_S = \emptyset$ and $\hat{\mathcal{J}} = \mathcal{J}$.
- **STEP 2**: Select the user

$$j^* = \arg\max_{j \in \hat{\mathcal{J}}} x_j(t) \ .$$

  Assign j^* to the closest base station, and schedule user j^* with the highest rate $r_{j^*}(t)$ that can be accommodated. Add user j^* to $\mathcal{J}_S$ and remove it from $\hat{\mathcal{J}}$.
- **STEP 3**: Choose the user

$$j^* = \arg\max_{j \in \hat{\mathcal{J}}} x_j(t) \ .$$

  For each base station $i \in \mathcal{I}$ compute the largest rate at which user j^* can be served. Assign user j^* to the base station that can schedule user j^* with the highest rate $r_{j^*}(t)$. Add user j^* to $\mathcal{J}_S$ and remove it from $\hat{\mathcal{J}}$.
- **STEP 4**: If $\hat{\mathcal{J}} \neq \emptyset$, go to STEP 3. Otherwise, stop.

IV. SIMULATION RESULTS

In this section, we evaluate the performance of the heuristic algorithms we have proposed in the previous section, using computer simulations. Typical results are presented to illustrate the performance enhancement achieved by jointly considering MAC layer queue state and physical layer spatial compatibility of users when scheduling users.

A. Single cell case

1) Simulation setup: We first consider a single-cell system where a base station transmits packets to $J = 10$ users. The users are angularly uniformly distributed in the cell and the distances of the users to the base station are uniformly distributed between zero and the radius of the cell. The BS is equipped with an antenna array with $M = 4$ elements and the distance between two closest elements is $d = \lambda/2$. The received power decays with distance l from the BS as l^{-4}. For each link corresponding to an antenna and a user receiver, multi-path fading is simulated with a 2-ray model. The angle of the first path θ_1 is uniformly distributed in $[0,\ 2\pi]$, while the angle of the second path θ_2 deviates from θ_1 by a random amount, uniformly distributed in $[0,\ 0.1\pi]$. The complex gain of each path is an independent log-normal random variable with standard deviation $\sigma = 6$ dB, which accounts for shadow fading.

An underlying time-slotted system is assumed. The packet arrivals at the BS in each time slot are i.i.d. Bernoulli random variables with the average rate vector $\mathbf{A} = a \cdot \mathbf{L}$, where $\mathbf{L}$ is a $J \times 1$ vector and a is the coefficient that is the control knob for the system load.

2) Comparative results: In Fig. 2 we show the average packet delay as a function of the system throughput in a single cell system with only one available transmission rate. Here we generate a random vector $\mathbf{L}$, and the system throughput on the x-axis is given by $a \cdot \sum_{j=1}^{J} \mathbf{L}_j$. We observe that for Algorithm II, the delay is almost identical for $P = 1$ and $P = 3$ cases. This means that the performance given by the average packet delay is not sensitive to the number of obtained rate vectors for this scenario. On the other hand, the delay is larger for Algorithm III, and it exhibits a rapid increase in the average delay at a smaller throughput than Algorithm II. This indicates that Algorithm II is able to maintain the system stability for a larger system throughput than Algorithm III.

Fig. 3 shows the performance of Algorithm II and Algorithm III for the same network scenario with the only difference being the availability of multiple transmission rates. The average packet delay is plotted as a function of system throughput. Either one or two packets are transmitted in one time slot when low or high transmission rate is chosen, respectively. We observe that Algorithm III performs better than Algorithm II with different values of P when the system throughput is low, and performs slightly worse than Algorithm II with $P = 3$ when the throughput is high. Algorithm II tries to balance the queue lengths of different users. Algorithm III, on the other hand, tries to assign a user that is most compatible with the users already assigned, and as a result Algorithm III can serve more users with higher sum transmission rate in each time slot and the queue lengths tend to be more unbalanced. When the throughput is low, the queue lengths are small on the average. In such situations Algorithm III can achieve smaller packet delay than Algorithm II because the uneven queue lengths can allow the users with enough packets backlogged in the queue to take advantage of higher transmission rates when possible and on the average a large number of packets are served in each time slot. Under Algorithm II that attempts to maintain even queue lengths, when the throughput is low the users may not have enough backlogged packets to take full advantage of high transmission rates. But, when the system throughput is high, Algorithm III performs slightly worse than Algorithm II with $P = 3$ because typically many users have enough packets to utilize high transmission rates when the throughput is high and Algorithm III balances the queues. However, multiple transmission rates allow Algorithm III to perform better than algorithm II with $P = 1$ since it can achieve higher instant sum transmission rates than Algorithm II due to the availability of only single candidate rate vector under Algorithm II.

Moreover, comparing Fig. 2 and 3 we can observe that, for a fixed delay, having multiple transmission rates increases the schedulable throughput region by about 70 percent. This is due to the better use of the transmission bandwidth achieved by multiple transmission rates. Although this gain is limited to our simulated scenario, this suggests much benefit may be gained through the use of multiple transmission rates through link adaptation.

B. Multiple cell case

1) Simulation setup: We consider a square area which is divided into four equal square cells. One BS is located in the

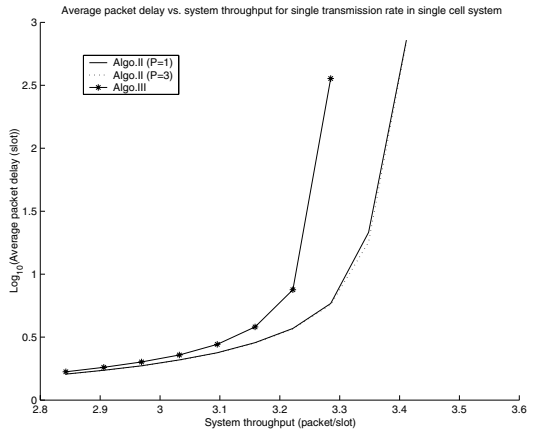

Fig. 2. Delay vs. throughput for single rate communication in single cell system

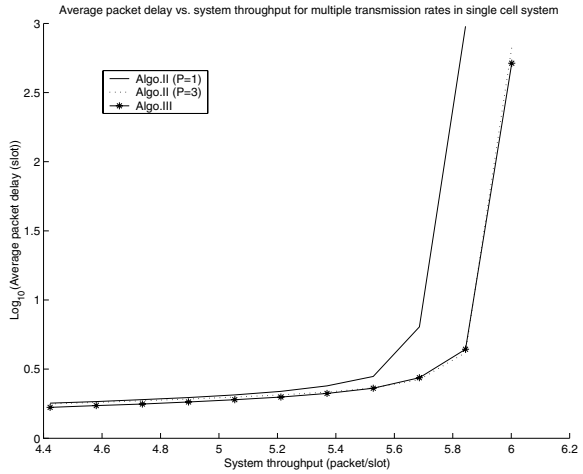

Fig. 3. Delay vs. throughput for multiple rate communication in single cell system

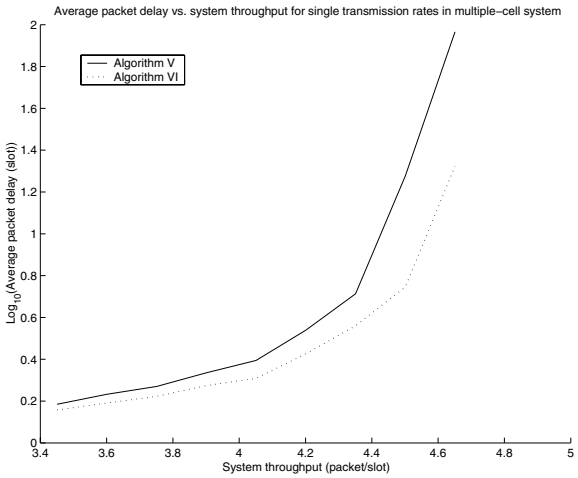

Fig. 4. Delay vs. throughput for single rate communication in multiple cell system

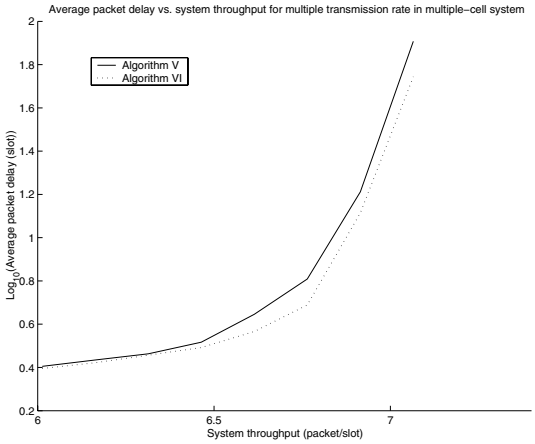

Fig. 5. Delay vs. throughput for multiple rate communication in multiple cell system

center of each cell. In the simulation 20 users are uniformly distributed in the square area. The links between a user and a BS is modeled as in the single cell system in the previous subsection, except that the angle of the first path with respect to a BS is determined by the relative location of the user and the BS. We assume that the four BSs are controlled by a single central controller. Packets arrive at the central controller according to i.i.d. Bernoulli rvs with the average rate $\mathbf{A} = a \cdot \mathbf{L}$ as in the single cell case.

2) Comparative results: In Fig. 4 we show the average packet delay for Algorithms V and VI for the multiple cell system where only single transmission rate is allowed. We observe that Algorithm VI achieves lower delay than algorithm V for different system throughput because dynamic base station assignment is able to allow more users to be served in each time slot by balancing the transmission load across different base stations. Similarly, we plot the average packet delay for Algorithms V and VI for multiple transmission rate case in Fig. 5. Again we observe that Algorithm VI performs better than Algorithm V for various system throughput. Moreover, by comparing Fig. 4 and 5, we observe that multiple transmission rates improve the maximum system throughput that can be supported for a fixed delay by about 50 percent.

V. Discussion

The use of antenna arrays at the base stations has been proposed in the past to improve the system throughput and provide QoS guarantees to mobile users in wireless networks. In this paper we studied the problem of wireless scheduling with antenna arrays to provide QoS guarantees in terms of throughput. We derived the optimal scheduling policy that results in maximum throughput region determined by the spatial separability of users. Due to the inherent difficulty in finding the optimal solution, heuristic algorithms must be adopted, which capture desired properties of the optimal solution.

We presented four algorithms for joint scheduling, beamforming and power control. The first two algorithms are proposed for single cell systems, while the last two are designed for multiple cell systems. The intuition behind these heuristic algorithms is to approximate the optimal scheduling algorithm with lower computational complexity. Simulation results indicate that this joint consideration of MAC layer scheduling algorithm and physical layer beamforming and power control yields significant performance improvement compared to the algorithms

that maximize instant system throughput as proposed in the literature.

REFERENCES

[1] F. Rashid-Farrokhi, K. R. Liu and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE Journal on Selected Areas in Communications,* vol.16, no.10, pp.1437-1450, October 1998

[2] H. Boche and M. Schubert, "SIR balancing for multiuser downlink beamforming: a convergence analysis," *IEEE ICC,* New York, NY, April 2002

[3] M. Bengtsson, "Jointly optimal downlink beamforming and base station assignment," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Salt Lake City, Ut, May 2001

[4] F. Shad, T.D. Todd, V. Kezys and J. Litva, "Dynamic slot allocation (DSA) in indoor SDMA/TDMA using a smart antenna basestation," *IEEE/ACM Transactions on Networking,* vol.9, no.1, pp.69-81, February 2001.

[5] I. Koutsopoulos, T. Ren and L. Tassiulas, "The impact of space division multiplexing on resource allocation: a unified approach," *IEEE INFOCOM*, San Francisco, CA, April 2003

[6] L. Tassiulas, A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control,* vol. 37, no. 12, pp.1936-1949, December 1992.

[7] L. Tassiulas, "Scheduling and performance limits of networks with constantly changing topology," *IEEE Transactions on Information Theory,* vol.43, no.3, pp.1067-1073, May 1997

[8] M. Neely, E. Modiano and C. Rohrs, "Power and server allocation in a multi-beam satellite with time varying channels," *IEEE INFOCOM,* New York, NY, June, 2002

[9] P. R. Kumar and S. P. Meyn, "Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies," *IEEE Transactions of Automatic Control*, vol.41, pp4-17, January 1996

[10] S. Asmussen, *Applied Probability and Queues*, Wiley 1987

[11] C. Farsakh and J. Nossek, "Spatial covariance based downlink beamforming in an SDMA mobile radio system," *IEEE Transactions on Communications,* vol.46, no.11, pp.1497-1506, November 1998