# QoS-Guaranteed Multiuser Scheduling in MIMO Broadcast Channels

Seung Hwan Lee, John S. Thompson, and Jin-up Kim

**This paper proposes a new multiuser scheduling algorithm that can simultaneously support a variety of different quality-of-service (QoS) user groups while satisfying fairness among users in the same QoS group in MIMO broadcast channels. Toward this goal, the proposed algorithm consists of two parts: a QoS-aware fair (QF) scheduling within a QoS group and an antenna trade-off scheme between different QoS groups. The proposed QF scheduling algorithm finds a user set from a certain QoS group which can satisfy the fairness among users in terms of throughput or delay. The antenna trade-off scheme can minimize the QoS violations of a higher priority user group by trading off the number of transmit antennas allocated to different QoS groups. Numerical results demonstrate that the proposed QF scheduling method satisfies different types of fairness among users and can adjust the degree of fairness among them. The antenna trade-off scheme combined with QF scheduling can improve the probability of QoS-guaranteed transmission when supporting different QoS groups.**

**Keywords: Multiple input multiple output (MIMO), broadcast channel, dirty paper coding (DPC), weighted sum-rate capacity, quality-of-service (QoS), multiuser scheduling, fairness.**

Manuscript received Jan. 12, 2009; revised June 2, 2009; accepted June 11, 2009.

Seung Hwan Lee (phone: +82 42 860 3876, email: lsh@etri.re.kr) and Jin-up Kim (email: jukim@etri.re.kr) are with Broadcasting & Telecommunications Convergence Research Laboratory, ETRI, Daejeon, Rep. of Korea.

John S. Thompson (email: john.thompson@ed.ac.uk) is with Institute for Digital Communications, University of Edinburgh, Edinburgh, UK.

doi:10.4218/etrij.09.0109.0017

## I. Introduction

In multiuser multiple-input multiple-output (MIMO) broadcast channels (BCs), that is, the channels from a base station to mobile users, dirty paper coding (DPC) [1] can achieve the sum-rate capacity [2]. However, because DPC is difficult to implement in real systems due to its excessively high complexity, practical precoding techniques such as Tomlinson-Harashima precoding (THP) and zero-forcing beamforming (ZF-BF) have been developed. Zero-forcing dirty-paper coding (ZF-DPC) [3] is a nonlinear suboptimal implementation of DPC based on interference pre-subtraction. THP applies a modulo operation to ZF-DPC to prevent a possibly large power increase due to the pre-subtraction of ZF-DPC. Linear ZF-BF is a simpler method than THP because no user ordering for pre-subtraction is required at the transmitter. Throughput comparisons between these MIMO precoding techniques can be found in [4] and [5].

Alongside the choice of the MIMO precoding technique, multiuser scheduling is one of the most important issues of a MIMO BC for satisfying different quality-of-service (QoS) requirements, such as throughput, delay constraint, and fairness among users. Many scheduling methods have been suggested [6]-[8]. A round robin (RR) is a simple algorithm that serves users in a cyclic fashion regardless of the channel conditions. Proportional fair (PF) scheduling [6] is designed to meet long-term throughput fairness among users by simultaneously considering the wireless channel conditions and the amount of past throughput in order to exploit the multiuser diversity effect. The modified largest weighted delay first (M-LWDF) [8] is a throughput optimal scheduling strategy that takes into account the channel conditions and state of the queues. With M-LWDF, different QoS requirements are satisfied in terms of the outage

  

probability.

A number of studies focusing on combined MIMO precoding techniques and multiuser scheduling methods for a MIMO BC have been carried out. In [9], the performance of ZF-BF with several scheduling algorithms with a multiuser selection algorithm of reduced-complexity has been analyzed. Joint ZF-BF and scheduling for optimal throughput with reduced complexity was considered in [10]. However, these studies do not consider different QoS requirements, such as different data rates, delay constraints, and different fairness requirements among users.

In this paper, a new QoS-guaranteed multiuser scheduling algorithm is proposed to simultaneously support a variety of different QoS groups while satisfying fairness among users in realistic MIMO BC scenarios, where a base station provides a variety of services to different users with different QoS requirements and fairness considerations. The proposed algorithm includes new fairness metrics for the achievement of different throughput or delay fairness among users in the same QoS group and is combined with the antenna trade-off scheme for QoS differentiation between different QoS groups. With the antenna trade-off scheme, a higher priority group takes precedence in using multiple antennas to satisfy its QoS requirement. The number of transmit antennas allocated to each QoS group can be determined by the wireless channel conditions and QoS requirements. After the number of transmit antennas allocated to each QoS group is determined using the antenna trade-off scheme, the proposed QoS-guaranteed multiuser scheduling algorithm finds user sets from the highest QoS group sequentially using the QoS-aware fair (QF) scheduling algorithm from each QoS group so that the final user set for transmission consists of different QoS users who can maximize the total throughput and satisfy the fairness requirements.

The rest of this paper is organized as follows. In section II, the system model is introduced. In section III, the proposed fairness scheduling algorithms are presented. Section IV briefly explains the antenna trade-off scheme used in this paper. Simulation results are presented in section V. Conclusions are drawn in section VI.

## II. System Model

We use boldface to denote matrices and vectors. For any general matrix $\mathbf{A}$, $\mathbf{A}^T$ denotes the transpose, $\mathbf{A}^H$ denotes the conjugate transpose, and Tr($\mathbf{A}$) denotes the trace. Also, $E[\cdot]$ denotes expectation. For any general set $B$, $|B|$ denotes the cardinality of the set.

Consider a narrowband temporally-correlated fading MIMO BC with $M_T$ transmit antennas at a base station and $K$ ($K \geq M_T$) users each with a single receive antenna. Let $\mathbf{h}_k(t) \in \mathbb{C}^{M_T \times 1}$ denote the channel at time $t$ between the transmit antenna array and the receive antenna for user $k$. The MIMO BC at time $t$ can then be represented as in [11] as

$$y_k(t) = \mathbf{h}_k^T(t)\mathbf{s}(t) + z_k(t), \qquad k = 1, \cdots, K, \tag{1}$$

where $\mathbf{s}(t) \in \mathbb{C}^{M_T \times 1}$ is the transmit signal vector with a power constraint $\mathrm{Tr}\left(E\left[\mathbf{s}(t)\mathbf{s}(t)^H\right]\right) \leq P$, $y_k(t)$ is the received signal for user $k$, and $z_k(t)$ is complex additive white Gaussian noise (AWGN) with zero mean and unit variance for user $k$ at time instant $t$. The temporal correlation of user $k$ with delay $\tau$ can be given by $\rho_k(\tau) = J_0\left(2\pi f_{\mathrm{d}}^k \tau\right)$, where $f_{\mathrm{d}}^k$ is Doppler spread and $J_0$ is the 0th-order Bessel function of the first kind [12]. The Rayleigh fading model [13] generates independent and identically distributed (i.i.d.) complex Gaussian random variables for the elements of $\mathbf{h}_k(t)$ with zero mean and unit variance.

## III. QF Scheduling Algorithms

### 1. Fairness Metrics for QF Scheduling

In real MIMO BC scenarios, a base station is likely to provide a variety of services to different users, each with different QoS requirements. In this case, throughput fairness among users does not mean the allocation of the same amount of bandwidth to all users.

To support such heterogeneous user channels, define the scaled deviation as

$$\Delta x = \frac{x - \bar{x}}{x}, \tag{2}$$

where $x$ is the observation and $\bar{x}$ is the required value of $x$. This can represent the relative degree of fairness regardless of the resource being considered. For example, there are two users with throughput requirements $R_1$ and $R_2$. It is assumed that the throughput requirement of the second user $X_2$ is twice as large as that of the first user $X_1$, that is, $R_2 = 2R_1$. If the data rates (observations) of $X_1$ and $X_2$ are $0.5R_1$ and $R_1$, respectively, at a certain instant, the scaled deviations with respect to the throughput requirements (the required value) are $\Delta x_1 = (0.5R_1 - R_1)/(0.5R_1) = -1.0$ for $X_1$ and $\Delta x_2 = (R_1 - 2R_1)/R_1 = -1.0$ for $X_2$. In this case, throughput fairness is said to be satisfied in terms of relative achievement.

To apply the concept of scaled deviation to the QF criterion, an exponential function taking the argument $\Delta x$ is used. According to the type of resource, either $\exp(\Delta x)$ or $\exp(-\Delta x)$ can be used as elements of the weight vector for a weighted sum-rate maximization rule. In the case of throughput fairness, $\exp(-\Delta x)$ is used so that any user with relatively smaller throughput than other users has a high weight value. On the

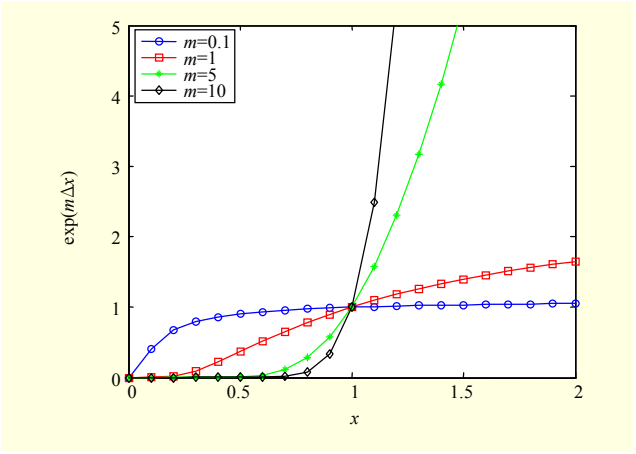


Fig. 1. Characteristics of scaled exponential function exp($m\Delta x$) with slope parameter $m$. $\bar{x} = 1$.

contrary, in the case of delay fairness, exp($\Delta x$) is used so that any user with relatively larger delay than other users has a high weight value.

Prior to applying the scaled deviation to fairness cases, let us introduce the other parameter, $m$, which changes the slope of the exponential function with a form of exp($m\Delta x$) or exp($-m\Delta x$). Later, this slope parameter is used to control the degree of fairness. Figure 1 shows the characteristics of the scaled exponential function exp($m\Delta x$) with $\bar{x} = 1$.

When $m = 0.1$, the output of the scaled exponential function is saturated near 1 when $x > 1$ regardless of the scaled deviation. However, as $m$ increases, the output of the scaled exponential function grows rapidly when $x > 1$. This makes a small change in $x$ cause a significant increase in the output.

For throughput fairness, define the throughput fairness metric $\mu_k^{\mathrm{t}}(t)$ for user $k$ at time $t$ as

$$\mu_k^{\mathrm{t}}(t) = \exp\left( m_{\mathrm{t}} \frac{\bar{R}(t) - a_k^{\mathrm{t}} \bar{r}_k(t)}{\varepsilon + a_k^{\mathrm{t}} \bar{r}_k(t)} \right), \tag{3}$$

where $m_{\mathrm{t}}$ ($m_{\mathrm{t}} \geq 0$) is the throughput slope of the exponential function, which determines the sensitivity of (2) toward throughput fairness. As might be expected from Fig. 1, smaller values of $m_{\mathrm{t}}$ mean less strict throughput fairness among users and typically give them only average throughput fairness. Indeed, when $m_{\mathrm{t}} = 0$, the weighted sum-rate maximization problem reduces to the normal sum-rate maximization problem without any fairness consideration. The scalars $a_k^{\mathrm{t}}$ are constants to allow different throughput requirements among users. The variable $\varepsilon$ is an appropriate small value for ensuring that the denominator is nonzero. The value $\bar{r}_k(t)$ is the exponential moving average of the past throughput for user $k$, which is updated as

$$\bar{r}_k(t+1) = \left(1 - \frac{1}{\alpha_{\mathrm{t}}}\right) \bar{r}_k(t) + \frac{1}{\alpha_{\mathrm{t}}} R_k(t), \tag{4}$$

where $\alpha_{\mathrm{t}}$ ($\alpha_{\mathrm{t}} > 0$) is the smoothing factor, and $R_k(t)$ is the data rate of user $k$ at time $t$. The scalar $\bar{R}(t)$ in (3) is defined as $\bar{R}(t) = (1/K)\sum_{k=1}^{K} \bar{r}_k(t)$. According to (3), any user with $a_k^{\mathrm{t}} \bar{r}_k(t) > \bar{R}(t)$ has a value less than 1 as its weight. In this case, that user might be excluded from a selected user set for transmission by the weighted sum-rate maximization rule for meeting the throughput fairness constraint among users despite its high channel gain.

For delay fairness, define the delay fairness metric $\mu_k^{\mathrm{d}}(t)$ for user $k$ at time $t$ as

$$\mu_k^{\mathrm{d}}(t) = \exp\left( m_{\mathrm{d}} \frac{a_k^{\mathrm{d}} d_k(t) - \bar{D}(t)}{\varepsilon + a_k^{\mathrm{d}} d_k(t)} \right), \tag{5}$$

where $m_{\mathrm{d}}$ ($m_{\mathrm{d}} \geq 0$) is the delay slope of the exponential function, which determines the sensitivity of delay fairness as in the case of throughput fairness. Again, small values of $m_{\mathrm{d}}$ mean less strict delay fairness among users. The scalars $a_k^{\mathrm{d}}$ are constants to allow different delay requirements among users. The scalar $d_k(t)$ denotes the head-of-line delay in the queue of user $k$, and $\bar{D}(t)$ is defined as $\bar{D}(t) = (1/K)\sum_{k=1}^{K} d_k(t)$.

## 2. Weighted Sum-Rate Maximization with THP

For the THP technique in a MIMO BC, a channel matrix $\mathbf{H}(S_0)$ is formed with a user set $S_0$ ($|S_0| \leq M_{\mathrm{T}}$) and decomposed into unitary transmit beamforming matrix $\mathbf{F}$ and lower triangular matrix $\mathbf{B}$ by taking the QR decomposition. Because of the lower triangular matrix $\mathbf{B}$, any interference caused by users $j > i$ on each user $i$ is forced to zero by pre-subtraction at the transmitter. However, due to the pre-subtraction, the transmit power increases, so that THP employs the modulo operation to minimize this situation. Denoting the transmit power allocated to user $k$ as $P_k$ and $b_{kk}$ as the $k$-th diagonal element of matrix $\mathbf{B}$, the achievable sum-rate capacity of THP is given by

$$\begin{aligned} C(S_0) &= \max \sum_{k \in S_0} \log_2 \left( 1 + \frac{b_{kk}^2 P_k}{\Gamma_{\mathrm{THP}}^k} \right), \\ &\text{subject to } P_k \geq 0, \sum_{k \in S_0} P_k \leq P, \end{aligned} \tag{6}$$

where $\Gamma_{\mathrm{THP}}^k$ denotes the modulo loss of user $k$ [14]. For simplicity, several assumptions are made before applying the achievable sum-rate capacity to the weighted sum-rate maximization rule. First, although the maximum sum-rate capacity of THP can be achieved by the optimal transmit power allocation [15], these optimal strategies are ignored for simplicity. Therefore, an equal power allocation over spatial channels is assumed. Second, the number of elements in a user set $S$ is assumed to be the same as that of transmit antennas

($|S| = M_T$). Finally, if the target bit-error rate (BER) of the system is very small (that is, BER$\leq 10^{-6}$), and a high SNR is assumed for all users, the modulo loss can be ignored [16] except for the shaping loss of 1.53 dB, which can be achieved by using multidimensional lattice codes rather than $M$-QAM modulation. Then, the sum-rate capacity of THP at time $t$ is approximated as

$$C_{\mathrm{THP}}(S,t) = \sum_{k=1}^{M_T} \log_2\left(1 + \frac{b_{kk}^2(t)P}{M_T}\right). \quad (7)$$

From (3), (5), and (7), the weighted sum-rate maximization rule considering throughput and delay fairness with THP at time $t$ can be respectively given as

$$S_{\max}^{\mathrm{t}}(t) = \arg \max_{S\subset U, |S|=M_T} \sum_{k=1}^{M_T} \mu_k^{\mathrm{t}}(t) \log_2\left(1 + \frac{b_{kk}^2(t)P}{M_T}\right), \quad (8)$$

$$S_{\max}^{\mathrm{d}}(t) = \arg \max_{S\subset U, |S|=M_T} \sum_{k=1}^{M_T} \mu_k^{\mathrm{d}}(t) \log_2\left(1 + \frac{b_{kk}^2(t)P}{M_T}\right), \quad (9)$$

where $U$ is the total user set, that is, $U = \{u_k \mid k = 1,\cdots, K\}$.

The weighting used for the existing PF algorithm for user $k$ is the inverse of the exponential moving average of the past throughput so that the weighted sum-rate maximization rule for throughput fairness with the PF algorithm is given by

$$S_{\max}^{\mathrm{t}}(t) = \arg \max_{S\subset U, |S|=M_T} \sum_{k=1}^{M_T} \frac{1}{\overline{r}_k(t)} \log_2\left(1 + \frac{b_{kk}^2(t)P}{M_T}\right). \quad (10)$$

For delay fairness with the existing LWDF rule, the weighted sum-rate maximization rule is of the form

$$S_{\max}^{\mathrm{d}}(t) = \arg \max_{S\subset U, |S|=M_T} \sum_{k=1}^{M_T} a_k^{\mathrm{d}} d_k(t) \log_2\left(1 + \frac{b_{kk}^2(t)P}{M_T}\right). \quad (11)$$

In [7], the constants $a_k^{\mathrm{d}}$ are defined by probabilistic QoS requirements. It is useful to note that the time index $t$ is omitted to simplify equations in the next section.

If there are two or more different QoS groups, the QoS-guaranteed multiuser scheduling algorithm determines the number of transmit antennas allocated to each QoS group. Then, the QF scheduling algorithm is applied at each stage to find a user set for that QoS group. In this case, the number of selected users from each QoS group is the number of transmit antennas allocated to each QoS group, which is determined by the antenna trade-off scheme introduced in the next section.

## IV. Antenna Trade-off between Different QoS Groups

In this section, the antenna trade-off scheme in [17] is briefly introduced for the proposed QoS-guaranteed multiuser scheduling algorithm. We assume that there are two QoS groups in a MIMO BC, namely, a best effort (BE) user set $U_{\mathrm{BE}} = \{u_k \mid k = 1,\cdots, K_{\mathrm{BE}}\}$ and a delay-constrained real-time (RT) user set $U_{\mathrm{RT}} = \{u_k \mid k = 1,\cdots, K_{\mathrm{RT}}\}$, where $K_{\mathrm{BE}}$ and $K_{\mathrm{RT}}$ are the number of BE and RT users, respectively, and $K_{\mathrm{BE}}+K_{\mathrm{RT}} = K$. It is assumed that RT users have a higher priority than BE users in terms of QoS requirements. The objective of any multiuser scheduling algorithm in this configuration is to satisfy the delay constraint of RT users while maximizing the total throughput of all users. A simple multiuser scheduling algorithm may reserve a certain fixed number of transmit antennas to each group to support BE and RT users simultaneously. A more advanced way to serve different QoS groups is to change the number of transmit antennas assigned to each QoS group according to certain criteria. This can increase the degree of adaptability to fading MIMO channels. For this adaptive scheme, each QoS group has its set of pre-assigned transmit antennas, but there is a priority for their actual use in transmission. RT users are assumed to have higher priority than BE users in using multiple transmit antennas. For example, when the average delay of RT users is within a certain delay threshold, RT and BE users use their pre-assigned transmit antennas for transmission. When the average delay of RT users exceeds the delay threshold, all transmit antennas are assigned to RT users to reduce the average delay of RT users. When this happens, BE users have no chance of using their pre-assigned antennas for transmission.

The antenna trade-off scheme determines the number of transmit antennas for RT users $M_{\mathrm{RT}}$ according to

$$M_{\mathrm{RT}} = M_{\mathrm{S}} + (M_T - M_{\mathrm{S}}) \cdot \mathrm{u}\left[\overline{D}_{\mathrm{RT}} - D_{\mathrm{TH}}\right], \quad (12)$$

where $\overline{D}_{\mathrm{RT}}$ denotes the average delay of RT users, $D_{\mathrm{TH}}$ denotes the delay threshold of RT users, $M_{\mathrm{S}}$ ($M_{\mathrm{S}} < M_T$) denotes the pre-assigned groups of antennas for RT users, and u[·] denotes the unit step function.

After the number of transmit antennas for each QoS group is determined, the proposed algorithm selects a user set from each QoS group sequentially. It is assumed that the numbers of selected BE and RT users are the same as the numbers of determined transmit antennas for BE and RT users, respectively. For the selection of RT users, QF scheduling with delay fairness among RT users is employed. Similarly, for the selection of BE users, QF scheduling with throughput fairness among BE users is considered. In other words, the proposed multiuser scheduling algorithm in this configuration not only considers throughput and delay fairness among BE and RT users respectively, but also includes the delay constraint of RT users. The proposed multiuser scheduling algorithm uses the antenna trade-off scheme to determine the number of transmit antennas for each QoS group and the weighted sum-rate maximization rules to select an optimal user set with fairness

considerations. The weighted sum-rate maximization rule with delay fairness finds a user set $S_{RT}^{d}$ among all possible user sets $S_{RT}$ ($S_{RT} \subset U_{RT}$, $|S_{RT}| = M_{RT}$) according to

$$S_{RT}^{d} = \arg\max_{S_{RT}} \sum_{k=1}^{M_{RT}} \mu_k^{d} \log_2 \left(1 + \frac{b_{kk}^2 P}{M_{RT}}\right). \quad (13)$$

For the selection of BE users, the weighted sum-rate maximization rule with throughput fairness finds a user set $S_{BE}^{t}$ among all possible user sets $S_{BE}$ ($S_{BE} \subset U_{BE}$, $|S_{BE}| = M_{BE}$) and the already selected RT user set $S_{RT}^{d}$ according to

$$S_{BE}^{t} = \arg\max_{S_{BE}} \left\{ \sum_{k=1}^{M_{BE}} \mu_k^{t} \log_2 \left(1 + \frac{b_{kk}^2 P}{M_{BE}}\right) + C\left(S_{RT}^{d}\right) \right\}, \quad (14)$$

where $C(S_{RT}^{d})$ is the weighted sum-rate capacity of the already selected RT users, which is of the form

$$C\left(S_{RT}^{d}\right) = \sum_{i \in S_{RT}^{d}} \mu_i^{d} \log_2 \left(1 + \frac{b_{ii}^2 P}{M_{RT}}\right). \quad (15)$$

After the user sets are selected from each QoS group, THP is performed on the final user set $S_{max}$ ($S_{max} = S_{RT}^{d} \cup S_{BE}^{t}$) for transmission.

## V. Simulation Results

Throughout the simulations, we assume the number of transmit antennas $M_T = 4$, the transmit power $P = 10$, and the Doppler spread $f_d = 10$ Hz for all users. The data arrivals in the queues for all users are assumed to be independent Poisson processes with arrival rate $\lambda$. If the queue dropping probability due to the queue overflow and the error rate through the MIMO BC are negligible, the throughput can be determined by the arrival rate and the sum-rate capacity. When the arrival rate is less than or equal to the sum-rate capacity, it defines the throughput. However, when the data arrival rate is greater than the sum-rate capacity, the sum-rate capacity determines the throughput.

Several fairness indices have been suggested for the measurement of fairness for different resource allocation schemes. The Jain fairness index [18] is one quantitative measure of fairness, which is given by

$$I_{Jain} = \frac{\left(\sum_{k=1}^{K} \gamma_k\right)^2}{K \sum_{k=1}^{K} \gamma_k^2}, \quad (16)$$

where $\gamma_k$ is the fraction of transmission resource allocated to user $k$.

### 1. Throughput Fairness Using QF Scheduling Algorithm

Figures 2 and 3 compare the throughput and Jain fairness index of the PF algorithm and the proposed QF algorithm for different smoothing factors $\alpha_t$ against throughput slope $m_t$. The number of users is $K = 8$, and the constants $a_k^t$ are assumed to be 1. We notice that PF-THP and QF-THP have the same Jain fairness index when the average throughout of each scheme is identical (cross-over point). For example, when $\alpha_t = 10$, these two algorithms have the same average throughput of about 10.9 [bits/channel use] with a fairness index of about 0.90. This means that these two schemes achieve basically the same performance in terms of average throughput and fairness. However, the proposed QF algorithm is able to control the degree of fairness by trading the average throughput for fairness among users. The larger $m_t$ is, the higher the throughput fairness is among users and the lower the throughput. The smoothing factor $\alpha_t$ determines the time

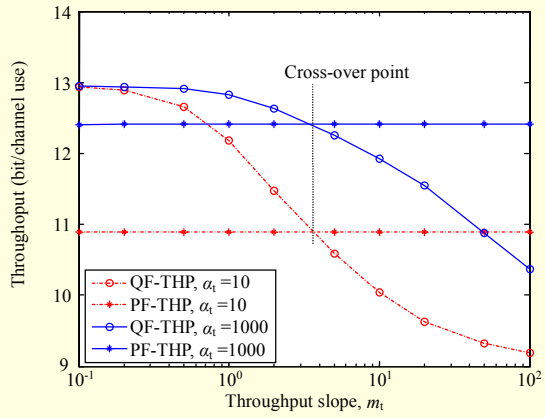


Fig. 2. Comparison of the PF algorithm and the proposed QF algorithm in terms of average throughput for different smoothing factors against the throughput slope $m_t$. $M_T = 4$, $K = 8$, $P = 10$ dB, $a_k^t = 1$, and $f_d = 10$ Hz.

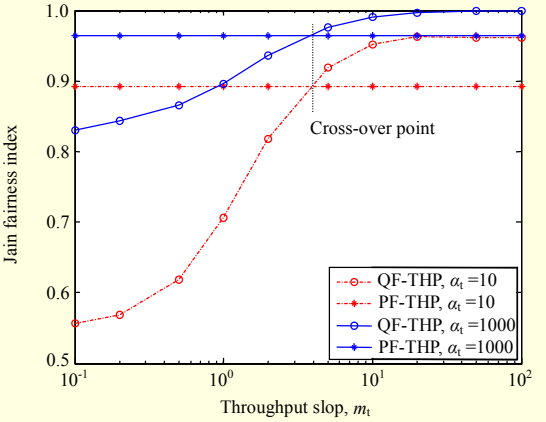


Fig. 3. Comparison of the PF algorithm and proposed QF algorithm in terms of the Jain fairness index for different smoothing factors against the throughput slope $m_t$. $M_T = 4$, $K = 8$, $P = 10$ dB, $a_k^t = 1$, and $f_d = 10$ Hz.

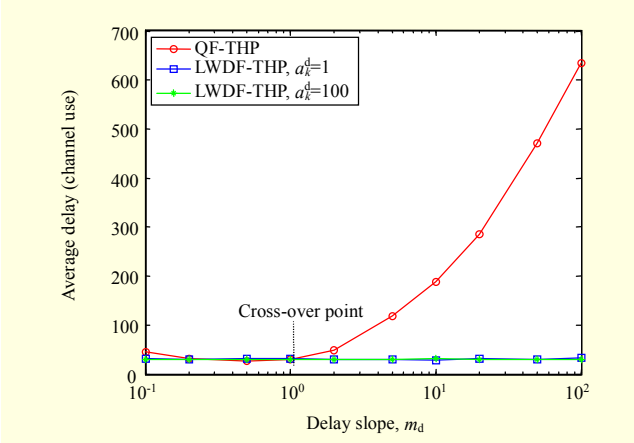


Fig. 4. Comparison of the LWDF algorithm and the proposed QF scheduling algorithm in terms of average delay against the delay slope $m_d$. $M_T = 4$, $K = 8$, $P = 10$ dB, $a_k^d = 1$, $f_d = 10$ Hz, and $\lambda$=1.25.

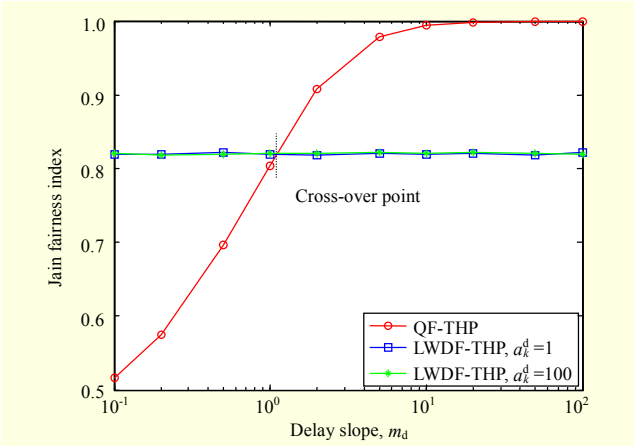


Fig. 5. Comparison of the LWDF algorithm and the proposed QF scheduling algorithm in terms of the Jain fairness index against the delay slope $m_d$. $M_T = 4$, $K = 8$, $P = 10$ dB, $a_k^d = 1$, $f_d = 10$ Hz, and $\lambda$=1.25.

window size for fairness. Larger time windows provide more flexibility than smaller windows in selecting users due to multiuser diversity. This increases the average throughput of all users. We notice that the case of $\alpha_t = 1{,}000$ has a larger average throughput and fairness index than the case of $\alpha_t = 10$.

## 2. Delay Fairness Using QF Scheduling Algorithm

Figures 4 and 5 compare the average delay and fairness, respectively, of the LWDF algorithm and proposed QF algorithm for different smoothing factors $\alpha_d$ against the delay slope $m_d$. The number of users is $K = 8$ and the constants $a_k^d$ are assumed to be 1. An arrival rate of $\lambda$=1.25 [bits/channel use] is assumed for all users. Unlike the LWDF algorithm, the

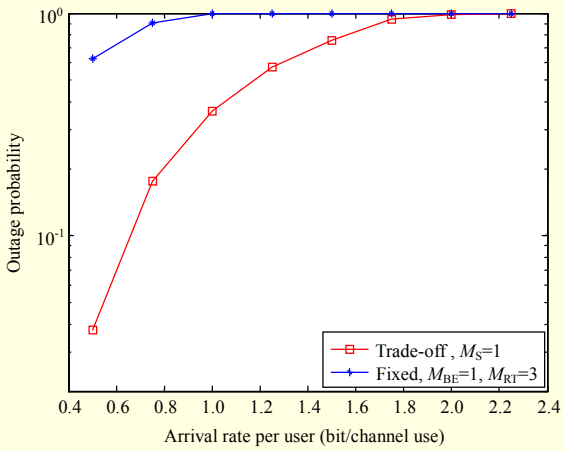


Fig. 6. Outage probability of the average delay of RT users exceeding delay threshold $D_{TH}$ for different antenna configurations. $M_T = 4$, $K_{RT} = 4$, $K_{BE} = 12$, $P = 10$ dB, $a_k^t = 1$, $a_k^d = 1$, $D_{TH} = 50$ , $\alpha_t = 1000$, and $f_d = 10$ Hz.

proposed QF algorithm can adjust the degree of fairness by changing the delay slope $m_d$. The average delay and fairness indexes of the LWDF algorithm are the same regardless of the constants $a_k^d$. Similar to the case of throughput fairness, the performance of these two algorithms in terms of the average delay and fairness index is the same when $m_d \approx 1$. The delay fairness metric increases the degree of fairness by increasing the average delay of all users. This results in a decrease of the average throughput of all users.

## 3. Antenna Trade-off between Different QoS Groups

For a performance analysis of the proposed QF scheduling algorithm with the antenna trade-off scheme, it is assumed that there are 4 RT users ($K_{RT} = 4$) and 12 BE users ($K_{BE} = 12$) for each QoS group. The pre-assigned transmit antenna to RT users, $M_S$, is assumed to be 1. For throughput fairness, $\alpha_t = 1{,}000$ is also assumed. All users in the BE and RT groups have the same throughput and delay requirements. For the delay threshold for RT users, $D_{TH} = 50$ [channel use] is assumed.

Figures 6 and 7 show the outage probability of the average delay of RT users exceeding the delay threshold $D_{TH}$ and the average throughput with fairness considerations for different transmit antenna configurations, respectively.

As seen in Fig. 6, the proposed algorithm with the antenna trade-off scheme performs better than that without the antenna-trade-off scheme (labeled 'Fixed') in terms of outage probability because it can adapt to the number of transmit antennas assigned to RT users according to the MIMO channel conditions. This enables the multiuser scheduling algorithm to minimize the outage probability of the average delay of RT

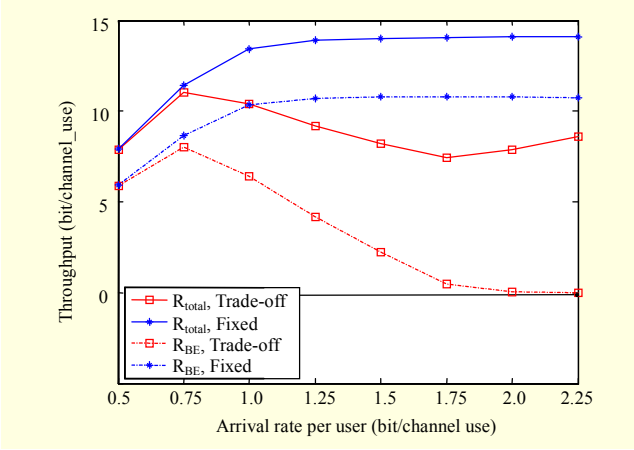


Fig. 7. Throughput performance for different antenna configurations, $M_T = 4$, $K_{RT} = 4$, $K_{BE} = 12$, $P = 10$ dB, $a_k^t = 1$, $a_k^d = 1$, $D_{TH} = 50$ , $\alpha_t = 1000$, and $f_d = 10$ Hz.

users exceeding delay threshold $D_{TH}$. However, the fixed scheme performs better than the antenna trade-off scheme in terms of the total throughput of all users ($R_{total}$) and that of BE users ($R_{BE}$) because it can always include BE users regardless of the delay status of RT users. This maximizes the multiuser diversity gain. Nevertheless, the throughput results of the fixed scheme are misleading because it cannot satisfy the delay constraints of RT users, unlike the antenna trade-off scheme. The throughput of BE users with the antenna trade-off schemes is almost zero when the arrival rate is greater than 2 bit/channel_ use. This is because RT users have higher priority, and they very frequently take more transmit antennas than pre-assigned value $M_S$ in order to satisfy the delay constraint. This results in less chance of transmission for BE users.

## VI. Conclusions

We proposed and analyzed a new QF scheduling to select users in the same QoS group with throughput or delay fairness among users. We applied an antenna trade-off scheme to the proposed QF scheduling algorithm to simultaneously support different QoS users with QoS differentiation. For the selection of a user set with the proposed QF scheduling algorithm, the weighted sum-rate maximization rule is exploited so that the selected user set satisfies the different fairness requirements whilst maximizing the sum-rate capacity. The exponential function with scaled fairness deviation as its argument is used for fairness among users in terms of throughput and delay. The proposed QF scheduling algorithm can also control the degree of fairness by adjusting the slope of the exponential fairness metric. In simultaneously serving BE and RT users with delay constraints, the antenna trade-off scheme performs better than the fixed scheme in terms of both the delay performance of RT users and the throughput performance of all users due to the adaption to the time-varying channels.

However, the computational complexity of the weighted sum-rate maximization rule used by the proposed multiuser scheduling algorithm grows rapidly as the number of users increases. Hence, additional work should be done to decrease the complexity of the proposed algorithm for practical implementation.

## References

[1] M. Costa, "Writing on Dirty Paper," *IEEE Trans. Information Theory*, vol. 29, no. 3, 1983, pp. 439-441.

[2] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, Achievable Rates, and Sum-Rate Capacity of Gaussian MIMO Broadcast Channels," *IEEE Trans. Information Theory*, vol. 49, no. 10, 2003, pp. 2658-2668.

[3] G. Caire and S. Shamai, "On the Achievable Throughput of a Multiantenna Gaussian Broadcast Channel," *IEEE Trans. Information Theory*, vol. 49, no. 7, 2003, pp. 1691-1706.

[4] R. Zhang, J. Cioffi, and Y. Liang, "Throughput Comparison of Wireless Downlink Transmission Schemes with Multiple Antennas," *IEEE Int'l Conf. Communications (ICC)*, vol. 4, 2005, pp. 2700-2704.

[5] T. Yoo and A. Goldsmith, "Sum-Rate Optimal Multi-antenna Downlink Beamforming Strategy Based on Clique Search," *Global Telecommunications Conference (GLOBECOM)*, vol. 3, 2005, pp. 1510-1514.

[6] A. Jalali, R. Padovani, and R. Pankaj, "Data Throughput of CDMA-HDR: A High Efficiency–High Data Rate Personal Communication Wireless System," *IEEE 51st Vehicular Technology Conference*, vol. 3, 2000, pp. 1854-1858.

[7] A. Stolyar and K. Ramanan, "Largest Weighted Delay First Scheduling: Large Deviations and Optimality," *The Annals of Applied Probability*, vol. 11, no. 1, 2001, pp. 1-48.

[8] M. Andrews et al., "CDMA Data QoS Scheduling on the Forward Link with Variable Channel Conditions," *Bell Labs Technical Memorandum*, Apr. 2000.

[9] T. Yoo and A. Goldsmith, "On the Optimality of Multiantenna Broadcast Scheduling Using Zero-Forcing Beamforming," *IEEE J. Selected Areas in Communications*, vol. 24, no. 3, 2006, pp. 528-541.

[10] C. Swannack, E. Uysal-Biyikoglu, and G. Wornell, "Low Complexity Multiuser Scheduling for Maximizing Throughput in the MIMO Broadcast Channel," *Proc. Allerton Conf. Communications, Control and Computing*, 2004.

[11] S. Lee and J. Thompson, "Simulation Models for Investigation of Multiuser Scheduling in MIMO Broadcast Channels," *ETRI Journal*, vol. 30, no. 6, 2008, pp. 765-773.

[12] M. Vu and A. Paulraj, "A Robust Transmit CSI Framework with Applications in MIMO Wireless Precoding," *Proc. 39th Asilomar Conf. Signals, Systems and Computers*, 2005, pp. 623-627.

[13] Y. Zheng and C. Xiao, "Simulation Models with Correct Statistical Properties for Rayleigh Fading Channels," *IEEE Trans. Communications*, vol. 51, no. 6, 2003, pp. 920-928.

[14] C. Fung, W. Yu, and T. Lim, "Precoding for Multi-antenna Downlink: Multiuser SNR Gap and Optimal User Ordering," *IEEE Trans. Communications*, vol. 55, no. 1, 2007, pp. 188-197.

[15] N. Jindal et al., "Sum-Power Iterative Water-Filling for Multi-antenna Gaussian Broadcast Channels," *IEEE Trans. Information Theory*, vol. 51, no. 4, 2005, pp. 1570-1580.

[16] R. Wesel and J. Cioffi, "Precoding and the MMSE-DFE," *28th Asilomar Conf. Signals, Systems and Computers*, vol. 2, 1994, pp. 1144-1148.

[17] S. Lee and J. Thompson, "Trade-Offs of Spatial Gain for QoS-Guaranteed Services in the MIMO Broadcast Channels," *IEEE Int'l Conf. Communications* (*ICC)*, 2007, pp. 4640-4645.

[18] R. Jain, D. Chiu, and W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems," *Digital Equipment Corp. Eastern Research Lab, DEC-TR-301*, 1984.

**Seung Hwan Lee** received the BS and MS degrees from Korea University, Seoul, Korea, in 1995 and 1997, respectively, and the PhD degree from the University of Edinburgh, Edinburgh, UK, in 2007. He has been with the Electronics and Telecommunications Research Institute as a senior member of engineering staff since 2001. His research includes multiuser scheduling, MIMO, cross-layer optimization, and cognitive radio systems.

**John S. Thompson** received his BEng and PhD degrees from the University of Edinburgh, in 1992 and 1996, respectively. From July 1995 to August 1999, he worked as a postdoctoral researcher at Edinburgh, funded by the UK Engineering and Physical Sciences Research Council (EPSRC) and Nortel Networks. Since September 1999, he has been a lecturer at the School of Engineering and Electronics at the University of Edinburgh. In October 2005, he was promoted to the position of reader. His research interests currently include signal processing algorithms for wireless systems, antenna array techniques, and multihop wireless communications. He has published approximately 170 papers to date including a number of invited papers, book chapters, and tutorial talks. He has also co-authored an undergraduate textbook on digital signal processing. He is currently editor-in-chief of the *IET Signal Processing Journal* and was recently a technical program co-chair for the *IEEE International Conference on Communications (ICC) 2007*, held in Glasgow in June 2007.

**Jin-up Kim** received the BS degree from Korea University, Seoul, Korea, in 1985, and the MS and PhD degrees from Korea Advanced Institute of Science and Technology, Seoul, Korea, in 1987 and 1996, respectively. He has been with Electronics and Telecommunication Research Institute since 1987. Also, he has been a professor with the University of Science and Technology in the field of wireless communications since 2005. He has researched in the field of the wireless communication systems. His recent research interests include digital RF, software defined radio, and cognitive radio technologies.