

Performance Analysis of SIMO Space–Time Scheduling With Convex Utility Function: Zero-Forcing Linear Processing

Vincent K. N. Lau, *Senior Member, IEEE*, and Yu-Kwong Kwok, *Senior Member, IEEE*

Abstract—In a multiple-antenna system, an optimized design across the link and scheduling layers is crucial toward fully exploiting the temporal and spatial dimensions of the communication channel. In this paper, based on discrete optimization techniques, we derive a novel analytical framework for designing optimal space–time scheduling algorithms with respect to general convex utility functions. We focus on the reverse link (i.e., client to base station) and assume that the mobile terminal has a single transmit antenna while the base station has n_R receive antennas. In order that our proposed framework is practicable and can be implemented with a reasonable cost in a real environment, we further assume that the physical layer involves only linear-processing complexity in separating signals from different users. As an illustration of the efficacy of our proposed analytical design framework, we apply the framework to two commonly used system utility functions, namely *maximal throughput* and *proportional fair*. We then devise an optimal scheduling algorithm based on our design framework. However, in view of the formidable time complexity of the optimal algorithm, we propose two fast practical scheduling techniques, namely the *greedy algorithm* and the *genetic algorithm* (GA). The greedy algorithm, which is similar to the one widely used in 3G1X and Qualcomm high-data-rate (HDR) systems (optimal when $n_R = 1$), exhibits significantly inferior performance when $n_R > 1$ as compared with the optimal approach. On the other hand, the GA is quite promising in terms of performance complexity tradeoff, especially for a system with a large number of users with even a moderately large n_R . As a case in point, for a system with 20 users and $n_R = 4$, the GA is more than 36 times faster than the optimal while the performance degradation is less than 10%, making it an attractive choice in the practical implementation for real-time link scheduling.

Index Terms—Fairness, genetic algorithm (GA), multiple antenna, optimal algorithm, scheduling, single-input–multiple-output (SIMO), utility functions.

I. INTRODUCTION

AS GOVERNED by the traditional design, the achievable bit rate of a wireless communication link is limited by the available bandwidth and power. Given a fixed power budget, the only way to increase the bit rate is to increase the transmission bandwidth. Recently, however, there is an intense interest in pursuing an interesting approach for increasing the bit rate without increasing the bandwidth or power budget by making use of

multiple transmit and receive antennas. The wireless communication channel with multiple antennas is then transformed into a multiple-input–multiple-output (MIMO) channel, for which the link capacity gain is contributed by the n spatial channels induced by the array of antennas. It is shown in [2] and [22] that the channel capacity of an MIMO system is proportional to $n = \min\{n_T, n_R\}$, where n_T is the number of transmit antenna and n_R is the number of receive antenna, with the help of intelligent signal-processing techniques [5], [6], [16], [24]. A plethora of research efforts have been devoted to enhancing the link level throughput by multiple-antenna technologies [19], [20]. For example, a number of contributions are recently reported for boosting the data rate of universal mobile terrestrial service (UMTS) to 20 Mb/s, utilizing MIMO techniques. There is also a remarkable volume of high-quality work on how to design space–time code with full link-level diversity order. For example, [21] describes space–time trellis codes designed with full rank criteria to achieve the maximum link-level diversity. Furthermore, [14] suggests space–time turbo codes that are also designed with full rank criteria. However, although these lines of research are crucial, optimizing the MIMO link performance is insufficient from a system perspective. Indeed, it is also very important to consider the resource-management dimension at the system level to effectively exploit the performance potential of multiple-antenna technologies.

In multiuser systems with bursty data sources, we are more concerned about the system-level performance. In addition to the physical (or link) layer, the scheduling algorithm in the multiple access control (MAC) layer plays a critical role in determining the multiuser system performance. Specifically, in a K -user system, system performance is usually expressed as $U(\bar{R}_1, \dots, \bar{R}_K)$, which is the value of a *utility function* of individual user throughput, denoted by R_k . Here, $\bar{R}_k = \mathcal{E}[R_k]$ is the average throughput of user k . For example, to maximize the system capacity, the utility function is given by

$$U_{\text{thp}}(\bar{R}_1, \dots, \bar{R}_K) = \sum_{k=1}^K \bar{R}_k. \quad (1)$$

Schedulers designed to optimize such an utility function will result in the highest system capacity. However, users with a poor channel condition could be discriminated (e.g., always get assigned a zero value of R_k in the optimization process) and, thus, could suffer from *starvation*. This unfair situation is clearly undesirable [17], [18]. On the other hand, *proportional fairness* (PF) [8] is another popular criterion that was introduced to strike a balance between system capacity and fairness among

Manuscript received December 15, 2002; revised July 14, 2003, November 5, 2003, and December 10, 2003. This work was supported by a grant from the Research Grants Council of the HKSAR under Project HKU 7162/03E.

V. K. N. Lau is with Bell Laboratories, Lucent Technologies, Whippany, NJ 07054 USA (e-mail: knlau@ieee.org).

Y.-K. Kwok is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: ykwok@eee.hku.hk).

Digital Object Identifier 10.1109/TVT.2004.823507

users. In formal terms, a scheduler is called proportional fair if it optimizes the utility function defined as

$$U_{\text{PF}}(\bar{R}_1, \dots, \bar{R}_K) = \sum_{k=1}^K \log_2(\bar{R}_k). \quad (2)$$

Traditionally, the design of the link and MAC layers is done separately (modular design to cope with the complexity involved). This approach is effective for fixed-line systems in which the physical layer is essentially time invariant. However, for wireless systems, the physical-layer (link-layer) performance is time varying (driven by the time-varying wireless channel conditions due to fading effects). Thus, optimizing the link level alone does not always result in an optimal system level performance. As reported in [9] and [10], a joint design of the MAC and link layers can achieve a significant performance gain over the isolated design approach for single-antenna systems. This gain is contributed by the multiuser diversity, which is achieved by allocating system resource adaptively to the user(s) with the best channel condition. For multiple-antenna systems, it is also recently shown in [11] and [27] that maximizing the link diversity order (which maximizes the link capacity) does not always result in the optimal system capacity. Thus, it is apparent that joint optimization between the MAC and link layers is very important to the system-level performance of wireless systems [12], [13], [26].

Unfortunately, cross optimization between the link and MAC layers is generally ignored in traditional MAC layer design due to the formidable complexity involved. Nevertheless, there have been some recent attempts that try to take advantage of the optimization gap by cooperative scheduling, which takes the link-level metrics into account in making scheduling decisions [1], [7]. For example, the scheduling efficiency is shown to be greatly enhanced in a single-input–single-output (SISO) system with a variable-throughput adaptive physical layer in [10] by a jointly adaptive approach in which a scheduling decision is made by considering the instantaneous throughput that is achievable by the user. However, all the abovementioned approaches are designed based on a heuristic approach and performance is evaluated through simulations. With the heuristic approach, we do not have an accurate design insight and we do not know how far is the performance of the algorithm from the optimal. In [23], an analytical framework has been proposed to optimize the scheduling algorithm for a SISO multiuser system with average power constraint based on an information theoretical approach with respect to the utility function U_{thp} . It is shown that the optimal scheduling policy is to allocate resources to, at most, a single user (with the best channel condition) at a time. In [5] and [24], downlink broadcast channel scheduling is considered. However, when multiple antennas are introduced to the multiuser system, the resource space is expanded to include the spatial dimension as well and the framework and results do not apply to multiuser multiple-antenna systems.

In this paper, we propose a general analytical framework for the optimal space–time scheduler design of multiuser single—input—multiple-output (SIMO) systems. The frame-

work is general in that it can be applied to capture the space–time scheduler design for a wide class of multimedia applications with different convex utility functions. We consider the reverse link performance with slow fading. Channel encoding and decoding frames are short bursts that are much shorter than the coherence time of the fading channel. This short-burst model¹ is realistic for practical systems [such as HSDPA and high data rate (HDR)] with slow mobility where a typical burst duration lasts for 2 ms, which is much shorter than the coherence time of fading channels with pedestrian mobility. In summary, the contributions of the paper are listed here.

- **An analytical framework for SIMO multiuser space–time scheduler design with respect to convex utility functions.** The framework is based on an information theoretic approach with one important practical constraint of linear-processing complexity at the physical layer of the base station. The significance of the analytical framework lies in providing an optimal performance reference for heuristic scheduler design so that we can have a better idea of the performance–complexity tradeoff.
- **Design of low-complexity space–time scheduler using U_{thp} and U_{PF} as examples.** Since the optimal algorithm is too complex to realize in practice, we have to use heuristic algorithms. In this paper, we will consider two heuristic algorithms, namely the *greedy-* and *genetic-based algorithms*. Note that the greedy-based algorithms² are widely used in the wireless industry (e.g., HS-DPA and 3G1x), which is optimal for $n_R = 1$. As we will illustrate in this paper, there is a large performance penalty with the greedy-type algorithm for $n_R > 1$. On the other hand, the genetic-based algorithm has a promising complexity–performance tradeoff.

This paper is organized as follows. In Section II, we outline the multiuser SIMO channel model, the multiuser link layer model, and the MAC layer model. In Section III, we formulate the analytical design framework for space–time scheduling with convex utility functions. The optimization problem is shown to be a mixed nonlinear integer programming problem and the optimal solution is derived. In Section IV, we propose two heuristic suboptimal algorithms for the optimization problem with low implementation complexity. Specifically, we illustrate with maximal throughput and proportional fair utility functions as examples. In Section V, we present numerical results to evaluate the performance of the optimal and low complexity space–time schedulers. Finally, we conclude with a brief summary of results in Section VI.

¹Shannon’s channel coding requires very large number of coded bits in a frame. At a given bandwidth and a frame duration constraint, the number of bits in a frame would be limited. However, in a practical sense [say, in high-speed downlink packet access (HSDPA)], a frame duration is 2 ms (which is short enough for constant fading assumption) and, with turbo coding, the error performance is already quite close to the Shannon’s limit. Therefore, we believe it is still within practical limit to have 1) frame duration small enough with respect to channel coherence and 2) a number of coded bits large enough for ideal channel coding.

²By greedy algorithm, we refer to those algorithms that make a scheduling decision by a single-pass inspection of all the K users. In 3G1x or UMTS, the scheduler used is proportional fair but is, at the same time, the *greedy* type because they select one user (with the largest r_k / \bar{R}_k) out of the K users in a single pass.

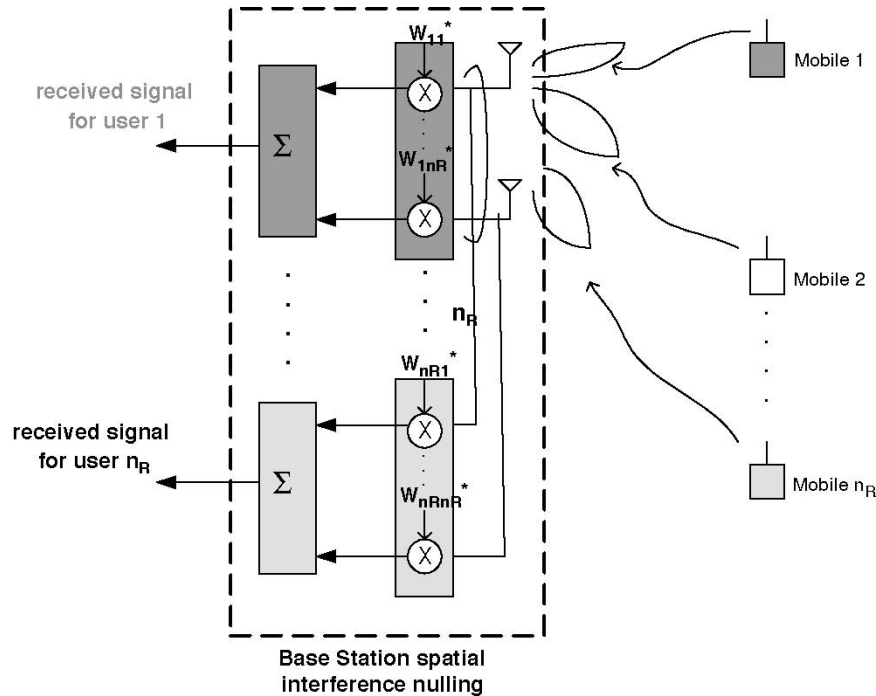


Fig. 1. Block diagram of the zero-forcing multiuser detection (MUD) at the base station with n_R receive antennas.

II. MULTIUSER SIMO SYSTEM MODEL

In this section, we describe the multiuser SIMO channel model, the link layer model, and the MAC layer model.

A. Channel Model

We consider a wireless communication system with K mobile terminals having single transmit antenna and a base station with n_R receive antennas. This is a realistic model because most mobile-client devices that are available today and in the foreseeable future can afford to only have single transmit antenna. The channel fading between different users and different antenna are modeled as independent identically distributed (i.i.d.) complex Gaussian process with unit variance. Furthermore, it is assumed that the encoding and decoding frames are short bursts that are much shorter than the coherence time of the fading channel. This is a reasonable assumption because most high-speed wireless data systems operate in short-burst mode. For example, in UMTS-HSDPA, a typical burst only lasts for 2 ms, which is much shorter than the coherence time of slow fading channels (at pedestrian speed). Thus, the fading coefficients are highly correlated within a coding frame. With rate feedback information to the mobile terminal (elaborated in Section II-C), nonzero ergodic capacity can be achieved.

The channel-fading coefficient between the k th mobile terminal and the base station is characterized by the $n_R \times 1$ dimension channel matrix \mathbf{H}_k . The received signal at the base station is

$$\mathbf{Y} = \sum_{k=1}^K \mathbf{H}_k X_k + \mathbf{Z} \quad (3)$$

where X_k is the transmit symbol from the k th mobile terminal, \mathbf{Z} is the $n_R \times 1$ -dimension Gaussian noise with covariance given

by $\sigma_z^2 \mathbf{I}_{n_R}$, and \mathbf{Y} is the $n_R \times 1$ -dimension vector of received symbols at the base station. We assume that the channel matrices of all K users are available to the base station at the start of scheduling. This can be practically achieved by the provisioning of pilot channels from the mobile terminals to the base station, similar to 3G1X, HDR, or UMTS systems.

B. Multiuser Link Layer Model

In order to isolate the performance of the link level from specific implementation details of modulation and channel coding, we take an information theoretic approach. Specifically, the maximum achievable rate at the link layer (with arbitrarily low error probability) is specified by the maximum mutual information and the transmitted power. We make use of the following assumptions in the parameters.

1) *Linear-Processing Multiuser Detection Constraint*: According to multiuser information theory, when K mobile terminals are transmitting simultaneously, individual signals from simultaneous transmitting mobile terminals can be separated through multiuser detection. However, the implementation complexity of the optimal multiuser detector is shown to be exponential [15] and is too complicated to realize with respect to today's technologies. In order to have a more realistic design, we assume a linear-processing constraint on the link layer of the base station. Specifically, at any fading block, a number of users are selected to transmit information to the base station (the indices of the selected users are contained in an admissible set \mathcal{A}). At the base station, signals from the selected users are obtained independently by forming strong beams to the desired user and nulls to the interfering users, which is illustrated in Fig. 1. Although the interference nulling will result in noise enhancement, such a scheme is shown to be optimal with respect to the *near-far* problem in [15] and [25]. Furthermore, asymptotically at large transmit power,

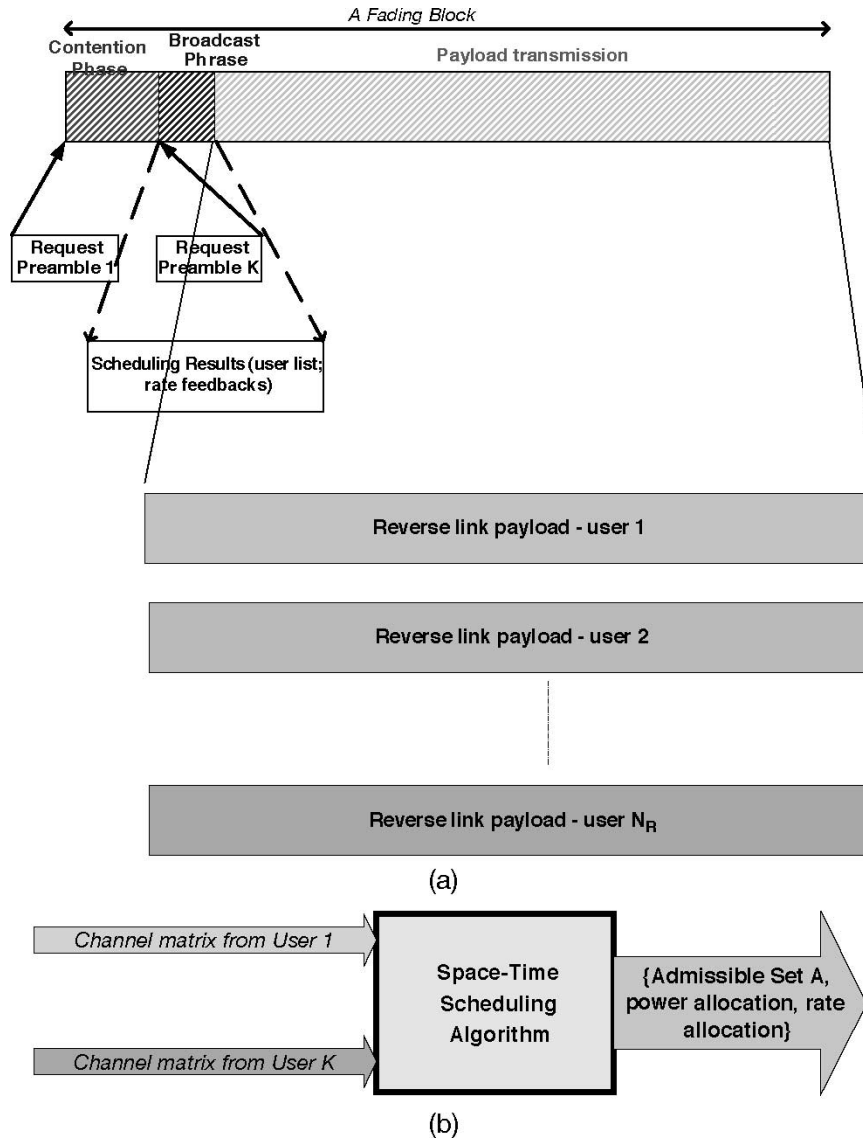


Fig. 2. Block diagrams of the MAC layer. (a) Timing of access and frame format. (b) Model of the scheduling layer.

the interference nulling scheme approaches the minimum mean-square error (MMSE) scheme.

Now, given the channel matrices of all users $\{\mathbf{h}_1, \dots, \mathbf{h}_K\}$, we multiply the received signal vector \mathbf{y}^* by an $n_R \times 1$ weight vector \mathbf{w}_k in order to extract information from user k (where $*$ denotes complex transpose). The received signal after the linear spatial processing for user k is given by

$$S_k = \mathbf{y}^* \mathbf{w}_k = \mathbf{h}_k^* \mathbf{w}_k X_k + \sum_{j \in \mathcal{A}, j \neq k} \mathbf{h}_j^* \mathbf{w}_k X_j + \mathbf{Z}^* \mathbf{w}_k \quad (4)$$

where the first term represents the desired signal, the middle term represents the interference terms, and the last term represents the residual noise after spatial processing. For interference nulling, the weight vector should be chosen according to

$$\mathbf{w}_k^* \mathbf{h}_j = \begin{cases} 1 & j = k \\ 0 & j \in \mathcal{A}, j \neq k. \end{cases} \quad (5)$$

Since there are n_R degrees of freedom and $|\mathcal{A}|$ equations, we can only form at most $n_R - 1$ nulls with n_R antennas. Thus,

the cardinality of \mathcal{A} is restricted to be less than or equal to n_R . When $|\mathcal{A}| = n_R$, (5) can uniquely determine³ the weight \mathbf{w}_k , which is given by

$$\mathbf{w}_k = (\mathbf{H}_{\mathcal{A}}^*)^{-1} \mathbf{e}_k, \quad \forall k \in \mathcal{A} \quad (6)$$

where $\mathbf{H}_{\mathcal{A}}$ is the $n_R \times |\mathcal{A}|$ -dimension aggregate channel matrix with columns obtained from the channel matrix indexed by the admissible set \mathcal{A} and \mathbf{e}_k is an $|\mathcal{A}| \times 1$ column binary matrix with all zero elements except the k th element. On the other hand, when $|\mathcal{A}| < n_R$, there are more degrees of freedom than the number of equations in (5) and, hence, we can further specify the weight \mathbf{w}_k by minimizing the noise variance in (4). That is

$$\mathbf{w}_k = \arg \min_{\mathbf{w}} \|\mathbf{w}\|^2 \quad (7)$$

where \mathbf{w} satisfies (5) as a constraint. Let $\mathbf{B}_{\mathcal{A},k}$ be an $n_R \times (n_R - |\mathcal{A}| + 1)$ -dimensional matrix with columns spanning the $(n_R - |\mathcal{A}| + 1)$ -dimension orthogonal subspace of the

³When $|\mathcal{A}| = n_R$, the rank of $\mathbf{H}_{\mathcal{A}}$ is n_R with probability 1.

$(|\mathcal{A}| - 1)$ -dimension *interference space* spanned by $\{\mathbf{h}_j, j \in \mathcal{A}, j \neq k\}$. That is

$$\mathbf{B}_{\mathcal{A},k}^* \mathbf{h}_j \begin{cases} \neq \mathbf{0} & j = k \\ = \mathbf{0} & j \neq k, j \in \mathcal{A}. \end{cases} \quad (8)$$

Observe that constraint (5) is satisfied if and only if \mathbf{w}_k belongs to the *orthogonal subspace* spanned by the columns of $\mathbf{B}_{\mathcal{A},k}$. Thus, $\mathbf{w}_k = \mathbf{B}_{\mathcal{A},k} \mathbf{u}$ for some $(n_R - |\mathcal{A}| + 1)$ -dimension column vector \mathbf{u} . The optimization problem in (7) reduces to

$$\mathbf{w}_k = \mathbf{B}_{\mathcal{A},k} \arg \min_{\mathbf{u}^* \mathbf{B}_{\mathcal{A},k}^* \mathbf{h}_k = 1} [\mathbf{u}^* \mathbf{B}_{\mathcal{A},k}^* \mathbf{B}_{\mathcal{A},k} \mathbf{u}]. \quad (9)$$

The solution to (9) is given by

$$\mathbf{w}_k = \frac{\mathbf{B}_{\mathcal{A},k} \left(\mathbf{B}_{\mathcal{A},k}^* \mathbf{B}_{\mathcal{A},k} \right)^{-1} \mathbf{B}_{\mathcal{A},k}^* \mathbf{h}_k}{\mathbf{h}_k^* \mathbf{B}_{\mathcal{A},k} \left(\mathbf{B}_{\mathcal{A},k}^* \mathbf{B}_{\mathcal{A},k} \right)^{-1} \mathbf{B}_{\mathcal{A},k}^* \mathbf{h}_k}. \quad (10)$$

The algorithm to obtain $\mathbf{B}_{\mathcal{A},k}$ is outlined in Appendix A.

Since the residual noise after linear spatial processing is still Gaussian, the input capacity achieving distribution for X_k is Gaussian and the maximum achievable data rate of the k th user during this fading block is given by

$$r_k = \begin{cases} \log_2 \left(1 + \frac{p_k}{\sigma_z^2 \mathbf{e}_k^* (\mathbf{H}_{\mathcal{A}}^* \mathbf{H}_{\mathcal{A}})^{-1} \mathbf{e}_k} \right) & |\mathcal{A}| = n_R \\ \log_2 \left(1 + \frac{p_k |\mathbf{h}_k^* \mathbf{\Omega}_{\mathcal{A},k} \mathbf{h}_k|^2}{\sigma_z^2 \mathbf{h}_k^* \mathbf{\Omega}_{\mathcal{A},k}^* \mathbf{\Omega}_{\mathcal{A},k} \mathbf{h}_k} \right) & |\mathcal{A}| < n_R \end{cases} \quad (11)$$

where p_k is the transmitted power of user k , $\mathbf{\Omega}_{\mathcal{A},k} = \mathbf{B}_{\mathcal{A},k} (\mathbf{B}_{\mathcal{A},k}^* \mathbf{B}_{\mathcal{A},k})^{-1} \mathbf{B}_{\mathcal{A},k}^*$.

2) *Mobile Power Constraint*: Due to the short burst duration, power adaptation within a coding frame is unnecessary because the channel fading will be highly correlated within the entire coding frame. The average transmitted power of the k th mobile terminal is assumed to be constrained (where the averaging window is over one encoding frame) to P_k . Thus, given any channel-fading realization, the transmitted power of user k , p_k , is constrained by

$$p_k \leq P_k. \quad (12)$$

C. MAC Layer Model

The scheduling algorithm in the MAC layer is responsible for the allocation of channel resources at every fading instance. The system resource is partitioned into frames, as illustrated by Fig. 2(a). A frame is divided into two parts, namely the request contention mini-slots and the payload transmission slot. At the start of every fading block, mobile terminals with backlogged data transmit access preambles to the base station. Based on the survived access preamble after contention, the base station measures the channel matrix \mathbf{h}_k from all the participating users. To decouple the scheduler design from the data source model and the contention dynamics, we assume that the data buffer in all the mobile terminals are sufficiently large so that there will be no empty payload slot due to empty source buffers in the mobile terminals. In other words, we assume that all the K mobile terminals are competing for system resource at every scheduling instance.

After the base station collects all the requests from the participating mobile terminals, the estimated channel states $\{\mathbf{h}_1, \dots, \mathbf{h}_K\}$ are passed to the scheduling algorithm in the

MAC layer, as illustrated in Fig. 2(b). The outputs of the scheduler consist of an admissible set $\mathcal{A} = \{k \in [1, K] : p_k > 0\}$, which is the set of user indices with nonzero power allocated at the current fading block, the corresponding power-allocation $\{p_k\}$, as well as the rate feedback⁴ for the selected users. Thus, the scheduler is modeled as an optimization algorithm with respect to a system-utility function. Note that with the linear processing constraint, the order of indices in the admissible set, is irrelevant to the utility function.

III. GENERAL FORMULATION OF THE SPACE-TIME SCHEDULER

In this section, we describe our formulation of the optimal space-time scheduling problem based on the channel, link-layer, and MAC models introduced in Section II. In general, the solution to the scheduling problem will be very tedious if the utility function is expressed in the general form $U(\bar{R}_1, \dots, \bar{R}_K)$ because it contains the average throughput $\bar{R}_k = \mathcal{E}[r_k]$. To simplify the problem, we shall restrict the class of utility functions to be of the form

$$U(\bar{R}_1, \dots, \bar{R}_K) = \mathcal{E}_{\{\mathbf{h}_1, \dots, \mathbf{h}_K\}} [G(r_1, \dots, r_K)] \quad (13)$$

where $G(r_1, \dots, r_K)$ is a convex function in (r_1, \dots, r_K) and the expectation is taken over all channel matrix realization. So that the optimization is well defined, we constrain the utility function $G(r_1, \dots, r_K)$ to be

$$\frac{\partial G}{\partial r_k} \geq 0 \quad \forall r_k \geq 0. \quad (14)$$

Introduce a binary indication variable $\alpha_k \in \{0, 1\}$ for user k . Note that $\alpha_k = 1$ if user $k \in \mathcal{A}$ and $\alpha_k = 0$ if $k \notin \mathcal{A}$. The optimization problem is given by the following.

Problem 1: Given any channel matrices realization for all mobile terminals $\{\mathbf{h}_1, \dots, \mathbf{h}_K\}$, determine the optimal $\{\alpha_1, \dots, \alpha_K\}$ and the corresponding power allocation $\{p_k \leq P_k : \alpha_k = 1\}$ so that the system-utility function $G(r_1, \dots, r_K)$ is maximized and $\sum_{k=1}^K \alpha_k \leq n_R$.

Observe that the optimization problem above is a mixed nonlinear integer-programming problem. The standard computational procedure for obtaining the optimal solution is to discretize the resource space so that discrete optimization techniques can be applied to the space-time scheduler design.

A. Optimal Solution—Integer Programming

Observe that the mixed nonlinear integer-programming problem in Problem 1 can be transformed into a pure integer-programming problem as a result of the lemma here.

Lemma 1: Given any admissible set \mathcal{A} and convex utility function $G(r_1, \dots, r_K)$, the optimizing power-allocation p_k is given by P_k for all $k \in \mathcal{A}$.

Proof 1: For any given admissible set \mathcal{A} , assume that the optimal power allocation is $\mathcal{P}^* = (p_1^*, \dots, p_K^*)$, where $p_k^* \neq P_k$ for some $k \in \mathcal{A}$ and the corresponding rate vector is given by $\mathcal{C}^* = (r_1^*, \dots, r_K^*)$. Consider another power allocation, $\mathcal{P}' = (p_1^*, \dots, P_k, \dots, p_K^*)$. From (11), the corresponding rate vector is given by $\mathcal{R}' = (r_1^*, \dots, r'_k, \dots, r_K^*)$ where $r'_k > r_k^*$. Since the utility function $G(r_1, \dots, r_K)$ is convex, we have $G(r_1^*, \dots, r_k^*, \dots, r_k^*) < G(r_1^*, \dots, r'_k, \dots, r_K^*)$.

⁴The rate feedback is used to ensure nonzero ergodic capacity at the current fading block.

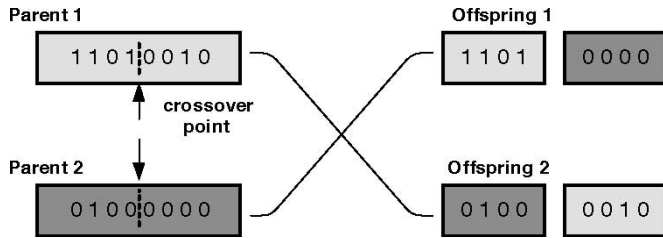


Fig. 3. Illustration of crossover operation in genetic algorithm (GA).

This implies that \mathcal{P}' is the optimizing power allocation, which contradicts the original assumption.

Due to the above lemma, the search space for the scheduling policies is discretized and we apply conventional discrete-optimization techniques [4] to solve the problem. The optimization problem is defined here.

Problem 2: Given any channel matrices realization for all mobile terminals $\{\mathbf{h}_1, \dots, \mathbf{h}_K\}$, find the optimal $\alpha_k \in \{0, 1\}$, $\forall k \in [1, K]$ such that $\sum_k \alpha_k \leq n_R$ and $G(r_1, \dots, r_K)$ is maximized. The power allocation is given by

$$p_k = \begin{cases} P_k & \alpha_k = 1 \\ 0 & \alpha_k = 0. \end{cases} \quad (15)$$

Computational complexity of the optimal solution depends critically on the form of the utility function. Obviously, the computational complexity is upper bounded by the size of the search space, which is given by $\sum_{m=1}^{n_R} \binom{K}{m}$. For even moderate K and n_R , the total search space becomes exceedingly large. Thus, low complexity suboptimal heuristics are much needed for feasible implementation.

IV. LOW-COMPLEXITY SPACE-TIME SCHEDULER DESIGN

While integer programming offers the optimal solution, the computational complexity exceeds the implementation limitation in most designs. In this section, we introduce two heuristic space-time scheduling algorithms with respect to the convex utility function. They are, namely, the *greedy algorithm* and the *genetic algorithm* (GA). We evaluate their performance by comparing them with the optimal solution in Section V.

A. Suboptimal Solution A—Greedy Algorithm

The greedy algorithm is widely used in commercial systems such as 3G1X and Qualcomm HDR due to its simplicity. The algorithm is described here.

Algorithm 1

- Step 1) For $k = 1, K$, initialize $\alpha(k) = (0, \dots, 1, \dots, 0)$ (the only nonzero element is in the k th position). Calculate $\mathbf{r} = (0, \dots, r_k, 0, \dots, 0)$ based on $\alpha(k)$.
- Step 2) Sort according to the descending order of $\{G_k\}$, calculated in Step 1).
- Step 3) The admissible set is obtained by including the first n_R indices in the sorted list in Step 2).

Observe that the greedy algorithm complexity involves K function evaluations plus sorting operation. Furthermore, it is the optimal algorithm when $n_R = 1$.

B. Suboptimal Solution B—GA

As we will illustrate in the next section, the performance gap between the greedy algorithm and the optimal algorithm is quite large for $n_R > 1$. Thus, this motivates the design of other suboptimal algorithms with better performance-complexity tradeoff. In this section, we describe our proposed GA to the space-time scheduling problem.

GA is a family of computational models inspired by evolution and is first investigated by Holland [3]. These algorithms encode a potential solution to a specific problem with a simple *chromosome-like* data structure and apply recombination operators to a set of these structures to preserve critical information. A chromosome, in its simplest form, is a string of bits that uniquely represents the independent variable of the optimization problem. An implementation of a GA begins with a population of random chromosomes. The main template of GA is illustrated here.

Algorithm 2

- Step 1) **Initialization:** Initialize a population with N_p chromosomes.
- Step 2) **Selection:** Construct an intermediate population based on selection according to the fitness of chromosomes in the initial population.
- Step 3) **Breeding:** Randomly select a pair of chromosomes in the intermediate population and recombine the two parents into two offspring according to crossover and mutation rules. The crossover operation is characterized by a crossover probability p_c . For every selected pairs of parents, there is p_c probability of performing a crossover operation. By crossover, a randomly selected crossover point (between 1 and K) is selected for the pair of chromosomes. The two parents are split, respectively, in the crossover point selected and the two offsprings are obtained by crossing the fragments of the two parents, as illustrated in Fig. 3. For every bit in the chromosomes of the offspring, there is a p_m (mutation rate) chance of toggling the bit. This is called the mutation operation.

Step 4) **Termination:** Replace the original population with the new population and repeat Steps 2) and 3) until the number of iterations reaches N_g .

Note that GA is well suited for the space-time scheduling problem. This is because the optimizing variable $(\alpha_1, \dots, \alpha_K)$ is a binary string and is represented naturally by the *chromosome* of GAs without extra encoding efforts. Second, the GA can be applied for general utility functions. Instead of the regular version of the GA, we apply adaptive a mutation rate based on the diversity of the population. Detailed steps of the GA specific to our problem is elaborated here.

Chromosome: A chromosome is a sample of the optimizing variable $(\alpha_1, \dots, \alpha_K)$, where $\alpha_k \in \{0, 1\}$.

Population Initialization: A population is a set of N_p chromosomes. The starting population is initialized with N_p randomly picked chromosomes satisfying the constraint $\sum_{k=1}^K \alpha_k \leq n_R$.

Selection: During Step 2), an intermediate generation (which is a set of N_p chromosomes) is formed based on the current population and the selection rule. The selection process we adopt is the *remainder stochastic sampling*. For each randomly selected chromosome in the current population $\alpha(i) = (\alpha_1, \dots, \alpha_K)$, we evaluate the *fitness* based on the utility function $G_i = G(r_1, \dots, r_K)$. Let $\bar{G} = (1/N_p) \sum_i G_i$ be the average fitness within the current population. The integer portion of G_i/\bar{G} indicates how many copies of that chromosome i are directly placed in the intermediate population. An additional copy of that chromosome is placed in the intermediate population with a probability equal to the fractional part of G_i/\bar{G} . For example, a chromosome with $G_i/\bar{G} = 1.36$ places one copy in the intermediate generation and receives a 0.36 chance of placing a second copy. The selection process carries on until all the N_p slots in the intermediate generation have been filled up. In this manner, we see that chromosomes that are *fitter* will be allowed more of a chance to propagate into the next population.

Adaptive Mutation Rate: After selection is completed, we perform breeding as in Step 3). The breeding process consists of two stages, namely the *crossover* and the *mutation*. These two processes introduce randomness into the intermediate generation so that the new population will be a combination of the best chromosome in the current population as well as some new random elements. This is necessary to avoid the algorithm getting stuck at local optima. In fact, the performance of the GA is found to be quite sensitive to the mutation probability. A large mutation probability introduces variability into the population, but too much is deleterious. The right value of p_m is usually problem specific. In this paper, we dynamically adjust the mutation rate based on the diversity of the fitness in the current population. The key to this strategy lies with recognizing that as long as the population is broadly distributed in the parameter space (large variance in fitness),

the crossover operation leads to a pretty efficient *search* as it recombines fragments of existing solutions, which are hopefully the “ingredients” of a better solution. However, once the population has converged (small variance in fitness), a high mutation rate is needed to reinject variability into the population. Specifically, the mutation rate of current generation is given by

$$p_m = \frac{1}{\beta_1 + \frac{\beta_2 \sigma_G}{\mu_G}} \quad (16)$$

where σ_G is the standard derivation and μ_G is the mean of the fitness of the current population (before selection). β_1 and β_2 are two constants.⁵

Elitism and Constraints: An additional feature is added to the GA to preserve the fittest member in the current population. Specifically, the fitness chromosome in the current population is saved and inserted into the next population. This is a safeguard against the possibility that the breeding process destroys the current best solution. Furthermore, all members of the next population are checked against the constraint $\sum_k \alpha_k \leq n_R$. If any chromosome violates this constraint, 0 is inserted into a randomly selected bit position in the violating chromosome until the constraint is satisfied.

The computation complexity of GA involves $N_g \times N_p$ function evaluations. As will be illustrated in the next section, this represents significant computational savings as compared with the optimal algorithm.

C. Example 1: Maximal System Capacity

We illustrate the space-time scheduler design based on a common utility function, the *total system capacity*. The utility function is given by

$$U_{\text{thp}}(\bar{R}_1, \dots, \bar{R}_K) = \mathcal{E} \left[\sum_{k=1}^K r_k \right]. \quad (17)$$

Maximizing the above utility function implies optimized system capacity. However, there might be problems of starvation when the users are *heterogeneous*. That is, mobile terminals at shadowed locations may suffer from discrimination due to persistently poor channel conditions. Thus, we usually operate the network with *fairness* constraint.

D. Example 2: Proportional Fair

Another commonly used utility function is *proportional fair* scheduler, which is applied in Qualcomm’s HDR [7] and 3G1X systems. The utility function is given by

$$U_{\text{PF}}(\bar{R}_1, \dots, \bar{R}_K) = \sum_{k=1}^K \log_2(\bar{R}_k). \quad (18)$$

Any scheduler that maximizes U_{PF} is called a proportional fair scheduler. The proportional fairness has the advantage of

⁵In this paper, $\beta_1 = 1.2$ and $\beta_2 = 10$ for the maximal throughput and proportional fair-utility functions.

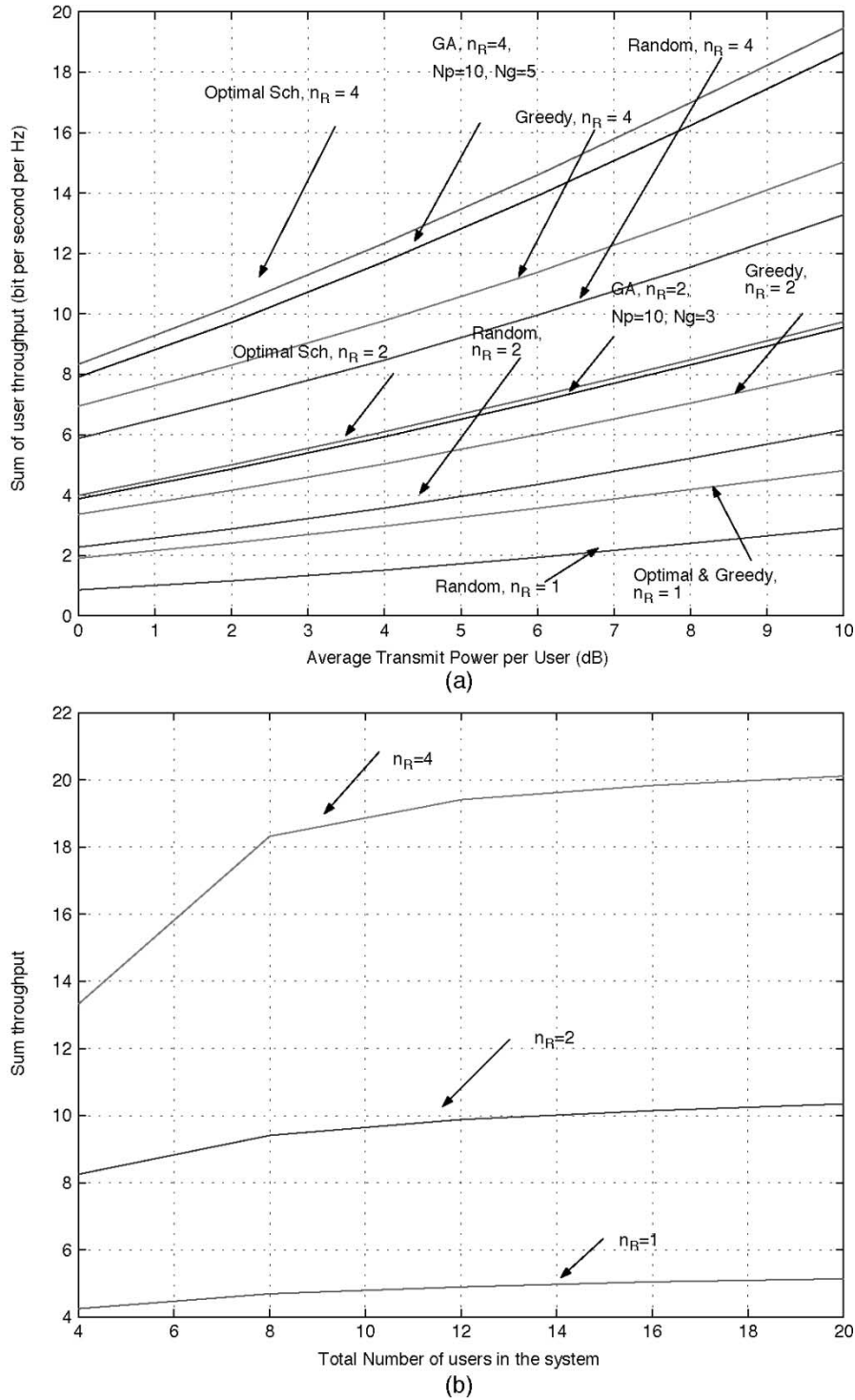


Fig. 4. Performance of the maximal throughput scheduler at various n_R . $K = 10$. (a) Total system capacity versus signal-to-noise ratio (SNR) and n_R . (b) Total system capacity versus K and n_R . (SNR = 10 dB).

achieving a reasonable tradeoff between fairness and system capacity. It also enjoys simple implementations for single-antenna systems. For instance, at any scheduling instance, the user with the largest (r_k/R_k) is selected to transmit. However, such a scheduling rule does not achieve proportional fairness for $n_R > 1$.

Note that the above utility function does not correspond to the general form discussed in the previous section. In order to

transform that into an appropriate form of the utility functions we considered, we consider the following lemmas.

Lemma 2: If a scheduler that achieves the average throughput $(\bar{R}_1^*, \dots, \bar{R}_K^*)$ satisfies

$$\sum_{k=1}^K \frac{\bar{R}_k - \bar{R}_k^*}{\bar{R}_k^*} \leq 0 \quad (19)$$

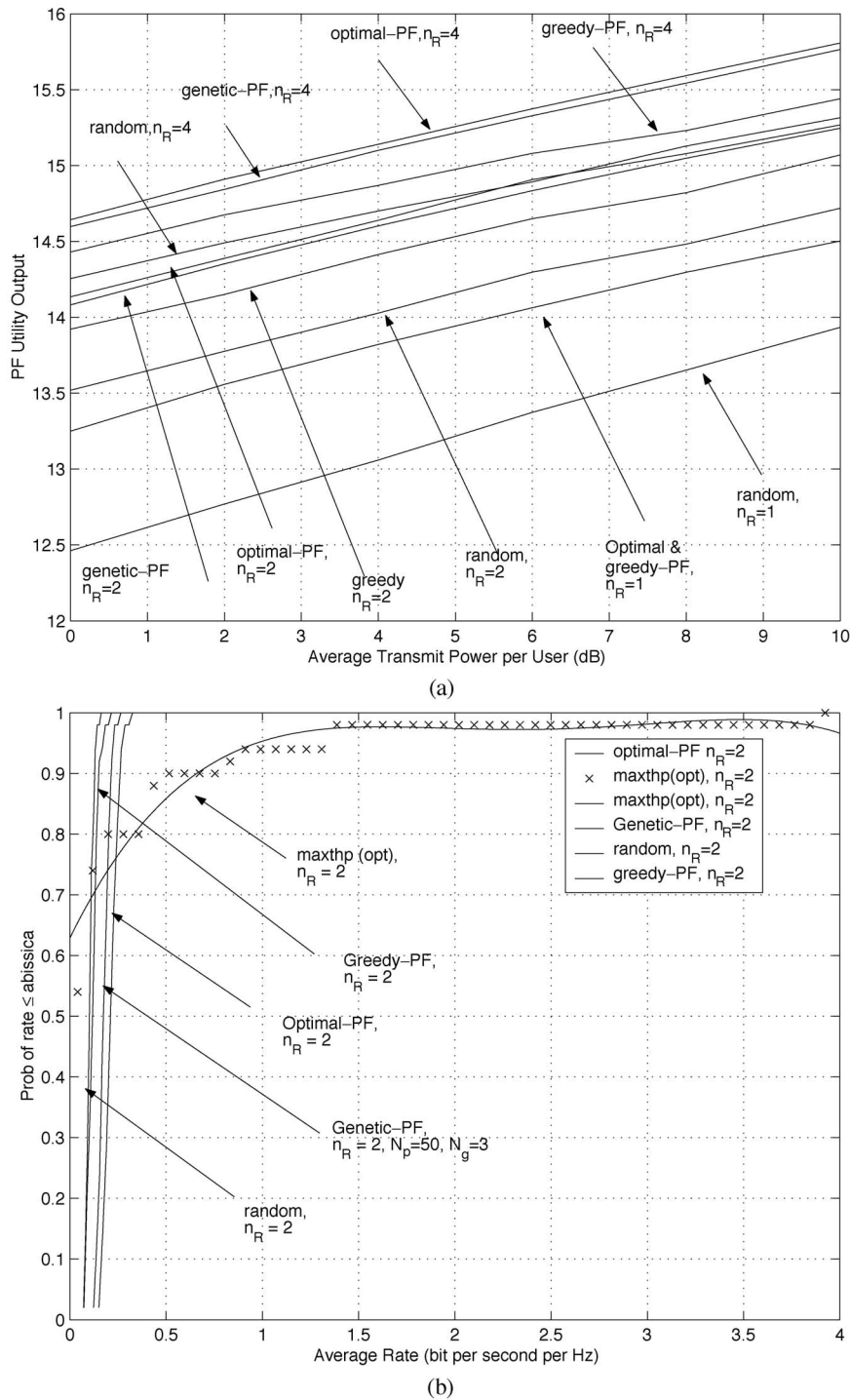


Fig. 5. Performance of proportional fair space-time scheduler at various n_R . $K = 10$ and $t_c = 1000$. (a) Utility function versus SNR and n_R . (b) Illustration of fairness. Cumulative distribution function (CDF) of mobile terminal's data rate at SNR = 10 dB, $n_r = 2$.

where $(\bar{R}_1, \dots, \bar{R}_K)$ is the average throughput of any other scheduler, then it maximizes U_{PF} .

Proof 2: This is proven in [8].

Thus, the equivalent utility function for proportional fairness is given by

$$U_{PF2} = \mathcal{E} \left[\sum_{k=1}^K \frac{r_k}{\bar{R}_k} \right] \quad (20)$$

where the terms \bar{R}_k is the average throughput of user k , which depends on the heterogeneous level among users $[1, K]$. Due to the causality requirement upon obtaining \bar{R}_k , we approximate \bar{R}_k with the *moving average throughput*. Let t denotes the t th scheduling instance, the *moving average throughput* at the $(t + 1)$ th scheduling instance $\bar{R}_k(t + 1)$ is defined as

$$\bar{R}_k(t + 1) = \left(1 - \frac{1}{t_c} \right) \bar{R}_k(t) + \frac{1}{t_c} r_k \quad (21)$$

where t_c is the *time constant* of the averaging window.

V. NUMERICAL RESULTS AND DISCUSSION

In this section, we compare the performance of the space–time schedulers with respect to the *maximal throughput* and the *proportional fair* utility functions. We compare the performance with respect to the *random* and *optimal* schedulers. The random scheduler selects randomly n_R users irrespective of their channel matrices at every fading block, which serves as performance lower bound. The optimal scheduler performance is used as a performance upper bound. The focus is on the performance of the heuristic algorithms.

A. Performance of Maximal System Capacity Scheduler

Fig. 4(a) illustrates the scheduling performance of the maximal throughput scheduler with respect to n_R and SNR. 10 000 realization of i.i.d. channel matrices for the K users are generated for each performance point. The channel matrices are zero-mean circular symmetric complex distributed. Upon each realization, the scheduling algorithm is applied to the K users and the assigned rates of the selected users are calculated based on (11). All mobile users are assumed to have the same transmit power. We observe that a significant gain in capacity is achieved by increasing the number of receive antenna n_R , which is attributed to the distributed $n_R \times n_R$ configuration. For example, there is a 2 and 4 times of total system capacity gain as compared with $n_R = 1$ at SNR = 10 dB for $n_R = 2, 4$, respectively. On the other hand, observe that the performance of the greedy algorithm coincides with the optimal scheduler when $n_R = 1$. However, there is a 2.5 and 4.5 dB of SNR penalties between the greedy and optimal performance at $n_R = 2, 4$, respectively, (at SNR = 5 dB). This motivates the GA. Observe that the GA has negligible performance loss as compared with the optimal scheduler.

Fig. 4(b) illustrates the scheduling performance at SNR = 10 dB with respect to K . As K increases, the efficiency of multiuser selection diversity increases because, at any scheduling instance, it is more likely to select a user with good channel conditions. However, supporting a large K would induce a large signaling overhead for channel estimation at the base station. From the results, we conclude that $K \approx 10$ would be sufficient to realize a significant portion of the multiuser diversity gain.

B. Performance of Proportional Fair Scheduler

Fig. 5(a) illustrates the scheduling performance of the proportional fair scheduler with respect to n_R and SNR. Mobile users are assumed to have heterogeneous path loss of variance 2 dB. The performance is represented by the utility function value. Similar to the maximal throughput utility function, significant gain is observed for $n_R = 2$ and $n_R = 4$. Similarly, the greedy algorithm achieves optimal performance only when $n_R = 1$. For large n_R , there is remarkably large performance gap between the optimal and greedy algorithms. On the other hand, GA can fill up a majority of the performance gap at various n_R .

Another way to visualize the fairness dimension of the scheduler performance is given in Fig. 5(b) where the cumulative distribution function (CDF) of a mobile terminal's throughput is plotted for $n_R = 2$ and SNR = 10 dB. The y axis illustrates

TABLE I
COMPARISON OF THE COMPUTATIONAL COMPLEXITY OF GREEDY, GENETIC, AND OPTIMAL ALGORITHMS

(K, n_R)	Greedy Algorithm	Genetic Algorithm ($N_p \times N_g$)	Optimal Algorithm
(10,2)	10 + sorting	$10 \times 3 = 30$	55
(10,4)	10 + sorting	$10 \times 5 = 50$	385
(20,2)	20 + sorting	$20 \times 3 = 60$	210
(20,4)	20 + sorting	$20 \times 5 = 100$	3645

the probability of mobile terminals acquiring a throughput less than or equal to the value in the x axis. Observe that while a maximal throughput scheduler can achieve the highest total system capacity, the chance of mobile terminals achieving such a throughput is very low. On the other hand, for the PF scheduler, although the absolute maximum throughput achieved by any mobile terminal is small than the maximal throughput scheduler, mobile terminals have higher chance of achieving a decent throughput. For example, at a CDF level of 90%, only 10% of the mobile terminals will have a throughput given by the x -axis value of the CDF curve. Thus, at 10% service guarantee (90% CDF level), the data rate of maximal throughput scheduler is 0.7 while that of the PF scheduler is just 0.25. On the other hand, at 90% service guarantee level (10% CDF level), the data rate of maximal throughput scheduler is essentially 0 while the rate of the PF scheduler is 0.2. This demonstrates the fairness dimension of the system performance. Furthermore, observe that the performance gap of the greedy algorithm against the optimal PF is quite large and is filled up by the GA.

C. Complexity Comparisons

Table I compares the number of function evaluations of the optimal algorithm, the greedy algorithm and the GA at various n_R and K . Observe that there is an 8 and 36 times speedup in the computation of GA (as compared with the optimal algorithm) when $(K, n_R) = (10, 4)$ and $(20, 4)$, respectively.

VI. CONCLUSION

In this paper, we introduce an analytical framework for the space–time scheduling problem with convex utility functions. There are K mobile terminals with single transmit antenna and one base station with n_R receive antennas. Furthermore, we assume short burst transmission and linear processing at the physical layer of the base station. The scheduling algorithm is modeled as an integer programming problem. Optimal performance is obtained for maximal throughput and proportional fair utility functions as a performance upper bound. For computational feasible solutions, we introduce two heuristics, namely the greedy algorithm and the GA, and compare their performances at different SNR and n_R . It is found that the widely used greedy algorithm, although optimal when $n_R = 1$, suffers enormous performance gap as compared with the optimal solution. This motivates the investigation of the second heuristic algorithm, namely the GA for $n_R > 1$. It is found that GA offers remarkable computational speedup as compared with the optimal algorithm and is an attractive candidate given its performance–complexity tradeoff.

APPENDIX A

 FORMATION OF ORTHOGONAL BASIS $\mathbf{B}_{\mathcal{A},k}$ GIVEN \mathbf{H} AND \mathcal{A}

Without a loss of generality, let $\mathcal{A} = \{1, 2, \dots, D, D + 1\}$ and $k = D + 1$. Let $\{\mathbf{e}_1, \dots, \mathbf{e}_D\}$ be the orthonormal basis of the vector subspace \mathcal{H} spanned by the column vectors $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_D\}$. (The orthonormal basis can be obtained easily by Gram-Schmidt process.) Thus, \mathbf{h}_j can be expressed as

$$\mathbf{h}_j = \sum_{i=1}^D (\mathbf{e}_i^* \mathbf{h}_j) \mathbf{e}_i \quad \forall j = [1, D]. \quad (22)$$

Let $\mathbf{B}_{\mathcal{A},k} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{n_R - D}]$ where \mathbf{b}_j is a $n_R \times 1$ column vector representing the basis of the orthogonal space of \mathcal{H} . We require

$$\mathbf{b}_j^* \mathbf{h}_i = 0 \quad \forall i \in [1, D], j \in [1, n_R - D]. \quad (23)$$

An obvious choice of \mathbf{b}_j is given by

$$\mathbf{b}_j = \mathbf{h}_{D+j} - \sum_{i=1}^D (\mathbf{e}_i^* \mathbf{h}_{D+j}) \mathbf{e}_i \quad (24)$$

which is the orthogonal projection of \mathbf{h}_{D+j} over \mathcal{H} .

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers (in particular, Reviewer A) for their insightful and constructive comments. They would also like to thank K. Molnar for his careful reading of our revised manuscript.

REFERENCES

- [1] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhusyana, and S. Viterbi, "CDMA/HDR: A band-width efficient high speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, pp. 70–77, July 2000.
- [2] G. J. Foschini and M. J. Gans, "On the limits of wireless communications in a fading environment," *Wireless Pers. Commun.*, vol. 6, pp. 315–335, 1998.
- [3] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor: Univ. of Michigan Press, 1975.
- [4] T. C. Hu and M. T. Shing, *Combinatorial Algorithms*, 2nd ed. New York: Dover, 2002.
- [5] H. C. Howard C. Huang, S. Venkatesan, and H. Viswanathan, "Downlink capacity evaluation of cellular networks with known interference cancellation," in *Proc. DIMACS Workshop Signal Processing for Wireless Transmission*, Oct. 2002.
- [6] H. Huang, H. Viswanathan, and G. J. Foschini, "Multiple antennas in cellular CDMA systems: Transmission, detection, and spectral efficiency," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 383–392, July 2002.
- [7] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. VTC'001*, vol. 3, May 2000, pp. 1854–1858.
- [8] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: Shadow prices, proportional fairness, and stability," *European Trans. Telecommun.*, pp. 31–37, Sept. 1997.
- [9] Y.-K. Kwok and V. K. N. Lau, "A quantitative comparison of multiple access control protocols for wireless ATM," *IEEE Trans. Veh. Technol.*, vol. 50, pp. 796–815, May 2001.
- [10] V. K. N. Lau and Y. K. Kwok, "On the synergy between adaptive physical layer and multiple access control for integrated voice and data services in a cellular wireless network," *IEEE Trans. Veh. Technol.*, vol. 51, pp. 1338–1351, Nov. 2002.

- [11] V. K. N. Lau, Y. J. Liu, and T. A. Chen, "Optimal space-time scheduling for wireless communications with partial power feedback," *Bell Labs Tech. J.*, Nov. 2002.
- [12] X. Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 2053–2064, Oct. 2001.
- [13] —, "A framework for opportunistic scheduling in wireless networks," *Comput. Networks*, vol. 41, no. 4, pp. 451–474, 2003.
- [14] Y. Liu, M. P. Fitz, and O. Takeshita, "Full rate space-time turbo codes," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 969–980, May 2001.
- [15] R. Lupas and S. Verdu, "Linear multiuser detectors for synchronous code-division multiple-access channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 123–136, Jan. 1989.
- [16] A. Mantravadi, V. V. Veeravalli, and H. Viswanathan, "Spectral efficiency of MIMO multiaccess systems with single-user decoding," *IEEE J. Select. Areas Commun.*, vol. 21, pp. 382–394, Apr. 2003.
- [17] P. Marbach, "Priority service and max-min fairness," in *Proc. INFOCOM'02*, vol. 1, June 2002, pp. 266–275.
- [18] L. Massoulie and J. Roberts, "Bandwidth sharing: Objectives and algorithms," *IEEE Trans. Networking*, vol. 10, pp. 320–328, June 2002.
- [19] L. H. Ozarow and A. G. Kogiantis, Forward link DMAS with phase information, in Lucent Internal Tech. Memo. (JW9130000-011 102-02TM), pp. 1–16, Nov. 2001.
- [20] A. Pandey, S. Emeott, J. Pautler, and K. Rohani, "Application of MIMO and proportional fair scheduling to CDMA downlink packet data channels," in *Proc. VTC'02*, vol. 2, Sept. 2002, pp. 1046–1050.
- [21] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inform. Theory*, vol. 44, pp. 744–765, Mar. 1998.
- [22] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 39, pp. 585–595, 1999.
- [23] D. Tse and S. V. Hanly, "Multiaccess fading channels—Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2796–2815, Nov. 1998.
- [24] H. Viswanathan and J. Balakrishnan, "Space-time signaling for high data rates in EDGE," *IEEE Trans. Veh. Technol.*, vol. 51, pp. 1522–1533, Nov. 2002.
- [25] J. H. Winters, "Optimal combining in digital mobile radio with cochannel interference," *IEEE J. Select. Areas Commun.*, vol. IT-2, pp. 528–539, July 1984.
- [26] M. Xiao, N. B. Shroff, and E. K. P. Chong, "A utility-based power control scheme in wireless cellular systems," *IEEE Trans. Networking*, vol. 11, pp. 210–221, Apr. 2003.
- [27] W. Yu, W. Rhee, and J. M. Cioffi, "Optimal power control in multiple access fading channels with multiple antennas," in *Proc. ICC'01*, vol. 2, June 2001, pp. 575–579.



Vincent K. N. Lau (M'98–SM'01) received the B.Eng degree (Distinction 1st Hons.) in electrical engineering from the University of Hong Kong in 1992 and the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 1997.

He was with HK Telecom, Hong Kong, China, where he was a System Engineer for three years, responsible for transmission systems design. He joined Lucent Technologies—Bell Laboratories, Whippany, NJ as Member of Technical Staff and was engaged in algorithm design, standardization, and prototype development of CDMA2000 systems. He joined the University of Hong Kong in 1999 as an Assistant Professor and was appointed the Co-Director of the information engineering program, as well as the Co-Director of the 3G Technology Center. In July 2001, he left the University and returned to the Wireless Advanced Technology Laboratory, Lucent Technologies. His research interests include digital transceiver design, adaptive modulation and channel coding, CDMA power control, soft handoff and CREST factor control algorithms, and jointly adaptive multiple access protocols, as well as short-range wireless *ad hoc* networking. He is currently working on BLAST—MIMO systems, iterative decoding, and UMTS call-processing protocol stack design.

Dr. Lau was awarded the Sir Edward Youde Memorial Fellowship and the Croucher Foundation in 1995.



Yu-Kwong Kwok (S'94–M'95–SM'03) received the B.Sc. degree in computer engineering from the University of Hong Kong in 1991 and the M.Phil. and Ph.D. degrees in computer science from the Hong Kong University of Science and Technology (HKUST) in 1994 and 1997, respectively.

He was a Visiting Scholar in the Parallel Processing Laboratory, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN. He currently is an Associate Professor in the Department of Electrical and Electronic

Engineering, University of Hong Kong. His research interests include high-performance wireless communications, wireless networking, distributed computing systems, and mobile computing.

Dr. Kwok is a Member of the ACM, the IEEE Computer Society, and the IEEE Communications Society.