Downlink Scheduling Schemes in Cellular Packet Data Systems of Multiple-Input Multiple-Output Antennas

Aimin Sang*, Xiaodong Wang†, Mohammad Madihian* and Richard D. Gitlin*
*NEC Laboratories America, Princeton, NJ 08536, Email: sang@nec-labs.com
†Electrical Engineering Department, Columbia University, New York, NY 10027

***Abstract*— High-speed cellular data systems demand fast downlink scheduling algorithms and Multiple-Input Multiple-Output (MIMO) techniques. The associated multiuser diversity and antenna diversity play a central role in achieving high system throughput and fair resource allocation among multiple users. For such systems we evaluate the cross-layer interactions between channel-dependent scheduling schemes and MIMO techniques, such as Space-Time Block Coding (STBC) or Bell Labs Layered Space-Time (BLAST), and propose a new scheduling algorithm named the Alpha-Rule. The evaluation shows that the STBC/MIMO provides reliable channel but at certain cost of spectral efficiency. Comparatively BLAST/MIMO provides larger capacity and enables higher scheduling throughput. Thus BLAST/MIMO may be a more suitable technique for high-rate packet data transmission at the physical layer. At the medium access control (MAC)-layer, the Alpha-Rule is shown to be more flexible or efficient to exploit the diversity gains than the exiting max-C/I or Proportionally Fair (PF) scheduling schemes. It enables online tradeoff between aggregate throughput, per-user throughput, and per-user resource allocation.**

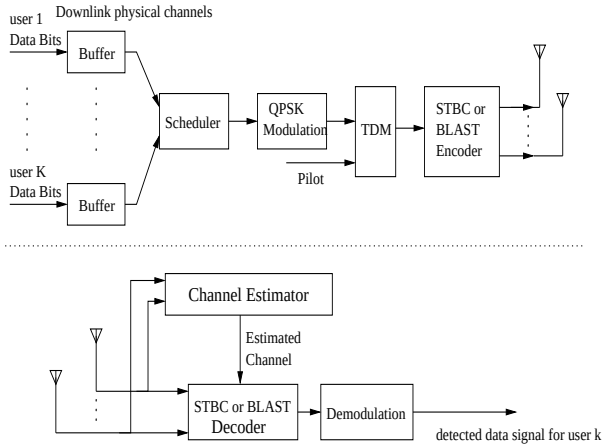

Fig. 1. The downlink transmitter structure at the base station and the receiver structure at a user terminal.

I. Introduction

High-speed downlink packet data services are the key to the success of 3G wireless cellular systems, such as CDMA2000 High Data Rate (HDR) [1] and WCDMA High Speed Data Packet Access (HSDPA) [2]. Both adopt the time-division multiple accessing (TDMA) technique for the downlink data channel shared by multiple users. To support such systems, critical technologies at different layers are being intensively studied. At the physical layer, Multiple-Input Multiple-Output (MIMO) antenna techniques can increase the channel capacity between base station (BS) and each individual user due to the *spatial (antenna) diversity*. At the media access control (MAC) layer, the opportunistic scheduler at each BS selects users for transmission according to their channel-state-information (CSI) feedback and throughput performance, characterizing the *multiuser diversity* [3]. Both diversities play a central role in such systems in achieving high throughput and fair resource allocation among all users. Our research focuses on the cross-layer interactions between scheduling and MIMO channel, and also the efficiency to exploit the two diversities in combination. In this paper, we take the HDR systems for an example, but our study applies to the HSDPA systems as well.

Multiple-Input Multiple-Output (MIMO) antenna techniques [4], [5], [6] have been studied extensively in the recent past. One technique is the orthogonal space-time block coding (STBC) [4], [7], which achieves both "full transmit diversity" and reliable communication, but fails to provide a linearly increasing channel capacity as the number of transmit and receive antennas grows simultaneously. Another technique is the Vertical Bell Labs Layered Space-Time (V-BLAST) [8], [9], which provides high-rate data transmission, but does not guard against deep fading. In this paper we will see their performance in supporting high-rate data packet services.

Over different antenna arrays and MIMO techniques, we evaluate several TDM-based downlink scheduling schemes. Our purpose is to see their efficiency in exploring the multiuser diversity, i.e., the independent channel dynamics of multiple users, and the STBC or BLAST-provided antenna diversity. Attacking the limitations of existing schemes, we propose the "Alpha-Rule" scheduling algorithm. By tuning a control variable α, the Alpha-Rule can migrate between algorithms such as the Maximum Carrier-to-Interference Ratio (max-C/I) scheduling [10], the Proportionally Fair (PF) scheduling [11], [12], and the Max-Min fair [13] scheduling. Note that those algorithms are all channel-dependent, with difference in per-user time-slot allocation, per-user throughput, and overall system (aggregate) throughput.

Our Study shows that the Alpha-Rule can effectively balance between aggregate throughput and fairness. STBC/MIMO may be more suitable than BLAST/MIMO to provide low to medium rate (voice) services and dedicated channels, while BLAST/MIMO is suitable for high-rate packet data access over a shared channel. Our study reveals the importance of cross-layer system optimization in coordinating channel structure and scheduling algorithms.

II. System Descriptions

Figure 1 shows a high-speed downlink packet data cellular system adopting STBC or BLAST with n_T transmit antennas and n_R receive antennas. There is infinite data backlog for each station (i.e., user). The output data bits are modulated (by QPSK, say) and then time-multiplexed with the pilot signal. For this system we assume capacity-achieving codes and Maximum Likelihood (ML) decoder. At the transmitter, the data sequences are STBC or BLAST coded and transmitted. At the receiver, signals are received from n_R receive antennas. After matched filtering and symbol-rate sampling, the received discrete-time signal at the kth of totally K users is

$$\begin{aligned} \boldsymbol{r}_k(t) &= \sqrt{\frac{E_s}{n_T}}\boldsymbol{H}_k\boldsymbol{c}_k(t)+\boldsymbol{n}_k(t), \quad t=1,\cdots,T, \\ \text{with} \quad \boldsymbol{r}_k(t) &= [r_{1,k}(t),\cdots,r_{n_R,k}(t)]^T, \\ \boldsymbol{H}_k &= \left[\boldsymbol{h}_{1,k}^T,\cdots,\boldsymbol{h}_{n_R,k}^T\right]^T, \\ \boldsymbol{h}_{j,k} &= [h_{1j,k},\cdots,h_{n_Tj,k}]^T, \\ \boldsymbol{c}_k(t) &= [c_{1,k}(t),\cdots,c_{n_T,k}(t)]^T, \\ \boldsymbol{n}_k(t) &= [n_{1,k}(t),\cdots,n_{n_R,k}(t)]^T, \end{aligned}$$

where

- $c_{i,k}(t),\ i=1,\cdots,n_T$, is the symbol transmitted to the kth user from the ith transmit antenna at the tth time slot;
- E_s is the average total transmission energy in one time slot, i.e., $tr(E\{\boldsymbol{c}_k(t)c_k^H(t)\}) < E_s$;
- $\boldsymbol{H}_k$ is a circularly symmetric complex matrix of dimension $n_R \times n_T$;
- The entry $h_{ij,k}$ of $\boldsymbol{H}_k$ represents the complex channel gain from the ith transmit antenna to the jth receive antenna of the kth user; It is a complex Gaussian random variable with mean 0 and variance 0.5 per dimension.
- $\boldsymbol{n}_k(t)$ is a complex Gaussian random vector with zero mean and covariance matrix $\sigma_k^2\boldsymbol{I}$, i.e., $\boldsymbol{n}_k(t) \sim \mathcal{N}_c\left(\boldsymbol{0},\sigma_k^2\boldsymbol{I}\right)$.

Due to rich scattering we assume the spatial paths $h_{ij,k}$ $(\forall i,j)$ are independent of each other.

Assume that the channel matrix $\boldsymbol{H}_k$ is known to the receiver of each user, but not the transmitter. By [5], [6], the "perfect-knowledge" instantaneous mutual information of the MIMO channel can be written as follows :

$$r_k(t) = \text{logdet}(\boldsymbol{I}_{n_R} + \frac{\rho_k}{n_T}\boldsymbol{H}_k\boldsymbol{H}_k^H), \tag{1}$$

where $\rho_k \triangleq \frac{E_s}{\sigma_k^2}$ is the mean SNR of user k; $\boldsymbol{H}_k$ is the instantaneous channel state at time t; the capacity unit is bits/s/Hz. In the remaining part we neglect the subscript k whenever possible without confusion.

A. STBC- or BLAST-coded MIMO Channels

For the downlink channel of each user, STBC scheme ([4], [7]) transmits an array of different data streams from multiple transmit antennas simultaneously. It then duplicates the streams in time domain and sends them out from an orthogonal arrangement of the transmit antennas. For the STBC scheme of $n_T = 2$ and $n_R = 1$ or 2, an orthogonal design named Alamouti code [4] is given with the code matrix $\mathcal{G}$ defined as follows:

$$\mathcal{G} = \begin{bmatrix} c_1 & c_2 \\ -c_2^* & c_1^* \end{bmatrix}. \tag{2}$$

Based on [7], [14], the STBC/MIMO channel capacity is as follows:

$$r(t) = R\ \log(1+\text{SNR}_{STBC}), \tag{3}$$

where $R = 1$ and $\frac{3}{4}$ for $n_T = 2$ and 3, respectively; SNR_{STBC} is the effective signal-to-noise ratio (SNR) at the output of the STBC decoder:

$$\text{SNR}_{STBC} \triangleq \frac{\rho}{n_T}\sum_{j=1}^{n_R}\sum_{i=1}^{n_T}|h_{ij}|^2. \tag{4}$$

Note that the mean SNR $\rho = \frac{E_s}{\sigma^2}$ may differ for different terminals. Obviously SNR_{STBC} is a Chi-square random variable with $2n_Tn_R$ degrees of freedom.

Bell-Labs Layered Space-Time (BLAST) is another MIMO technique. In contrast to STBC, BLAST transmits heterogeneous symbol streams to the MIMO channel overlapping in both time and frequency. Thus it provides higher communication rate at the cost of reliability. Readers can refer to [5], [8], [9] for further details. Under the constraint of $n_T \le n_R$, the instantaneous mutual information of BLAST/MIMO channel follows (1). Numerical studies show that the pdfs of BLAST are quite close to the Chi-square distribution. Generally speaking, BLAST has a larger variance and higher mean, and offers a higher transmission rate than STBC, making it more suitable for high-rate data services. However, for BLAST, the decoding constraint of $n_T \le n_R$ and the lack of guards against deep fading may limit its application, especially in the downlink environment where base stations have more antennas than mobile terminals.

III. Utility Function-based Scheduling Formulation

In light of the network economy for elastic traffic of best-effort services, the utility function would be increasing, strictly concave, and continuously differentiable [15]. Thus a natural rule of scheduling would be to maximize the total "utility" $U_k(\cdot)$ generated by per-user's mean throughput $\tilde{r}_k$:

$$\max_{U_k}\sum_{k=1}^{K}U_k(E[r_k(t)\mathbf{1}_{(k^*(t)=k)}]) = \max_{\{\tilde{r}_k\}}\sum_{k=1}^{K}U_k(\tilde{r}_k(t)), \tag{5}$$

where $\mathbf{1}_{(k^*(t)=k)}$ is the instantaneous scheduling decision:

$$\mathbf{1}_{(k^*(t)=k)} = \begin{cases} 1, & \text{scheduler picks user } k \text{ at slot } t \\ 0, & \text{otherwise} \end{cases};$$

The optimization is taken over all possible solution set of $\{\tilde{r}_k(t)\}$ under the constraint $\sum_{k=1}^{K}\mathbf{1}_{(k^*(t)=k)} = 1$.

Assume we know the history up to time t, e.g., $\tilde{r}_k(t+1) \triangleq (1-\frac{1}{t_c})\tilde{r}_k(t) + \frac{1}{t_c}r_k(t)\mathbf{1}_{(k^*(t)=k)}$, where t_c is the exponential

filtering factor. Then we need to make the next-step decision. Take the steepest gradient ascent of $U(t)$ as the optimized direction of the controlled system evolution, and assume each time slot $\triangle t$ is infinitesimal and $t_c \triangle t$ constant. Thus the decision becomes a continuous-time fluid-flow process. Take the derivative in time domain:

$$\frac{dU(t)}{dt} = \sum_{k=1}^{K} \frac{\partial U_k(\tilde{r}_k(t))}{\partial \tilde{r}_k(t)} \frac{d\tilde{r}_k(t)}{dt} = \sum_{k=1}^{K} \frac{\partial U_k(\tilde{r}_k(t))}{\partial \tilde{r}_k(t)} \tilde{r}_k(t)'.$$

From the discrete-time $\tilde{r}_k(t)$, we have $\tilde{r}_k(t+\triangle t) = (1 - \frac{1}{t_c})\tilde{r}_k(t) + \frac{1}{t_c} r_k(t)\mathbf{1}_{(k^*(t)=k)}$. So $\tilde{r}_k(t)'$ is approximately:

$$\frac{\tilde{r}_k(t+\triangle t) - \tilde{r}_k(t)}{\triangle t} = \frac{r_k(t)\mathbf{1}_{(k^*(t)=k)} - \tilde{r}_k(t)}{t_c \triangle t}.$$

Therefore, the steepest gradient ascent of $U(t)$ at time t is obtained by picking the user as $k^* =$

$$\arg\max_k \{\sum_{k=1}^{K} \frac{\partial U_k(\tilde{r}_k(t))}{\partial \tilde{r}_k(t)} \frac{r_k(t)\mathbf{1}_{(k^*(t)=k)} - \tilde{r}_k(t)}{t_c \triangle t}\}. \quad (6)$$

This is our utility-based scheduling rule, while the utility function is to be defined according to practical requirements. In practice, $r_k(t)$ is the instantaneous "supportable channel rate" fed back to the BS through data rate control (DRC) channel by individual MS (k).

Note that although seemingly similar, our optimization target is actually different from the one adopted by Liu *et al.* [16] in that our utility function depends on long-term per-user mean throughput whereas [16] defines an "instantaneous" utility function and tried to maximize the expectations of the total utility under certain long-term time-fraction constraints. We argue that the long-term throughput performance may be more relevant to revenue-generation in best-effort services.

A. The Alpha-Rule Scheduling Algorithm

There are many different fairness criteria in the Internet rate control or bottleneck link sharing. The most popular one is the Max-Min fairness [13], which in terms of our problem means the feasible set of mean throughput $\{\tilde{r}_k\}$, of which any user i can not increase its mean throughput $\tilde{r}_i$ without decreasing some $\tilde{r}_j$ that is no bigger than $\tilde{r}_i$. An example to achieve near-optimum Max-Min fairness among TCP and UDP users is [17]. Later, Kelly *et al.* [18] proposed proportional fairness criterion, and Mo *et al.* [19] extended it to (p, α)-proportionally fair (see Definition 1 in [19]). Mapping (p, α)-proportionally fair into our notations, the (w, α)-proportional fairness says that given a positive $w = [w_1, \ldots, w_K]$ and a non-negative α, the vector of $\{\tilde{r}_k^*\}$ is (w, α)-*proportionally fair* if under the link-sharing capacity constraint, it satisfies

$$\sum_{k=1}^{K} w_k \frac{\tilde{r}_k - \tilde{r}_k^*}{\tilde{r}_k^{*\alpha}} < 0 \quad (7)$$

for any other non-negative and feasible vector $\{\tilde{r}_k\}$ under the same constraint. It is noted that such a $\{\tilde{r}_k^*\}$ maximizes the utility function given by $U_k(\tilde{r}_k) = w_k \frac{\tilde{r}_k^{1-\alpha}}{1-\alpha}$, where $w_k > 0$, $\alpha \geq 0$, and $U_k(\cdot)$ is a strictly concave and increasing function of $\tilde{r}_k(t)$. Yet in our scenario, there is no static capacity constraint of link-sharing among K users, but a constraint on time slot sharing instead. Following the logic in the previous section, and adopting $U_k(\tilde{r}_k) = w_k \frac{\tilde{r}_k^{1-\alpha}}{1-\alpha}$ in (6), where w_k is the *weight* of user k in the total utility, we have the following maximization target:

$$\sum_{k=1}^{K} \frac{w_k}{\tilde{r}_k(t)^\alpha} \frac{r_k(t)\mathbf{1}_{(k^*(t)=k)} - \tilde{r}_k(t)}{t_c \triangle t}$$
$$= \sum_{k=1}^{K} \frac{w_k}{t_c \triangle t} \frac{r_k(t)}{\tilde{r}_k(t)^\alpha} \mathbf{1}_{(k^*(t)=k)} - \sum_{k=1}^{K} \frac{w_k}{t_c \triangle t} \tilde{r}_k(t)^{1-\alpha}.$$

Since $\tilde{r}_k(t)$ as the history value is independent of the future $r_k(t)$ and $\mathbf{1}_{(k^*(t)=k)}$, we can neglect the second part and transform the maximization into the following scheduling scheme, which we named the *Alpha-Rule*:

$$k^* = \arg\max_k \{w_k \frac{r_k(t)}{\tilde{r}_k(t)^\alpha}\}. \quad (8)$$

B. Tradeoff between Throughput and Fairness

In contrast to wired networks, resources in wireless networks, such as the time slots, link capacity, and power, are separate and orthogonal resources among users. For this reason, per-user *throughput* is not equal to per-user (time slot) resource sharing. Although one always expects a higher *aggregate throughput* $r = \sum_{k=1}^{K} \tilde{r}_k$, the scheduler has to take care of two per-user fairness criteria, based either on the time-slot allocation or per-user throughput over the shared channel. Adopt the *fairness index* in [20]:

$$F = \frac{(\sum_{k=1}^{K} x_k)^2}{K \sum_{k=1}^{K} x_k^2}, \quad (9)$$

where x_k is the per-user metrics. F is a continuous function ranging between 0 and 1. Larger F denotes better fairness: $F = \frac{1}{K}$ is the extremely unfair case when only one x_k is nonzero, whereas $F = 1$ means completely fair since x_k's are all equal. When $x_k = r_k$, F denotes the performance-based fairness index. When x_k is the time-fraction allocated to user k, F is the resource-based fairness index.

Suppose users are equally weighted, i.e., $w_k = 1, \ \forall k$. We then look at the following special cases.

- $\alpha = 0$: The optimization target becomes $\max_{\{\tilde{r}_k\}} \sum_{k=1}^{K} \tilde{r}_k$. By (8), the Alpha-Rule reduces to $k^* = \arg\max_k \{r_k(t)\}$, i.e., the well-known max-C/I scheme [10] that always picks the user of the best channel and starves the worst-channel users, say, those who are the remotest to the base station. Obviously it maximizes the throughput without any consideration of fairness.
- $\alpha \longrightarrow 1$: The target becomes $\max_{\{\tilde{r}_k\}} \sum_{k=1}^{K} \log \tilde{r}_k$. The Alpha-Rule reduces to $k^* = \arg\max_k \{\frac{r_k(t)}{\tilde{r}_k(t)}\}$, i.e., the Proportionally Fair (PF) scheduling [11], [12] that picks the best user by its "relative" channel status. PF scheduling asymptotically guarantees an equal sharing of

time slots among all users [12], i.e., its resource-based fairness index is 1.

- $\alpha = 2$: The target is to minimize $\sum_{k=1}^{K} \frac{1}{\bar{r}_k}$ and the rule minimizes the "potential delay" of all users with equal packet length. The corresponding scheduling policy is $k^* = \arg\max_k \{\frac{r_k(t)}{\bar{r}_k(t)^2}\}$. With this rule, users of poorer channels tends to get more time slot. As a result, its aggregate throughput is lower than PF and even RR.
- $\alpha \to \infty$: This extreme case attains max-min fairness in that the scheduler achieves strict fairness about throughput performance. In other words, the scheduler tends to pick the user of the smallest mean throughput at each time slot. So a significant fraction of time goes to users of noisy channels. As a result, the aggregate throughput is the lowest among all the cases.

In (8), weight is to differentiate individual users: a larger weight implies more shares of time slot allocation. There are other efforts of weight-designing (say, for minimum rate guarantee), such as the Modified Largest Weighted Delay First (M-LWDF) scheme [21] and the Exponential Rule [22]. Beyond the focus of this paper, we note that the weight design is tightly related to per-user willingness to pay for the resources. On the other hand, our α is used to control the overall system performance. With the increase of α, scheduler allocates more time slots to users of weaker channel condition at the sacrifice of total throughput. Although it is difficult to get a closed-form solution of α for a specified target of r or F, a closed-loop tuning of α based on online measurement of them is quite intuitive.

IV. Multiuser Diversity and Scheduling

Multiuser diversity [3] means that almost at any moment, one of a large number of independent MSs has its channel status near the optimum. BS can increase its aggregate throughput by picking the best user at any moment for transmission. Denote γ as the effective SNR at the output of the decoder when k is the only user in the system:

$$\gamma \stackrel{\triangle}{=} \text{SNR}_{STBC} = \frac{\rho}{n_T} \sum_{j=1}^{n_R} \sum_{i=1}^{n_T} |h_{ij,k}|^2. \quad (10)$$

With the QPSK modulation, $E_s = 2E_b$ for 2bits/symbol, $\rho = \frac{E_s}{\sigma^2} = \frac{2E_s}{\sigma^2}$. As we know, γ is a chi-square random variable with $2n_T n_R$ degrees of freedom and pdf given by $f_\gamma(\gamma) = \frac{\mu^{n+1}}{n!}\gamma^n e^{-\mu\gamma}$, where $\mu = \frac{n_T\sigma^2}{E_s}$, $n = n_T n_R - 1$. Now suppose there are K users. Firstly, assume the channel-independent Round-Robin (RR) scheduling is adopted to randomly pick the users in a uniform way. Therefore, the pdf of the effective SNR after the scheduling remains unchanged. So there is no gain of throughput for serving more than one users. Secondly, suppose max-C/I is used to select the user. Therefore, with K-fold multiuser diversity the effective SNR after max-C/I scheduling would be

$$\tilde{\gamma} = \max_{k\in\{1,2,\cdots,K\}} \frac{E_s}{n_T\sigma^2} \sum_{j=1}^{n_R} \sum_{i=1}^{n_T} |h_{ij,k}|^2. \quad (11)$$

By order statistics [23], the pdf of $\tilde{\gamma}$ is $g_{\tilde{\gamma}}(y) = Kf_\gamma(y)F_\gamma(y)^{K-1}$, where $F_\gamma(y)$ denotes the cumulative distribution function (cdf) of γ:

$$F_\gamma(y) = \int_0^y f_\gamma(\gamma)d\gamma = 1 - \mu^{n+1}e^{-\mu t} \sum_{i=0}^{n} \frac{y^i}{i!\mu^{n-i+1}}. \quad (12)$$

With an increasing number of users, the effective SNR after max-C/I scheduling is getting bigger, which translates to a higher MAC-layer throughput. Finally, suppose PF scheme is adopted instead. Given a symmetric distribution of the effective SNR around its mean, each user gets asymptotically equal share ($\frac{1}{K}$) of the total time slots. When all the users have i.i.d. channel statistics, the PF scheme reduces to max-C/I and enjoys the same multiuser diversity gain. When the channel statistics are not identical, PF still benefits from the multiuser diversity: when the number of users is big, there is always someone whose channel is near its own peak at any moment [12]. Generally speaking, the diversity gain remains with the channel-dependent Alpha-Rule, but decreases with increasing α because the scheduler is sacrificing efficiency for fairness.

A. Multiuser and Antenna Diversity

Channel-dependent scheduling schemes, e.g., the max-C/I and PF, "rides the peak" of different user channels to transmit data. Therefore the aggregate throughput depends not only on the channel mean SNR, but the range of its dynamics as well. The larger of the channel fluctuations, i.e., where pdf has longer tails, the more the multiuser-diversity is. This phenomenon has been explained in [12]. Since STBC and BLAST MIMO explore the antenna diversity in different ways, they present different channel statistics to schedulers. STBC delivers a smoother channel of lower mean SNR than BLAST. Both degrade the throughput for smaller multiuser diversity. Therefore, for bursty data services, BLAST may be more suitable than STBC.

Refer to (10) for STBC/MIMO channels. The effective SNR of user k has mean and variance of $E[\gamma] = \rho n_R$ and Var $= \frac{n_R\,\rho^2}{n_T}$, respectively. Its coefficient of variance is $\frac{1}{\sqrt{n_T n_R}}$. With an increasing n_R, the mean SNR increases. This improves the channel throughput. On the other hand, both n_R and n_T in (10) destructively smooth the SNR dynamics including peaks due to the fading-compensation mechanism inherent in the STBC/MIMO channel. Therefore, in contrast to BLAST, antenna diversity in STBC may increase the mean SNR but reduce the multiuser diversity. The dual impacts limit the increase of scheduling throughput with antenna diversity.

V. Performance of the Scheduling Schemes over MIMO

We simulate the HDR systems over different MIMO schemes. Our focus is to evaluate the cross-layer system performance under multiuser and spatial diversity. The performance metrics are the aggregate downlink scheduling throughput and fairness among heterogeneous users at the MAC layer.

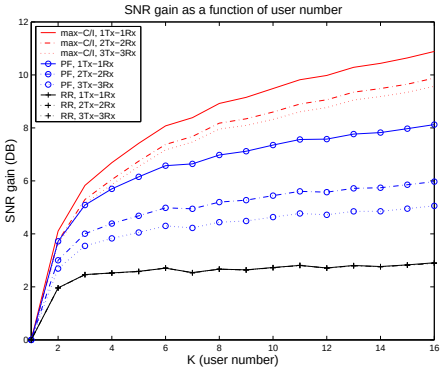

Fig. 2. STBC/MIMO: effective SNR gain as a function of scheduling and K-fold multiuser diversity.

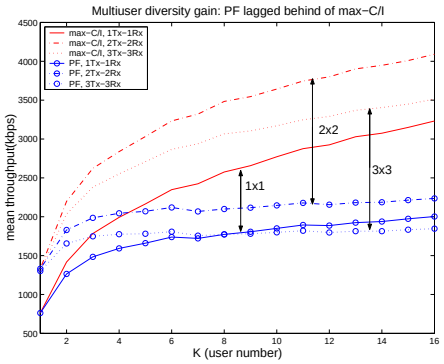

Fig. 3. STBC/MIMO: mean aggregate throughput as a function of user number.

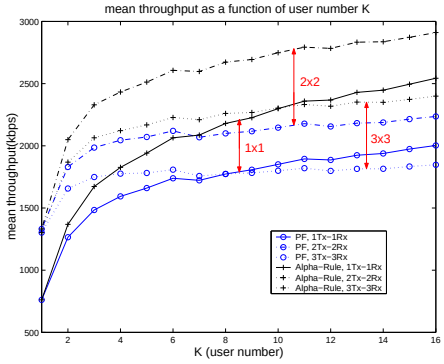

Fig. 4. STBC/MIMO: the throughput of Alpha-Rule ($\alpha = 0.5$) and PF v.s. user number.

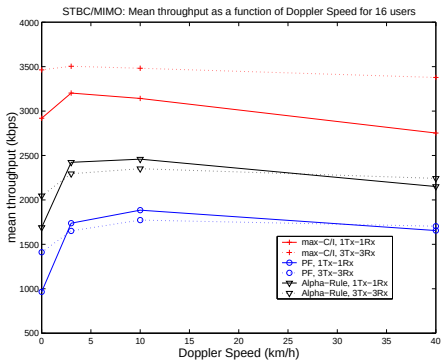

Fig. 5. STBC/MIMO: the mean aggregate throughput as a function of Doppler speed.

Simulation setup: Following the settings in [11] and [1], mobile users are uniformly located within a cell. The STBC- or BLAST-coded MIMO channels, characterized by (3) and (1), respectively, have the same statistics except non-identical ρ_k of users. The mean effective SNR ρ_k is initialized according to the cdf of the CDMA/HDR system [1]. In the tests, users may or may not move with Doppler fading for each spatial channel. For a given number of users K, 50 runs of simulations are executed to get a stable value. For each run, ρ_k is re-initiated. Each run lasts for 30 seconds, i.e., 18000 time slots. Each slot of forward channel is 1.667 ms, i.e. the slot frequency is 600. The exponential filtering interval t_c is 1000 in unit of slots. So the filtering latency timescale is 1.667 s. Unless explicitly mentioned, we assume an ideal case where "perfect" channel state is fed back without delay while the channel dynamics is memoryless.

Scheduling over STBC/MIMO channel: First take a look at the effective SNR of the multiuser STBC/MIMO downlink shared channel after scheduling. Fig. 2 clearly shows the multiuser diversity gain given multiple (K) users. As we can see, the effective SNR given both PF and max-C/I increases monotonically with K, but the gain is getting smaller from 1Tx-1Rx (denoted 1×1) to 2Tx-2Rx and 3Tx-3Rx. It is because of the decreasing multiuser diversity as stated before. Hence the STBC technique blocks the scheduling throughput to increase linearly with the number of MIMO antennas. In contrast, the gains of RR remain largely flat for increasing K and thus does not exploit the multiuser diversity. Obviously RR is not a good choice of scheduling.

Now look at the MAC-layer throughput performance of PF over STBC/MIMO. Fig. 3 illustrates two aspects. First, the system benefits from the spatial diversity offered by MIMO technique in that the throughput increases from 1Tx-1Rx to 2Tx-2Rx or 3Tx-3Rx. The lower throughput of 3Tx-3Rx than 2Tx-2Rx is due to the lack of full-rate orthogonal STBC codes with 3Tx-3Rx. Second, there is an increasing gap between PF and max-C/I given more users or higher order of antenna arrays. To shrink the gap, we may adopt the Alpha-Rule and set α less than 1 (PF) but larger than 0 (max-C/I). As an example, Fig. 4 shows that the Alpha-Rule with $\alpha = 0.5$ achieves throughput somewhere in between the PF and max-C/I regardless of antenna setups. In other words, it successfully controls the tradeoff between per-user fairness and aggregate throughput.

Fig. 5 shows the impact of mobility on mean aggregate throughput given different velocities: 0 km/h, 3 km/h, 10 km/h, and 40 km/h. The chip rate in the Doppler fading is 1.2288 Mbps with 2048 chips per slot (1.667 ms). The carrier frequency is 1.88 GHz. Note that our previous studies are for an ideal case of the following two assumptions: instantaneous

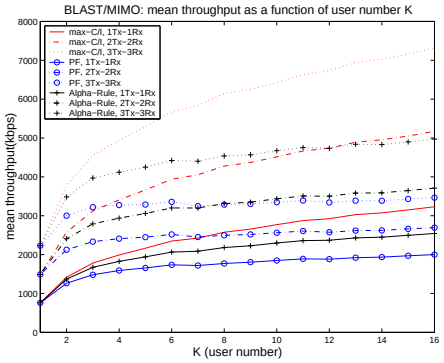

Fig. 6. BLAST/MIMO: the mean aggregate throughput of Alpha-Rule ($\alpha = 0.5$) and others scheduling v.s. user number.

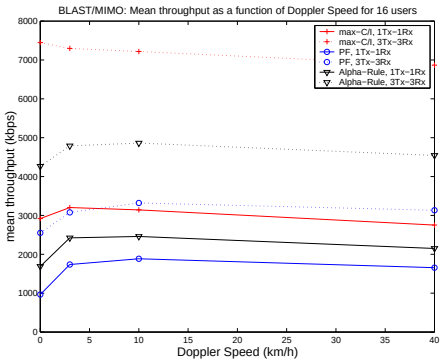

Fig. 7. BLAST/MIMO: the mean aggregate throughput as a function of Doppler speed.

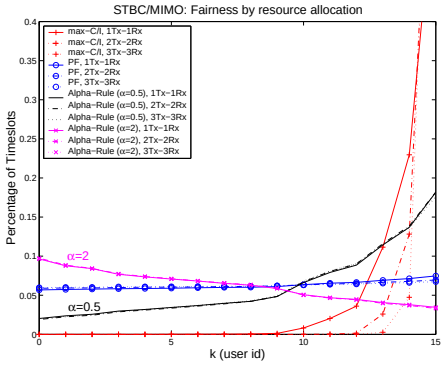

Fig. 8. STBC/MIMO: per-user resource allocation of scheduling schemes for different users.

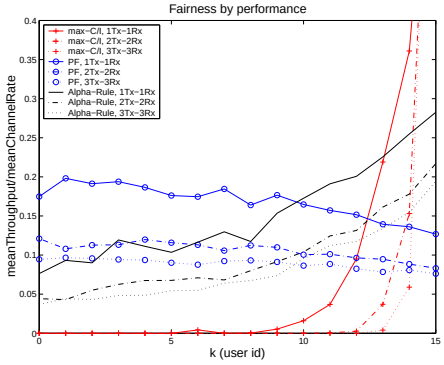

Fig. 9. STBC/MIMO: per-user throughput of different scheduling schemes for different users, where $\alpha = 0.5$ for the Alpha-Rule.

CSI feedbacks; memoryless channel fading across consecutive slots. The first one denotes a perfect channel prediction, while the latter implies the best multiuser diversity. Given the user speed of zero, each channel is static and infinitely correlated. So the first assumption holds but the second fails. With mobility gradually increasing from 0 km/h to 3 km/h and 10 km/h, the channel correlation decreases, which increases the multiuser diversity. This mobility-induced multiuser diversity improves the throughput as the figure shows. When the speed increases further, the inaccuracy in CSI feedbacks starts to deteriorate the throughput. Therefore, the mobility has dual impacts on the throughput by alternating the channel condition between the two assumptions. However, the throughput degrades in Fig. 5 is acceptable for speeds up to 40 km/h, reflecting the robustness of the HDR system to a degree of mobility.

Scheduling over BLAST/MIMO channel: Regardless of MIMO techniques, the PF scheduling throughput is also lower than max-C/I, as illustrated by STBC/MIMO before and BLAST/MIMO in Fig. 6. Again the Alpha-Rule can tradeoff fairness for throughput. A comparison between Fig. 6 and Fig. 4 shows that the scheduling throughput over BLAST/MIMO is higher than STBC/MIMO due to higher transmission capacity and more multiuser diversity. In particular, the full-rate codes in BLAST/MIMO are not limited to 1×1 to $2 \times n_R$, e.g., the throughput above the 3×3 BLAST/MIMO channel is actually the highest given the same scheduling scheme. The results show that BLAST/MIMO may be a better choice for the high-speed packet data services.

Similar performance with mobility tests given BLAST/MIMO is shown in Fig. 7. An exception is 3×3 max-C/I, comes from the simulation granularity and also the fact that with a highly bursty 3×3 BLAST channel, the CSI inaccuracy easily dominates the impacts of mobility on the aggregate throughput.

Fairness versus throughput: To check the resource-based or performance-based fairness, we randomly pick 16 users of different location, i.e., different mean SNR (ρ). Their mean supportable rate ranges from 153kbps to 3.767Mbps. Then we label them by the mean SNR: users closer to the BS have a bigger IDs. Note that a strict fairness is represented by equivalent fairness indices among all users.

Our results show that the fairness performance depends on the scheduling rather than the physical-layer transmission schemes. Take STBC/MIMO for example. Both Fig. 8 and 9 reveal the extreme unfairness of max-C/I scheme: users of small IDs (less than 9) are starved for resource and their throughput is actually zero. Comparatively PF has approximately equal fairness indices among all users for any antenna array, e.g., 1Tx-1Rx, 2Tx-2Rx, and 3Tx-3Rx. The Alpha-Rule

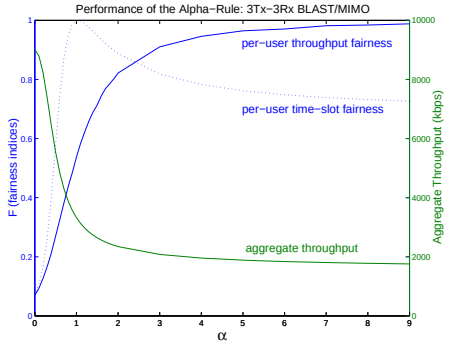

Fig. 10. Alpha-Rule: the performance of aggregate throughput and fairness as a function of α.

with $\alpha = 0.5$ falls in between PF and max-C/I. With a larger α, say $\alpha = 2$, resource allocation is biased towards users of weak channels in an effort to equalize their potential transmission delay with users of strong channels.

The tuneable α controls the flexibility of the Alpha-Rule in the tradeoff between throughput and fairness. The sensitivity is seen from Fig. 10, where 3Tx-3Rx BLAST/MIMO is adopted. As explained in Section III-B, the aggregate throughput decreases monotonically with α because more time slots will be assigned to users of weak channels. For the same reason the throughput-based fairness indices increase monotonically with α. When $\alpha < 1$, the slot allocation favors users of strong channels. An extreme case is the max-C/I ($\alpha = 0$) that has the highest aggregate throughput but the worst fairness in terms of both per-user throughput and per-user time-slot allocation. When $\alpha = 1$ (PF), the time-slot allocation is the fairest ($F = 1$). This is a well-known property of the PF scheduling. When $\alpha > 1$, the slot allocation favors the users of weak channels and as a result, the throughput-based fairness also improves. When α goes to infinity, we can expect that all users achieve the same throughput regardless of their channel condition and the Max-Min fairness is satisfied. We note that the most effective tradeoffs between the aggregate throughput and fairness occur for α less than 2, where slight fairness can be traded for much improvement in the aggregate throughput.

VI. Conclusions

In this paper, we evaluated the opportunistic scheduling algorithms given STBC or BLAST MIMO channels in high-rate packet data cellular systems. We also proposed a flexible and more generic scheduling scheme, the Alpha-Rule, for high-rate packet data services. Compared to the existing PF and max-C/I schemes, the Alpha-Rule enables MAC-layer adaptations to physical-layer channel and flexible online tradeoff between throughput and fairness. In addition, a cross-layer evaluation of the MAC-layer scheduling and the physical-layer MIMO schemes reveals the role of multiuser diversity and antenna diversity in determining the global system performance. By our evaluations, the amount of multiuser diversity depends on the channel characteristics while the gain depends on the combination of scheduling policy and transmission technique. Comparatively, STBC/MIMO is more suitable for point-to-point dedicated channels of voice or mission critical services, while BLAST/MIMO better supports the high-speed downlink shared channel for packet data services.

References

[1] P. Bender et al., "CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, pp. 70–77, July 2000.

[2] 3GPP Technical Specification 25.308 version 5.2.0, *High Speed Downlink Packet Access (HSDPA): Overall description*, March 2002.

[3] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Trans. Networking*, vol. 10, no. 4, pp. 477–486, Aug. 2002.

[4] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998.

[5] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 41–59, Autumn 1996.

[6] I. E. Telatar, "Capacity of multi-antenna gaussian channels," *European Trans. on Telecommun.*, vol. 10, pp. 585–595, Nov.-Dec. 1999.

[7] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inform. Theory*, vol. 45, no. 5, pp. 1456–1467, July 1999.

[8] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-blast: An architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proc. Int. Symp. Sig. Sys. Elect. (ISSSE)*, Pisa, Italy, Sept. 1998.

[9] G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," *IEEE J. Select. Areas Commun.*, vol. 17, no. 11, pp. 1841–1852, Nov. 1999.

[10] R. Knopp and P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, June 1995, pp. 331–335.

[11] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE Veh. Technol. (VTC)*, May 2000, pp. 1854–1858.

[12] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. on Inform. Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[13] D. Bertsekas and R. Gallagar, *Data Networks*, Prentice-Hall, Englwood Cliffs, N.J., 1992.

[14] B. Hassibi and B. M. Hochwald, "High-rate codes that are linear in space and time," *IEEE Trans. Info. Theory*, vol. 48, no. 7, pp. 1804–1824, July 2002.

[15] S. Shenker, "Fundamental design issues for the future internet," *IEEE J. Select. Areas Commun.*, vol. 13, no. 7, pp. 1176–1188, Sept. 1995.

[16] 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, no. 10, pp. 2053–2064, Oct. 2001.

[17] A. Sang, H. Zhu, and S.-q. Li, "Weighted fairness guarantee for scalable diffserv assured forwarding," in *Proc. IEEE Int. Conf. Commun. (ICC)*, June 2001, pp. 2365–2369.

[18] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237–252, July 1998.

[19] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Networking*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[20] D. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, vol. 17, pp. 1–14, 1989.

[21] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyer, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, pp. 150–154, Feb. 2001.

[22] S. Shakkottai and A. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," in *Proc. 17th Int. Teletraffic Congress (ITC-17)*, Sept. 2001.

[23] H. A. David, *Order Statistics, 2nd Ed.*, John Wiley & Sons Inc., 1981.