Optimal transmission scheduling with base station antenna array in cellular networks

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

***Abstract*— We study the downlink scheduling problem in a cellular wireless network. The base stations are equipped with antenna arrays such that a base station can transmit to more than one mobile user at each time instant because of the spatial filtering capability of antenna array. A number of users can receive packets correctly provided they 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 throughput. In this paper, we consider finite user traffic demands. A scheduling policy depends 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 optimal scheduling policy that achieves the stability of the system if it is achievable by 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 realistic multi-path wireless channel model, we evaluate the performance of these algorithms through computer simulations. The results show the benefits of consideration of queue length and dynamic base station assignment.**

I. Introduction

Wireless communication has been experiencing rapid development during the last decade. The increasing need for providing fast wireless access and high-speed wireless links to users has become the driving force for active research in the telecommunications area. At present, wireless communication is undergoing the transition from conventional 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 applications such as wireless multimedia transmission, wireless Internet access and video conferencing is only the first sign of the projected demand for rapid and reliable wireless data access.

New network structures and protocols are proposed to support data applications in wireless networks. In such systems, the networks are in a cellular structure where the final interface between the mobile user and the network is wireless through access points (APs) or base stations (BSs) that are wired to the backbone network. For instance, 3G protocols have been standardized and are being implemented to provide mobile users with wireless data access. The most challenging goal in the design of 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 in different communication layers. QoS in physical layer is expressed as an acceptable signal to interference and noise ratio (SINR) or corresponding bit error rate (BER) at the receiver. In the MAC layer, QoS is usually expressed in terms of achievable bit rate or packet error rate (PER), while at higher layers QoS can be perceived as a minimum throughput or maximum delay requirement. The ability of the network infrastructure to fulfill QoS requirements and ultimately enhance system capacity depends on procedures in several layers.

A wide spectrum of approaches are proposed to reuse the communication resources in time, frequency or space domain, to provide 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 the most promising one and the last frontier for future capacity improvement of wireless networks. This is because of the beamforming capability of the antenna arrays that can form the beam pattern directed to the desired user while nulling the others. In this way, co-channel interference can be greatly compressed and spatial 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, minimizing the total transmission power subject to the constraints that each user's SINR requirement 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 above 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. Algorithm that assigns each user to the optimal base station and computes the corresponding transmit beam pattern for each user is designed.

The second class of research is on the MAC layer with consideration of physical layer user separability constraints. We have a set of users and we would like to put as many as possible users into one channel and compute the beamforming weights for each selected user under the constraints that the SINR value of each selected user is above a threshold. In this way, the throughput of the network is maximized. The channel can be a time slot in a TDMA system, a subcarrier in an OFDM system or a code in a CDMA system. Algorithms aiming at maximizing total throughput are proposed in the literature [4]. These algorithms are based on the same idea of insertion of users into a channel sequentially, and vary in the criteria that determine the order in which users are inserted. This problem is extended to be combined with other multiplexing schemes such as TDMA, OFDM and CDMA in [5]. A common assumption in these works is infinite packet backlog for any user. The major drawbacks of these works are the limitation of the focus on instant total throughput maximization and lack of consideration of upper layer QoS requirement of each individual user. Thus, the assignment of users in each channel only reflects the feasibility in the physical layer, 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 long term system performance. Therefore higher layer QoS requirement has to be taken into consideration for 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 scheduling problem at a number of BS's controlled by one central controller and each BS is equipped with antenna array. Packets arrive at the central controller for transmission to different mobile users. Buffer occupancy and thus traffic demand of each user are considered explicitly. In addition to feasibility of users sharing the same channel, the scheduling policies depend on 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. SINR requirement constraints are imposed on the selection of users that can be served in each time slot. Instead of maximizing instant throughput, we seek for optimal scheduling policy that stabilizes the system if it is stabilizable. Specifically, under this optimal scheduling policy the user throughput requirements are satisfied and thus the long term total system throughput is maximized.

Similar queueing system is used to model other scenarios in [6][7][8] and is first considered in [6] for a multihop radio network where the SINR requirement demands that two links can be active simultaneously only if they are separate for at least a 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 which stabilizes a system whenever it is stabilizable is identified. In this paper, we follow the same direction as in [8]. However, the optimal scheduling policy is exponentially complex in the number of users and no practical sub-optimal scheduling policy is proposed in [6] [7] [8]. For our problem, we will propose scheduling policies of polynomial complexity that achieve sub-optimal performance.

This paper is organized as follows. In Section II, we 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. Performance evaluation of these algorithms are presented in Section IV. Section V concludes the paper with discussions.

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

A. System model

We consider a wireless network which consists of several base stations. Each base station is equipped with an antenna array such that several users can be served simultaneously. These base stations are coordinated by a sin-

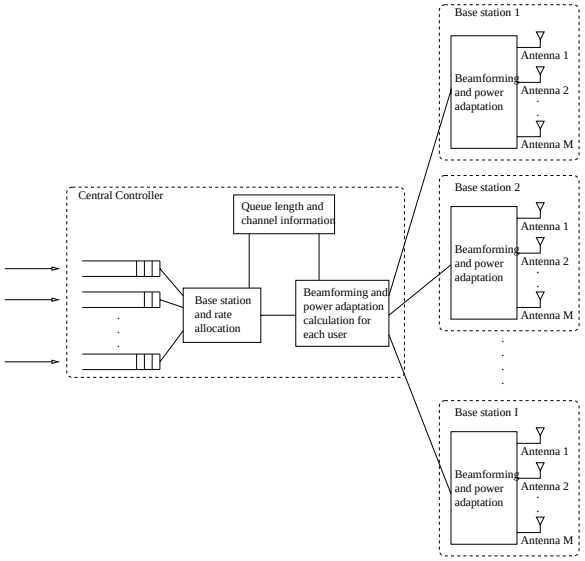

Fig. 1. The multiple cellular communication system

gle central controller. Mobile users in the network are able to receive data packets from any of these base stations. However, at each instant, one mobile user can receive data packet from only one base station. The central controller maintains a separate queue for incoming data packets destined to each mobile user. We assume a time slotted system where the transmission time of each packet equals to one time slot if lowest transmission rate is applied. In each time slot, the central controller collects the information of the wireless links of each user to different base stations. Based on this information and the packet backlog condition, the central controller assigns base stations to the users with respective transmission rates and calculate the beamforming weights which 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 allocations 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. A beam is formed by a dedicated transceiver and a power is assigned to a 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

A central controller coordinates the operation of I base stations. Each base station is equipped with an M-element antenna array. J users receive data packets from these base stations. We denote $\mathcal{I}$ and $\mathcal{J}$ as the sets of base stations and users respectively.

Several rates can be applied for the transmission to a user. We denote $\mathcal{V}$ as the set of available rates. We assume each rate is a positive integer number. If rate $v \in \mathcal{V}$ is applied, v packets can be transmitted in one time slot. We denote $|\mathcal{V}| = V$.

Packets arrive at the central controller for transmission to different users. The central controller maintains a separate queue for each user. Let $a_j(t)$ denote the number of packets that arrive at queue j in time slot t. $a_j(t)$ is an i.i.d. random variable with finite second moment distribution, $E[a_j(t)^2] < \infty$, for $j = 1, 2, \cdots, J$. We denote the number of backlog packets for user j at the start of time slot t as $x_j(t)$.

We assume the arrival process is ergodic and time invariant, such that

$$A_j = E[a_j(t)] = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} a_j(\tau), \qquad (1)$$

And we call $\mathbf{A} = (\mathrm{A}_1, \mathrm{A}_2, \cdots, \mathrm{A}_\mathrm{J})^\mathrm{T}$ an arrival vector.

We assume the central controller has perfect channel information for each user with regard to every base station. At each time slot, the central controller assigns the base stations to the users with respective rates. The calculations of beamforming weights for each base station are also performed at the central controller. A scheduling decision can be expressed as an $I \times J$ matrix $\mathbf{R}$ where element r_{ij} 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 evolution process as a Markov chain with stationary distribution π. Each channel state is represented by the set of all the feasible rate matrices in this state. Let $\mathcal{S}$ be the channel state space. Our problem is to find the optimal scheduling policy which selects the feasible rate matrix in each time slot given the queue lengths, such that the system achieves maximum throughput. We define and characterize the throughput region in the following subsection.

C. Throughput region

Definition: an arrival vector $\mathbf{A}$ is within throughput region $\mathcal{A}$ if there exists a scheduling policy such that

$$\lim_{c\to\infty} Pr(\frac{\sum_{\tau=1}^{t}\mathbf{1}(x_j(\tau)>c)}{t}) = 0, \text{ for } j = 1, 2, \cdots, J \quad (2)$$

where $x_j(\tau)$ is the number of backlog packets in queue j at the start of time slot τ. We say $\mathbf{A}$ is stable under this scheduling policy.

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

Proposition 1: for an arrival vector $\mathbf{A}$, the necessary and sufficient condition for $\mathbf{A} \in \mathcal{A}$ is: there exists a scheduling policy that achieves

$$\mathbf{A} \leq \mathbf{D} = \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} c_{\mathbf{SR}}\mathbf{R}^T\mathbf{1}_{I\times 1} \quad (3)$$

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

proof: [8].

D. Optimal scheduling policy

We are interested in the optimal scheduling policy that achieves stability for each $\mathbf{A} \in \mathcal{A}$. Specifically, we consider the following scheduling policy given backlog vector $\mathbf{X(t)}$ and system channel state $\mathbf{S(t)}$.

$$\mathbf{R(t)} = arg\ max_{\mathbf{R}\in\mathbf{S(t)}}\mathbf{X(t)^T}(\mathbf{R^T}\mathbf{1}_{I\times 1}) \quad (4)$$

where tie is broken arbitrarily.

The backlog process $\mathbf{X(t)}$ is an J-dimensional Markov process with infinite and countable state space given that the scheduling policy is stationary. Define Lyapunov function

$$L(\mathbf{X(t)}) = \sum_{\mathbf{j=1}}^{\mathbf{J}} \mathbf{x_j(t)^2} \quad (5)$$

Through the negative drift of Lyapunov function when backlog is large, we can prove the existence of the steady distribution of $\mathbf{X(t)}$ and hence the stability of the system. Formally, we rely on the following theorem to establish stability property.

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

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

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

Now we prove the following proposition which establishes the optimality of scheduling policy (4).

Proposition 2: scheduling policy (4) stabilizes the system if the arrival vector $\mathbf{A}$ is interior to the throughput region.

Proof: the one step drift of the backlog vector $\mathbf{X(t)}$ is

$$\mathbf{X(t+1)} = max(\mathbf{X(t)} + \mathbf{A(t)} - \mathbf{D(t)}, 0) \quad (6)$$

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

$$x_j^2(t+1) \leq (x_j(t) + a_j(t) - d_j(t))^2$$

$$\leq x_j(t)^2 - 2x_j(t)d_j(t) + 2x_j(t)a_j(t) + d_j(t)^2 + a_j(t)^2$$

$$E[L(X(t+1)) - L(X(t))|X(t)] \leq$$

$$\sum E[a_j(t)^2|X(t)] + \sum E[d_j(t)^2|X(t)]$$

$$-2\sum x_j(t)E[d_j(t) - a_j(t)|X(t)] \leq$$

$$B - 2\sum x_j(t)(E[d_j(t)|X(t)] - A_j(t))$$

where $B = \sum E[a_j(t)^2|X(t)] + Jv_M^2$ since $\max \sum E[d_j(t)^2|X(t)] \leq Jv_M^2$, where $v_M = \max_{v\in\mathcal{V}} v$.

Since $\mathbf{A}$ lies within the throughput region, we have

$$\sum_{j=1}^{J} x_j(t)A_j \leq \sum_{j=1}^{J} x_j(t) \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} c_{\mathbf{SR}}\mathbf{R}^T\mathbf{1}_{I\times 1}$$

$$= \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} c_{\mathbf{SR}} \sum_{j=1}^{J} x_j(t)r_j$$

$$\leq \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \max_{\mathbf{R}\in\mathbf{S}} \sum_{i=1}^{N} (x_j(t)r_j)$$

$$= \sum_{j=1}^{J} x_j(t)E[d_j(t)|X(t)]$$

where r_j is the j's element of vector $\mathbf{R}^T\mathbf{1}_{I\times 1}$.

We are able to find a vector ε=$(\epsilon, \epsilon, \cdots, \epsilon)$ such that $\mathbf{A} + \varepsilon$ is also within the throughput region and satisfies:

$$\sum x_j(t)(A_j + \epsilon) \leq \sum x_j(t)E[d_j(t)|X(t)]$$

Therefore

$$\sum_{j=1}^{J} x_j(t)(E[d_j(t)|X(t)] - A_j)$$

$$= \sum_{j=1}^{J} x_j(t)(E[d_j(t)|X(t)] - (A_j + \varepsilon) + \varepsilon)$$

$$\geq \epsilon \sum_{j=1}^{J} x_j(t)$$

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 define the compact region

$$\Sigma = \{\mathbf{X}(\mathbf{t}) \in \Re^J | \sum x_j(t) \leq (\frac{B+\alpha}{2\epsilon})\}$$

such that condition 2 in Theorem 1 is satisfied.

III. Heuristic downlink scheduling algorithms with base station antennas

We have derived the optimal scheduling policy. The next question is how to find the optimal rate allocation and base station assignment given the queue length of every user in each time slot. If the user channels are constant, the central controller can exhaustively search for all possible combinations offline, and select the one that maximizes (4) in each slot. However if the channels vary with time, this exhaustive search is not implementable even for a single cell system because the number of possible rate vectors is

$$c_1 = \sum_{j=1}^{J} \left(\begin{array}{c} J \\ j \end{array} \right) V^j \tag{7}$$

which is exponential in the number of users. Therefore, we want to design online scheduling policies which give sub-optimal performance with lower complexity.

In the previous sections, we abstract the spatial separability of users as feasible rate vectors. This abstraction hides the physical channel characteristics. In the following, we will present the physical channel model we adopt. Based on this model, we will propose joint scheduling and beamforming, power control algorithms with polynomial complexity in the number of users and study their performances in terms of average packet delay. We start with single cell system followed by multiple cell system.

A. Single cell system

1) Physical channel model and downlink beamforming algorithm: The first step is to test whether a set of users with their respective rates can be served at the same time. We will describe the adopted physical channel model [11] followed by the introduction of the downlink beamforming algorithm.

The multi-path channel between antenna m and user j is

$$h_j^m(t) = \sum_{\ell=1}^{L} \beta_{j,\ell}\, \delta\left(t - \tau_{j,\ell} + \tau_{j,\ell}^m\right), \tag{8}$$

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 sequel, we assume that the major limitation is cochannel interference rather than noise, so that SINR is approximated by SIR.

The received signal at the receiver of user j is,

$$y_j(t) = \sum_{k \in \mathcal{U}} \sqrt{P_k} \sum_{m=1}^{M} u_k^m \sum_{\ell=1}^{L} \beta_{j,\ell}(\omega_0) e^{j\omega_0 \tau_{j,\ell}^m} s_k(t - \tau_{j,\ell}), \tag{9}$$

where $\mathcal{U}$ is the cochannel set of users. 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 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 received power in channel of user j due to k is

$$E\left\{ |\sqrt{P_k} \sum_{m=1}^{M} u_k^m \sum_{\ell=1}^{L} \beta_{j,\ell}(\omega_0) e^{j\omega_0 \tau_{j,\ell}^m} s_k(t - \tau_{j,\ell})|^2 \right\} \tag{10}$$

$$= P_k \mathbf{u_k}^H \Big(\sum_{\ell_1=1}^{L} \sum_{\ell_2=1}^{L} \mathbf{v}(\theta_{j,\ell_1}) v^H(\theta_{j,\ell_2}) E\{\beta_{j,\ell_1}(\omega_0)\beta_{j,\ell_2}^*(\omega_0)\}$$

$$E\{s_k(t - \tau_{j,\ell_1}) s_k^*(t - \tau_{j,\ell_2})\}\Big)\mathbf{u_k} \tag{11}$$

$$= P_k(\mathbf{u}_k{}^H \mathcal{H}_j \mathbf{u}_k) \tag{12}$$

Observe that

$$E\left\{\beta_{j,\ell_1}(\omega_0)\beta_{j,\ell_2}^*(\omega_0)\right\} = \left\{ \begin{array}{ll} 0, & \text{if } \ell_1 \neq \ell_2 \\ A_{j,\ell}, & \text{if } \ell_1 = \ell_2 = \ell, \end{array} \right. \tag{13}$$

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

$$\mathcal{H}_j = \sum_{\ell=1}^{L} A_{j,\ell}\, \mathbf{v}(\theta_{j,\ell}) \mathbf{v}^H(\theta_{j,\ell}). \tag{14}$$

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, W_j is,

$$W_j = \frac{P_j \left(\mathbf{u}_j^H \mathcal{H}_j \mathbf{u}_j\right)}{\sum\limits_{\substack{k\in\mathcal{U}\\ k\neq j}} P_k \left(\mathbf{u}_k^H \mathcal{H}_j \mathbf{u}_k\right)}. \tag{15}$$

Consider the system of N users. Define as $\mathbf{U}$ the ensemble of computed beamforming vectors for all users, i.e, $\mathbf{U} = \{\mathbf{u}_j : j \in \mathcal{U}\}$. Then, we define the $(|\mathcal{U}|) \times |\mathcal{U}|)$ matrix $\mathbf{A}(\mathbf{U})$ where a_{ij} specifies the cochannel interference caused by the j-th to the i-th user, normalized by the useful signal power of i. That is,

$$a_{ij} = \begin{cases} 1, & \text{if } i = j \\ \frac{T(v_i)\mathbf{u}_j^H \mathcal{H}_i \mathbf{u}_j}{\mathbf{u}_i^H \mathcal{H}_i \mathbf{u}_i}, & \text{otherwise.} \end{cases} \tag{16}$$

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 $\mathbf{A}(\mathbf{U})$, $\lambda_{max}(\mathbf{A}(\mathbf{U}))$.

We introduce matrix $\mathbf{B}(\mathbf{U})$, whose elements are related to those of $\mathbf{A}(\mathbf{U})$ as follows,

$$b_{ij} = \begin{cases} a_{ij}, & \text{if } i \neq j \\ a_{ii} - 1, & \text{if } i = j, \end{cases} \tag{17}$$

Hence, $\mathbf{B}(\mathbf{U})$ is the interference matrix between users. A system in which all users achieve a common SIR γ_c in the downlink is described by the set of linear equations

$$\mathbf{B}(\mathbf{U}) \cdot \mathbf{p} = \frac{1}{\gamma_c} \cdot \mathbf{p} \tag{18}$$

where $\mathbf{p}$ is the power vector. Thus, γ_c is a reciprocal eigenvalue of $\mathbf{B}(\mathbf{U})$ and is actually 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 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 SIR then

$$\gamma_c^* = \frac{1}{\min\limits_{\mathbf{U}} \lambda_{max}(\mathbf{B}(\mathbf{U}))} \tag{19}$$

Holger and coworkers proposed a downlink beamforming algorithm [2] where the power level and beamforming weights are computed iteratively. This algorithm is optimal in the sense that convergence to the optimal beamforming weights and power levels are guaranteed if the problem is feasible. However the iterative porcess could be time consuming. In this paper, we apply a simple algorithm that calculates the beamforming weights and power levels only once. The pseudo-code of this algorithm is as follows.

ALGORITHM I

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

$$\mathbf{u}_j = \arg \max_{\|\mathbf{u}_j\|=1} \frac{\mathbf{u}_j^H \mathcal{H}_j \mathbf{u}_j}{\mathbf{u}_j^H \mathcal{R}_j \mathbf{u}_j}, \ \forall j \in \mathcal{U}. \tag{20}$$

  where

$$\mathcal{R}_j = \sum_{\substack{k\in\mathcal{U}\\ k\neq j}} \mathcal{H}_k \tag{21}$$

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

$$\mathbf{B}^T(\mathbf{U}) \cdot \mathbf{p} = \lambda_{max} \cdot \mathbf{p}. \tag{22}$$

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

2) Heuristic downlink scheduling algorithms: According to (4), to maintain the stability of the system, the users with large queue length should be given high service priority. However, these users may not be spatially separable such that they can not be served together. On the other hand, we may choose a set of compatible users such that the total throughput is maximized. Actually, this kind of algorithms have already been proposed in [4], [5]. Based on the relative priority of queue length or throughput, we propose the following algorithms. We can start with the user with longest queue, and try to schedule users sequentially in the decreasing order of their queue lengths. Each new user is allocated the highest possible rate such that SIR requirement is satisfied for the new rate vector.

When we insert users into the channel sequentially according to their queue lengths, it is possible that one inserted user prevents a number of other users from accessing the channel. To further improve performance and keep the complexity linear, we will consider several rate vectors and select the one that optimizes (4). Specifically, we will consider P out of all possible rate vectors. These P rate vectors form a subset of the set of all rate vectors. We expect this subset to consist of the most important rate vectors in terms of maximizing (4).

Therefore, we start to form the p-th activation set with the user of the pth longest queue. Then we are able to select the rate vector that maximizes (4) out of these P rate vectors. Denote by $\mathcal{J}$ the set of users. Let $J = |\mathcal{J}|$. Let $x_j(t)$ be the queue length of user j at time t. The pseudo-code of the algorithm is as follows.

ALGORITHM II

- **STEP 1**: For $p = 1$ to P do
  - **STEP 1.1**: Initialize $\mathcal{J}$ as the set containing all users. Assume j^* is the user with pth longest queue, schedule user j^* with the highest rate $r_{j^*}^p$, $\mathcal{K} = \mathcal{J} \setminus \{j^*\}$.
  - **STEP 1.2**: Select user
$$j^* = arg \max_{j \in \mathcal{K}} x_j(t)$$
  - **STEP 1.3**: Schedule user j^* with the highest rate $r_{j^*}^p$ that can be accommodated, remove user j^* from $\mathcal{K}$.
  - **STEP 1.4**: If the number of scheduled users with positive rates is less than M and $|\mathcal{K}| > 0$, go to **STEP 1.2**
- **STEP 2**: Among all the obtained rate vectors, select $\mathbf{r_o}$ as
$$\mathbf{r_o} = \mathrm{a}rg\ max_{\mathbf{r} \in \mathbf{S}'} \sum_{j=1}^{J} r_j x_j(t) \qquad (23)$$
where $\mathbf{S}'$ is the set of all rate vectors obtained.

The complexity of Algorithm II is

$$c_2 = PJMV \qquad (24)$$

Intuitively, Algorithms III tries to serve the users with larger queue lengths which is consistent with (4). 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. On the other hand, if a large number of compatible users with smaller queue lengths are assigned to the channel with their SIR requirements satisfied, a larger value of (4) can be obtained. Therefore, there exists the tradeoff between serving-long-queues-first policy (Algorithms III) and maximizing-total-throughput policy. In [5], algorithm that implements the maximizing-total-throughput policy is presented. The basic idea is to search through all the users and the user that is most compatible with already scheduled users is selected. In that paper, only single transmission rate is allowed. We extend the algorithm to the multiple transmission rate case and the pseudo-code of this algorithm is presented as follows. This algorithm has complexity $c_3 = (JV)^M$ due to the search process.

ALGORITHM III

- **STEP 1**: Select user
$$j^* = arg \max_{j \in \mathcal{J}} x_j(t)$$
Schedule user j^* with the highest rate $v_{j^*}(t)$ that can be accommodated, remove user j^* from $\mathcal{J}$.
- **STEP 2**: Let v^* be the highest rate with which a user $j \in \mathcal{J}$ can be accommodated in the channel with the rates of the already assigned users unchanged.
If $v^* = 0$, STOP.
Else, let $\mathcal{J}'$ be the set of users that can be assigned to the channel with rate v^*.
- **STEP 3**: Schedule user $j^* \in \mathcal{J}'$ that results in maximal minimum SIR for the set of already scheduled users plus the user under test with corresponding rates. Schedule userj^* with rate v^*. Remove j^* from $\mathcal{J}$.
- **STEP 4**: If the number of scheduled users with positive rates is less than M and $|\mathcal{J}| > 0$, go to **STEP 2**

For the above Algorithms II and III, Algorithm I is called for testing whether a rate vector is acceptable.

B. Multiple cell case

When we have I base stations to serve J users. It is beneficial to dynamically assign users to base stations in comparison with the fixed assignment of users. In the following, we study the performance enhancement achieved by dynamical assignment.

We first extend Algorithm I for the multiple base station case. We are given the set of base stations $\mathcal{I}$, the set of users $\mathcal{J}$, spatial covariance matrices $\mathcal{H}_j^i$ for $\forall i \in \mathcal{I}$ and $\forall j \in \mathcal{J}$ and user j is assigned to base station i_j, we try to calculate the beamforming weights and transmission power for each user such that the common SIR for all users is maximized.

ALGORITHM IV

- **STEP 1**: Solve a set of J decoupled generalized eigenproblems.
$$\mathbf{u}_j = \arg \max_{\|\mathbf{u}_j\|=1} \frac{\mathbf{u}_j^H \mathcal{H}_j^{i_j} \mathbf{u}_j}{\mathbf{u}_j^H \mathcal{R}_j^{i_j} \mathbf{u}_j},\ \forall j \in \mathcal{J}. \qquad (25)$$

where

$$\mathcal{R}_j^{i_j} = \sum_{\substack{k \in \mathcal{J} \\ k \neq j}} \mathcal{H}_k^{i_j} \tag{26}$$

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

$$\mathbf{B}^T(\mathbf{U}) \cdot \mathbf{p} = \lambda_{max} \cdot \mathbf{p} \tag{27}$$

where

$$b_{ij} = \begin{cases} 0, & \text{if } i = j \\ \frac{T(v_i)\mathbf{u}_j^H \mathcal{H}_i^{i_j} \mathbf{u}_j}{\mathbf{u}_i^H \mathcal{H}_i^{i_i} \mathbf{u}_i}, & \text{otherwise.} \end{cases}$$

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

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 user assignment is dynamic in algorithm VI.

ALGORITHM V

- **STEP 1**: To each user $j \in \mathcal{J}$, assign base station i_j that is the closest one to user j.
- **STEP 2**: Select user

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

- **STEP 3**: Schedule user j^* with the highest rate r_{j^*} that can be accommodated, remove user j^* from $\mathcal{J}$.
- **STEP 4**: If $|\mathcal{J}| > 0$, go to **STEP 2**, else STOP.

Algorithm VI presented in the following is different with 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**: Select user

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

Assign j to the closest base station and schedule user j^* with the highest rate $r_{j^*}(t)$ that can be accommodated, remove user j^* from $\mathcal{J}$.

- **STEP 2**: Select user

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

Assign j to base station that schedules user j^* with the highest rate $r_{j^*}(t)$ that can be accommodated, remove user j^* from $\mathcal{J}$.

- **STEP 3**: If $|\mathcal{J}| > 0$, go to **STEP 2**, else STOP.

IV. SIMULATION RESULTS

In this section, we evaluate the performance of the heuristic algorithms we proposed using computer simulations. Typical results are presented to illustrate the performance enhancement achieved by jointly considering MAC layer queueing 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 0 and the radius of the cell. The BS is equipped with an antenna array with $M = 4$ elements and $d = \lambda/2$. 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 numbers of packets that arrive at the BS in each time slot are identically and independently distributed random variables with Bernoulli distribution of 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 to the system load.

2) Comparative results: In fig.2, we show the average packet delay as a function of the system throughput in the single transmission rate scenario in single cell system. We observe that for algorithm II, the delay is almost identical for $P = 1$ and $P = 3$ cases. That means the performance in terms of packet delay is not sensitive with the number of obtained rate vectors for this scenario. On the other hand, the delay is larger for algorithm III, and the delay becomes large for 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.

For the same network scenario as for fig.2, we present the performance of algorithm II and algorithm III for a multiple-rate case in fig.3, where the average packet delay is shown as a function of system throughput. Either one or two packets are transmitted in one time slot when low or high transmission rate is applied respectively. We observe that algorithm III performs better than algorithm II with different values of P when the system throughput is low,

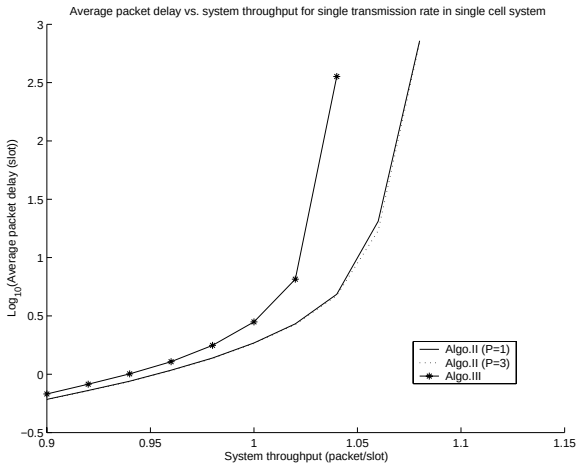

Fig. 2. Delay vs. throughput for single rate communication in single cell system

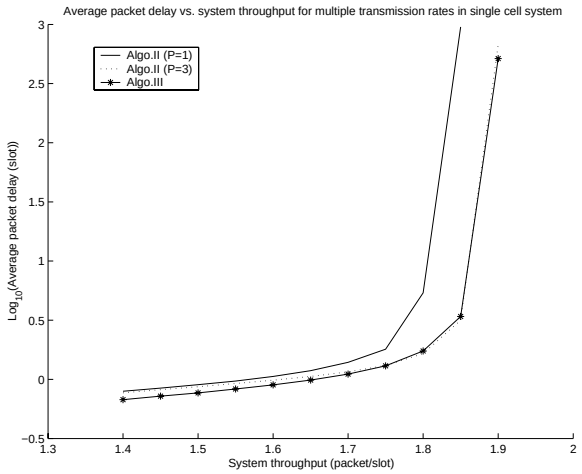

Fig. 3. Delay vs. throughput for multiple rate communication in single cell system

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. On the contrary, algorithm III tries to assign a user that is most compatible to the users already assigned, thus algorithm III can assign 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, algorithm III can achieve smaller packet delay than algorithm II, because the uneven queue lengths could be beneficial since multiple rates are possible and more users can be served with higher transmission rates compared to algorithm II where the even queue lengths make the users lack packets to receive high transmission rate. However, when the system throughput is high, algorithm III performs slightly worse than algorithm II with $P = 3$, because users have enough packets to receive high transmission rate when the throughput is high and algorithm III balances the queues. However, multiple transmission rates make algorithm III perform better than algorithm II with $P = 1$, because the instant sum transmission rates can be much higher than algorithm II with $P = 1$.

Moreover, comparing fig.2 and fig.3, we can observe that multiple transmission rates enlarge the system throughput region by about 70%. This is due to the better use of the transmission bandwidth achieved by multiple transmission rates.

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 center of each cell. $J = 20$ users are uniformly distributed in the square area. The links between each user and each 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 an i.i.d. Bernoulli process with the average rate being presented by $\mathbf{A} = a \cdot \mathbf{L}$ as in the single cell case.

2) Comparative results: In fig.4, we show the average packet delay for algorithm V and algorithm 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 assign more users in each time slot by balancing the transmission load across different base stations. Similarly, we show the average packet delay for algorithm V and algorithm VI for multiple transmission rate case in the multiple cell system in fig.5. Again we observe that algorithm VI performs better than algorithm V for different system throughput. Moreover, by comparing fig.4 and fig.5, we observe that multiple transmission rates improve the maximum system throughput by about 80%.

V. Discussion

We investigated the impact of antenna array on scheduling algorithms in order to increase system rate and provide QoS to users in the form of guaranteed 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 by adopted, which capture desired properties of a good solution. In

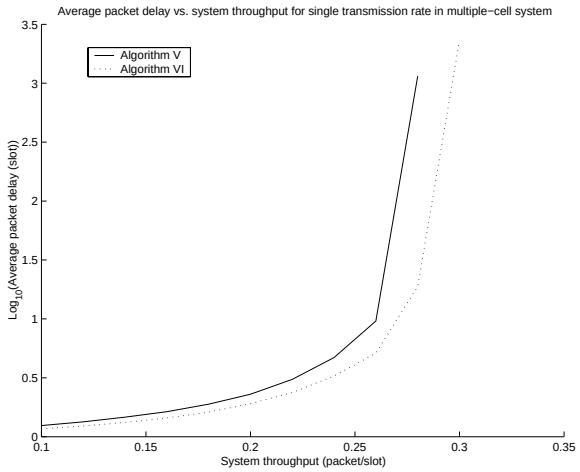

Fig. 4. Delay vs. throughput for single rate communication in multiple cell system

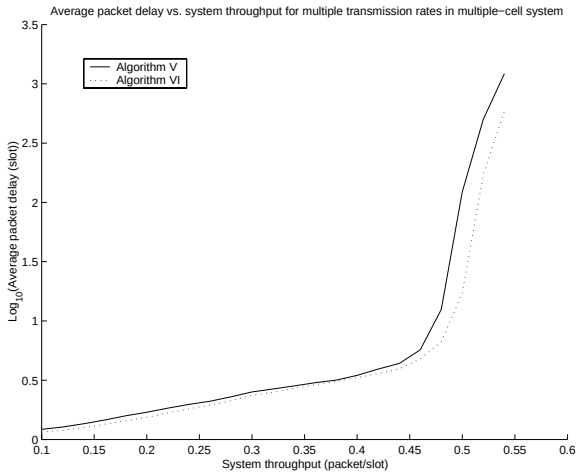

Fig. 5. Delay vs. throughput for multiple rate communication in multiple cell system

this paper, 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 for multiple cell systems. The intuition behind these heuristic algorithms is to approximate the optimal scheduling algorithm with lower computational complexity. Performance results indicate that this joint consideration of MAC layer scheduling algorithm and physical layer beamforming and power control yields significant system improvement as opposed to 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, Feburary 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 schdeuling 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