EURASIP Journal on Wireless Communications and Networking

Fairness in Radio Resource Management for Wireless Networks

Guest Editors: Mohamed Hossam Ahmed, Alagan Anpalagan, Kwang-Cheng Chen, Zhu Han, and Ekram Hossain

Fairness in Radio Resource Management for Wireless Networks

EURASIP Journal on
Wireless Communications and Networking

Fairness in Radio Resource Management for Wireless Networks

Guest Editors: Mohamed Hossam Ahmed,
Alagan Anpalagan, Kwang-Cheng Chen, Zhu Han,
and Ekram Hossain

Editor-in-Chief

Associate Editors

Contents

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 642878, 3 pages
doi:10.1155/2009/642878

Editorial

Fairness in Radio Resource Management for Wireless Networks

Mohamed Hossam Ahmed,[1] Alagan Anpalagan,[2] Kwang-Cheng Chen,[3] Zhu Han,[4] and Ekram Hossain[5]

[1] *Electrical and Computer Engineering, Memorial University of Newfoundland, St. John's, NL, Canada A1C 5S7*
[2] *Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada M5B 2K3*
[3] *Electrical Engineering department, National Taiwan University, Taipei 10617, Taiwan*
[4] *Electrical and Computer Engineering, University of Houston, Houston, TX, 77004, USA*
[5] *Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada R3T 2N2*

Correspondence should be addressed to Mohamed Hossam Ahmed, mhahmed@engr.mun.ca

Received 31 December 2008; Accepted 31 December 2008

Radio resource management (RRM) techniques such as admission control, scheduling, subcarrier allocation, channel assignment, power allocation, and rate control are essential for maximizing the resource utilization and providing quality of service (QoS) in wireless networks. In many cases, the performance metrics (e.g., overall throughput) can be optimized if opportunistic algorithms are employed. However, opportunistic RRM techniques always favor advantaged users who have good channel conditions and/or low interference levels. The problem becomes even worse when the wireless terminals have low mobility since the channel conditions become slowly varying (or even static), which might lead to long-term unfairness. The problem of fair resource allocation is more challenging in multihop wireless networks (e.g., wireless mesh networks and multihop cellular networks). This special issue addresses some fairness issues and solutions in using RRM techniques in modern wireless communication systems.

We received an overwhelming response to our call for paper of this special issue. From the large number of high quality submissions we received, we have selected sixteen papers grouped in six subtopics, namely, (1) *fairness of RRM in WiMAX networks*, (2) *fairness of RRM in OFDM/OFDMA systems*, (3) *fairness of RRM in CDMA/UMTS systems*, (4) *fairness of RRM in MIMO systems*, (5) *fairness of RRM in multihop and mesh networks*, and finally (6) *fairness of multiuser resource allocation*.

In the first group, three papers address the fair resource management in WiMAX networks.

The first paper titled "Fair adaptive bandwidth and subchannel allocation in the WiMAX uplink" by A. Morell, G. S. Granados, and J. L. Vicario proposes an uplink scheduling mechanism for mobile WiMAX networks. The scheduling mechanism implements a dynamic bandwidth allocation solution in a network utility maximization framework. The problem is decomposed into two subproblems, namely, a flow allocation subproblem and a subchannel allocation subproblem. To solve the optimization problem, the authors apply the mean value cross-decomposition method, which results in an implementation-friendly solution. The second paper titled "Fairness and QoS guarantees of WiMAX OFDMA scheduling with fuzzy controls" by C. L. Chen et al. proposes a fuzzy control-based scheduling mechanism for WiMAX. The objective of the proposed scheduling mechanism is to provide delay and jitter control for real-time connections, and throughput control for non-real-time connections. The scheduling method provides intra- and interclass fairness with QoS guarantees, and it has low implementation complexity. With intraclass fairness, the connections within the same class achieve equal degree of QoS. With interclass fairness, the connections with QoS requirements achieve their demands and those without QoS requirements equally share the remaining resources.

The third paper titled "CDIT-based constrained resource allocation for mobile WiMAX systems" by F. Brah, J. Louveaux, and L. Vandendorpe addresses the problem of subchannel assignment and power allocation for mobile WiMAX systems. The authors consider a fast fading environment, where the transmitter has only the channel distribution information (CDI) instead of the full instantaneous channel state information. The objective is to maximize the ergodic weighted sum rate under long-term fairness, minimum data rate requirement, and power budget constraints. The authors formulate the problem as a nonlinear stochastic constrained optimization problem and provide an efficient analytical solution based on Lagrange dual decomposition framework. For the proposed CDIT-based resource allocation framework, the trade-off between reduction in computational complexity and performance degradation is analyzed.

The papers in the second group consider the fair resource allocation in OFDM/OFDMA systems.

The first paper titled "Cross-layer resource scheduling for video traffic in the downlink of 4G wireless multicarrier networks" by F. Bokhari et al. presents a cross-layer scheduling scheme which is designed for packet scheduling and resource (subcarrier) allocation in the downlink of 4G wireless multicarrier networks. The authors propose an adaptive method for parameter selection which integrates packet scheduling with resource mapping. The performance of the proposed scheme is compared to that of the Round Robin and the Score-Based schedulers, considering varying interference and network loading conditions in a multicell environment. The authors further analyze the proposed scheme with different fairness indices available in the literature in order to quantify the achieved fairness as compared to the reference schemes.

The second paper titled "Busy bursts for trading-off throughput and fairness in cellular OFDMA-TDD" by B. Ghimire, G. Auer, and H. Haas proposes a decentralized interference management algorithm for OFDMA operating in TDD cellular systems. Interference aware allocation of time-frequency slots is accomplished by letting receivers transmit a busy burst (BB) in a time-multiplexed minislot, upon successful reception of data. A link adaptation method using BB signaling is proposed, where the transmission format is dynamically adjusted based on the channel conditions.

The third paper titled "A fair opportunistic access scheme for multiuser OFDM wireless networks" by C. Gueguen and S. Baey proposes a new access scheme for efficient support of multimedia services in OFDM wireless networks. Access to the medium is granted based on a system of weights that dynamically accounts for both the experienced QoS and the transmission conditions. This new approach enables the full support of multimedia services with the adequate traffic and QoS differentiation while maximizing the system capacity and keeping a special attention on fairness.

In the third group, three papers investigate the fair resource management in CDMA/UMTS networks.

The first paper titled "Decentralized utility maximization in heterogeneous multi-cell scenarios with interference limited and orthogonal air-interfaces" by Ingmar Blau et al. treats the problem of resource allocation in terms of optimum air-interface and cell selection in cellular multi-air-interface scenarios. The adopted model applies to arbitrary heterogeneous scenarios, where the air-interfaces belong to the class of interference limited systems like UMTS or to a class with orthogonal resource assignment such as TDMA-based GSM or WLAN. The performance of the dynamic algorithm is then evaluated for a heterogeneous UMTS/GSM scenario.

The second paper titled "Joint throughput maximization and fair uplink transmission scheduling in CDMA systems" by C. Li and S. Papavassiliou studies the optimal scheduling for uplinks of a code division multiple access wireless system while satisfying the quality of service requirement and maintaining fairness among users. The throughput maximization problem is formulated as a multiconstraint optimization problem and then expressed as a weighted throughput maximization problem under power and QoS weight constraints that nicely relate the fairness. The authors use the concept of power index capacities to convert the problem under investigation to a binary knapsack problem, and then the optimal solution is obtained through a global search mechanism using a two-step approach.

The third paper titled "Spatial and temporal fairness in heterogeneous HSDPA-enabled UMTS networks" by A. Mader and D. Staehle investigates spatial and temporal fairness aspects in HSDPA-enabled UMTS networks for different link level scheduling schemes. Spatial fairness refers to the spatial distribution of perceived data rates among users while temporal fairness refers to the long-term time-average user throughput. A flow-level simulation that is used for this study considers traffic dynamics for both QoS flows and best-effort (or elastic) flows. The impact of network-wide interference and multipath propagation effects is also considered.

In the fourth group, three papers address the fair resource management in multihop and mesh networks.

The first paper titled "Outage probability versus fairness trade-off in opportunistic relay selection with outdated CSI" by J. L. Vicario et al. analyzes the performance of opportunistic relay selection in a decode and forward cooperative relaying wireless network. In order to achieve global balance in terms of performance and tradeoff, a relay selection strategy has been proposed based on max-normalized SNR criterion. The tradeoff in terms of system performance (outage probability) versus fairness (relay node power consumption) among relays is studied for different relay selection strategies using portfolio theory. The impact of availability of accurate channel state information on the performance is also investigated.

The second paper titled "Cross-layer optimal rate allocation for heterogeneous wireless multicast" by A. Mohamed and H. Alnuweiri addresses the problem of rate allocation for heterogeneous multicast sessions over multihop wireless networks. The problem is formulated as a nonlinear optimization problem with an objective to optimizing resource allocation while providing system-wide fairness for end-to-end multirate multicast flows.

Based on primal-dual and pricing methods, the problem is decomposed into subproblems, which are easier to solve in a modular structure. The authors propose an iterative algorithm to solve the problem in a distributed ad hoc network environment with asynchronous computations.

The third paper titled "A novel approach to fair routing in wireless mesh networks" by J. Matti, H. Määttä, and T. Braysy proposes a novel centralized routing algorithm for wireless mesh networks. The proposed scheme can assure fairness, leads to a feasible scheduling, and does not collapse the aggregate network throughput with a strict fairness criterion.

The papers in the fifth group address the problem of fairness in RRM for MIMO systems.

The first paper in this group titled "On throughput-fairness trade-off in virtual MIMO systems with limited feedback" by A. A. Dowhuszko et al. investigates the performance of channel-aware scheduling algorithms designed for the downlink of a wireless communication system. The study focuses on a two-transmit antenna cellular system, where the base station can only rely on quantized versions of channel state information to carry out scheduling decisions. Virtual MIMO system selects at each time instant a pair of users that report orthogonal (quantized) channels. Closed-form expressions for the achievable sum-rate of three different channel-aware scheduling rules are presented using an analytical framework.

The second paper titled "Throughput versus fairness: channel-aware scheduling in multiple antenna downlink" by E. A. Jorswieck, A. Sezgin, and X. Zhang studies the trade-off of using four channel-aware scheduling algorithms using majorization theory for a space-time coded multiple antenna downlink system, where TDMA-based scheduling is employed and spatial diversity is exploited. The scaling laws of average sum rate and of average worst case delay are derived. The impact of user distributions on the system performance and the average worst case delay are analyzed.

The papers in the last group deal with the problem of multiuser fair resource allocation.

The first paper titled "Optimal and fair resource allocation for multiuser wireless multimedia transmissions" by Z. Guan, D. Yuan, and H. Zhang proposes an optimal and fair strategy for multiuser multimedia radio resource allocation based on copetition, a mixture of cooperation and competition. The copetition strategy is formulated as sum utility maximization under constraints from both APP and PHY and is shown to be effective to allocate power among multiple video users.

The second paper titled "Performance analysis of SNR-based scheduling policies in asymmetric broadcast ergodic fading channels" by J. Perez et al. analyzes the performance of SNR-based scheduling algorithms in broadcasting ergodic fading channels by exploiting multiuser selection diversity. At each fading state, the base station transmits to the user of the highest SNR. By arranging weights to users according to a specific scheduling policy, QoS or fairness can be achieved.

We hope the readership will find the papers in this special issue useful for their research. Finally, we would like to thank the authors of all submissions, the reviewers for their enormous help with the review process, and the editorial staff of EURASIP JWCN for their support during all phases of this special issue.

Mohamed Hossam Ahmed
Alagan Anpalagan
Kwang-Cheng Chen
Zhu Han
Ekram Hossain

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 918261, 13 pages
doi:10.1155/2009/918261

Research Article

Fair Adaptive Bandwidth and Subchannel Allocation in the WiMAX Uplink

Antoni Morell, Gonzalo Seco-Granados, and José López Vicario

Telecommunications and System Engineering Department (TES), Autonomous University of Barcelona (UAB), 08193 Bellaterra, Spain

Correspondence should be addressed to Antoni Morell, antoni.morell@uab.cat

Received 2 July 2008; Revised 22 November 2008; Accepted 29 December 2008

Recommended by Ekram Hossain

In some modern communication systems, as it is the case of WiMAX, it has been decided to implement Demand Assignment Multiple Access (DAMA) solutions. End-users request transmission opportunities before accessing the system, which provides an efficient way to share system resources. In this paper, we briefly review the PHY and MAC layers of an OFDMA-based WiMAX system, and we propose to use a Network Utility Maximization (NUM) framework to formulate the DAMA strategy foreseen in the uplink of IEEE 802.16. Utility functions are chosen to achieve fair solutions attaining different degrees of fairness and to further support the QoS requirements of the services in the system. Moreover, since the standard allocates resources in a terminal basis but each terminal may support several services, we develop a new decomposition technique, the coupled-decompositions method, that obtains the optimal service flow allocation with a small number of iterations (the improvement is significant when compared to other known solutions). Furthermore, since the PHY layer in mobile WiMAX has the means to adapt the transport capacities of the links between the Base Station (BS) and the Subscriber Stations (SSs), the proposed PHY-MAC cross-layer design uses this extra degree of freedom in order to enhance the network utility.

1. Introduction

The wireless community has recently directed much attention on a variety of topics related to Worldwide Interoperability for Microwave Access (WiMAX) technologies as a broadband solution. Two different standards are under this commercial nomenclature: the IEEE 802.16 [1], with its extension to mobile scenarios IEEE 802.16e [2], and the ETSI HiperMAN [3]. Operating in the range of 2 GHz to 11 GHz, WiMAX enables a fast deployment of the network even in remote locations with low coverage of wired technologies, such as the Digital Subscriber Loop (DSL) family, and it can be used, among others, for wireless backhaul or last-mile applications.

The IEEE 802.16 standards family provides manufacturers with basically four different physical (PHY) layers [4]. Two of them are based on single carrier transmissions and use Time Division Multiple Access (TDMA) whereas the other two are based on multicarrier modulations and use either TDMA or Orthogonal Frequency Division Multiple Access (OFDMA). Within the multicarrier subgroup, the WirelessMAN Orthogonal Frequency Division Multiplexing (OFDM) uses a 256-point Fast Fourier Transform- (FFT-) based OFDM modulation together with a TDMA scheme to deploy a Point-to-Multipoint (PMP) subnetwork in the frequency range from 2 GHz up to 11 GHz in Non-Line-of-Sight (NLOS) propagation conditions. This PHY layer has been accepted for fixed WiMAX applications, and it is often termed as fixed WiMAX. Finally, WirelessMAN OFDMA exploits the multicarrier principles to implement a more flexible OFDMA access scheme. As in WirelessMAN OFDM, it is intended for NLOS PMP applications in the 2 GHz–11 GHz range. However, it uses a variable-size FFT ranging from 128 up to 2048 subcarriers. This PHY layer has been accepted for mobile WiMAX applications, and it is usually termed mobile WiMAX.

Concerning network topology, the basic configuration is PMP with a Base Station (BS) serving many Subscriber Stations (SSs). Not with standing, there is also a mesh mode available where SSs can be linked directly to the BS or routed through other SSs. This last mode is out of the scope of this paper, where we consider the design of appropriate scheduling mechanisms in uplink using the WirelessMAN OFDMA PHY layer and a PMP network. The conceived scheduling mechanism is based on a Demand Assignment Multiple Access (DAMA) strategy that implements a Dynamic Bandwidth Allocation (DBA) solution (where bandwidth is understood as rate in a wide sense). Jointly with flow allocation, we consider the adjustment of the transmission parameters of the OFDM system, and hence, the joint approach proposes a cross-layer interaction between PHY and Medium Access Control (MAC) system layers.

Previous works related to Radio Resource Management (RRM) in WiMAX networks address a variety of scenarios, from PMP to mesh, from TDMA to OFDMA access types, and distinguishing single channel or multichannel networks, most of them from a physical (PHY) layer perspective, where the goal is to properly configure the transmission parameters. At the best of our knowledge, two main approaches are found in literature, namely: (i) formulate the problem in a mathematical optimization framework and (ii) develop heuristic algorithms. In the sequel, we briefly review some of the works. In [5], the author proposes an heuristic solution for the case of a single cell OFDMA WiMAX network that maximizes the network sum-rate under some fairness considerations by means of performing subcarrier and power allocation. The authors in [6] analyze how concurrent transmissions boost performance in mesh type networks by proposing an interference-aware routing and scheduling mechanism. In [7], the reader can find a discussion about the advantages of a multichannel network. Finally, [8] contributes with a mathematical optimization solution that falls into the Network Utility Maximization (NUM) framework, where a distributed optimal solution to the established NUM problem is obtained using a convex decomposition approach. The authors extend in [9] their original work to generic OFDMA mesh networks, and the contributions in [10–12] are within the same context. A common feature in the last three references is that they split the global rate control and resource allocation problem into independent and smaller subproblems in order to alleviate the complexity of the solution at the expenses of a certain loss in optimality.

Our work follows the NUM framework to define the underlying optimization problem as in [8] but modifies the formulation in order to exactly fit the DAMA process that is envisaged for the WiMAX uplink. The problem is then decomposed (without any loss in optimality) using the Mean Value Cross (MVC) decomposition method [13]. It allows to separate the original joint problem into a flow optimization problem (given fixed link capacities) and a radio resource optimization problem (given fixed values of transmission rates). The latter results in a linear program that can be solved centrally at the BS, whereas a distributed solution that uses the novel proposed coupled-decompositions method is applied to the former.

The rest of the paper is organized as follows. Section 2 describes the system model. Section 3 reviews the MVC decomposition technique and introduces the novel coupled-decompositions method, whereas Section 4 solves the proposed joint problem in Section 2. Finally, Section 5 gives some numerical results, and Section 6 ends the paper with the conclusions.

2. System Model

Let us consider a PMP OFDMA WiMAX network as depicted in Figure 1, where a number of SSs share a subset of the subchannels in the system. A subchannel in WiMAX is made up of some of the system subcarriers and lasts for several OFDM symbols in time. There exist different ways to gather subcarriers into subchannels, which depend on the permutation types (see in [4] a good review on WiMAX aspects). In this work, we assume that the transmitting power per subchannel as well as the set of subcarriers that form it is given. Therefore, the different powers are not variables of our allocation problem. Furthermore, each terminal allocates the amount of power at each subchannel among the inner subcarriers in order to optimize the transmitting rate. This assumption can be found in [14], where the authors take into account intercell interference to constrain the subchannel transmitting powers. Note that one interesting extension is then the inclusion of subchannel power allocation but it is beyond the scope of this paper. In our framework, given a specific allocation of subchannels to terminals $\{\boldsymbol{\rho}_i\}$ (top left part of the figure), each terminal is able to transmit at a rate $c^i(\boldsymbol{\rho}_i)$, which is the sum of the rates that the SS attains in its active subchannel subset (the subset allocated to the terminal).

We further assume (as described in the IEEE 802.16 standard documents) that each terminal negotiates the resource allocation for all traffic flows that go through it, that is, it jointly requests transmission opportunities for the ongoing connections without doing it on a flow basis. The advantage of this procedure is that signaling is reduced, specially when a significant number of connections have to be managed. The disadvantage is that, depending on the particular mechanism used to find the solution of the problem, it may not be optimal. In that sense, solutions derived from distributed optimization do not sacrifice optimality. The price to pay is the time required to get the solution, and therefore, we are interested in techniques that converge fast. In Figure 1, the rate of the jth flow at the ith SS is labeled as r_j^i.

The IEEE 802.16 standard defines five different scheduling services that will provide Quality of Service (QoS) differentiation among the multiple traffic types. These services are [4] (i) the Unsolicited Grant Service (UGS) (ii) the real-time Polling Service (rtPS) (iii) the non-real-time Polling Service (nrtPS) (iv) the Best-Effort (BE) service, and (v) the extended real-time Polling Service (ertPS). Let us model the DAMA solution implemented in the WiMAX uplink by means of a convex program [15] where the different scheduling services are mapped using three parameters: the minimum rate that has to be allocated to the connection (the jth flow at the ith terminal) or m_j^i, the rate requested or d_i^j, and the priority of the service or p_j^i. The desired QoS degree of each service depends then on both m_j^i and p_j^i. For example, the UGS that needs a constant rate can be requested just by plugging

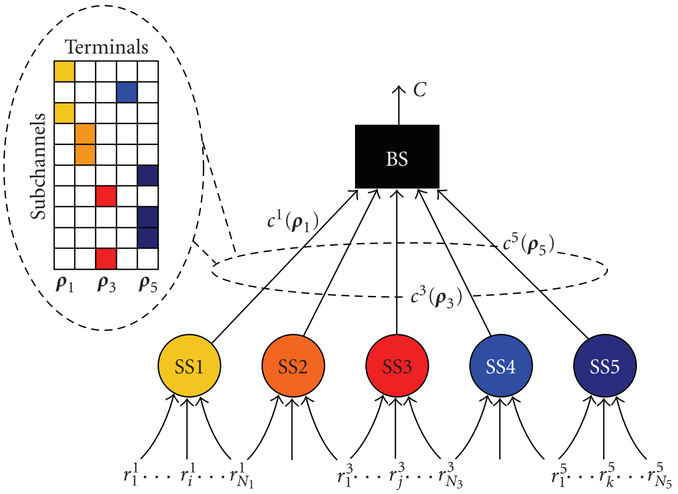

Terminals
Subchannels
ρ_1 ρ_3 ρ_5
C
BS
$c^1(\rho_1)$
$c^5(\rho_5)$
$c^3(\rho_3)$
SS1
SS2
SS3
SS4
SS5
$r_1^1 \cdots r_i^1 \cdots r_{N_1}^1$
$r_1^3 \cdots r_j^3 \cdots r_{N_3}^3$
$r_1^5 \cdots r_k^5 \cdots r_{N_5}^5$

Figure 1: Reference model.

that rate into m_j^i and fixing $d_j^i = m_j^i$ regardless the value of p_j^i. Another example is the ertPS that can be requested with some amount of m_j^i for the fixed allocation part and some $d_j^i > m_j^i$ for the variable rate part. The value p_j^i is then used to prioritize this flow against other competing connections.

In summary, the cross-layer system model used to characterize the DBA part of WiMAX, including PHY and MAC layer issues, responds to the following convex optimization problem in maximization form [15, Section 4.1.3]:

$$\begin{aligned} \max_{\{r_j^i\},\Gamma} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N_i} U_j^i(r_j^i; p_j^i) \\ \text{s.t.} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N_i} r_j^i \leq C, \\ & \sum_{j=1}^{N_i} r_j^i \leq c^i(\boldsymbol{\rho}_i), \quad i = 1,\dots,N, \\ & m_j^i \leq r_j^i \leq d_j^i, \quad \forall i,\ \forall j, \\ & \Gamma\mathbf{1} \preccurlyeq \mathbf{1}, \\ & \boldsymbol{\rho}_i \succcurlyeq \mathbf{0}, \quad i = 1,\dots,N, \end{aligned} \tag{1}$$

where $U_j^i(r_j^i; p_j^i)$ is the function that measures the utility perceived by the connection when the rate r_j^i is allocated. The function has p_j^i as a parameter. Furthermore, $\Gamma = [\boldsymbol{\rho}_1,\dots,\boldsymbol{\rho}_N]$ collects the subchannel allocation per user ($\boldsymbol{\rho}_i$), and the symbols $\preccurlyeq$ and $\succcurlyeq$ stand for component-wise non-strict inequalities. Finally, $c^i(\boldsymbol{\rho}_i) = \boldsymbol{\rho}_i^T\mathbf{c}_i$, where $\mathbf{c}_i$ contains the achievable rates of SS$_i$ at each possible subchannel, and C is the rate at which the BS can transmit. Note that in principle the allocation variables within each vector $\boldsymbol{\rho}_i$ should take the integer values 0 and 1 so that a given subchannel is completely allocated to a certain SS, whereas the constraint $\Gamma\mathbf{1} \preccurlyeq \mathbf{1}$ forces that no more than one terminal gets the subchannel. As it has been done in other works in literature [16], we relax the integer constraints to $\rho_i^k \geq 0$, which allows us to represent the problem as a convex one (easy to solve). Once the solution of the relaxed problem is found, a suboptimal solution to the original problem (with integer constraints) is obtained by means of employing rounding algorithms. However, in the WiMAX scenario and taking into account that an allocation is kept during several time-slots, real-valued allocation variables have sense in practice (by time sharing of subchannels). Indeed, if we consider that the allocation lasts for T time slots, then it is possible to use values in Γ with a granularity of $1/T$.

Not with standing, the problem in (1) itself does not guarantee a fair allocation of resources. Fortunately, such distribution can be attained by means of employing adequate utility functions, and a general formulation for fairness was introduced in [17] under the nomenclature of $(\mathbf{p},\alpha)$-proportional fairness. A feasible rate vector $\mathbf{r}^\dagger$ (i.e., it attains the generic network constraints $\mathbf{A}\mathbf{r}^\dagger \preccurlyeq \mathbf{c}$) is said to be $(\mathbf{p},\alpha)$-proportionally fair (where $\mathbf{p} = [p_1,\dots,p_{N'}]^T$ and α are positive real numbers) if, given any other feasible rate vector $\mathbf{r}^\ddagger$, it holds that

$$\sum_{i=1}^{N'} p_i \frac{r_i^\ddagger - r_i^\dagger}{\left(r_i^\dagger\right)^\alpha} \leq 0, \quad \forall \mathbf{r}^\ddagger \text{ s.t. } r_i^\ddagger \geq 0,\ \mathbf{A}\mathbf{r}^\ddagger \preccurlyeq \mathbf{c}. \tag{2}$$

Accordingly, the utility functions that accomplish this fairness criterion are [17]

$$U_i(r_i; p_i, \alpha) = \begin{cases} p_i \log(r_i), & \alpha = 1, \\ p_i \dfrac{r_i^{(1-\alpha)}}{1-\alpha}, & \alpha \neq 1. \end{cases} \tag{3}$$

The reader can find in Figure 2 the plots of $U_i(r_i; p_i, \alpha)$ for $\alpha = 0.1$, $\alpha = 1$, and $\alpha = 3$ (equal p_i value).

Let us fix $\mathbf{p} = [1,\dots,1]^T$ and move from $\alpha \to \infty$ to $\alpha = 0$. With $\alpha \to \infty$, the solution is said to be max-min fair [18, Section 6.5], and it is not possible (given feasibility, i.e., $\mathbf{A}\mathbf{r} \preccurlyeq \mathbf{c}$) to increase any rate in the network, say r_j, without decreasing another rate $r_p < r_j$. On the other hand, when $\alpha \to 0$, the flow allocation problem leads to a max sum-rate approach, and therefore, it drastically favors the users

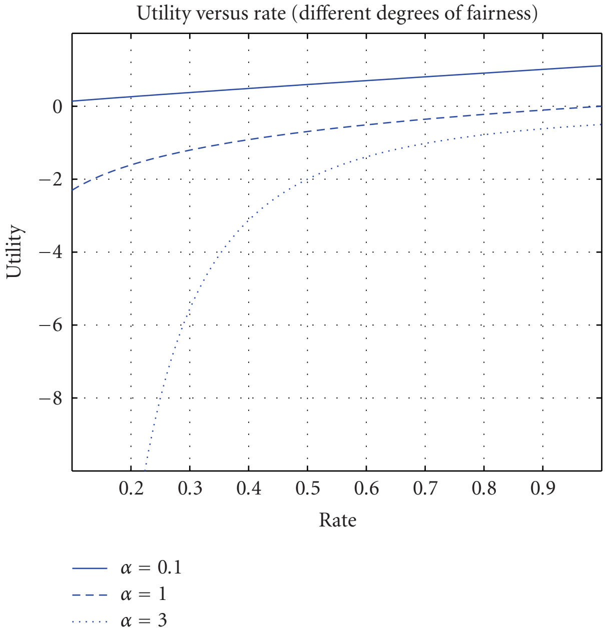


Figure 2: Different degrees of fairness (α) in the definition of utility functions.

with better links (it is then unfair). Intermediate solutions allow a certain decrease in r_p at the expenses of a greater increase in r_j depending on α. Note that in Figure 2 the bigger the α value is, the higher the increase in r_j will be in order to compensate a utility loss in r_p. A common adopted solution in literature is $\alpha = 1$, and it was termed by Kelly [19] as proportional fair. Moreover, this solution coincides with the Nash Bargaining one, and therefore, it accomplishes the recognized, axioms in game theory [20] of linearity, irrelevant alternatives and symmetry [21].

We can conclude that there is no unique criterion to define fairness but a series of them are explicitly characterized with the utility functions in (3). Furthermore, some flows can be prioritized over the others within a specific fairness framework (fixed by α) by particular adjustment of the scale thanks to the parameters $\{p_i\}$. In general, proportional fairness ($\alpha = 1$) provides a reasonable trade-off between fairness and resource utilization (network throughput).

3. Decomposition in Convex Programming

Decomposition techniques are used to break down a given optimization problem into a number of smaller problems, usually termed the subproblems. The most used decomposition methods in communications literature and in relation to convex optimization are primal and dual decompositions [22, 23]. It is usual to employ these decomposition techniques as a tool to obtain distributed solutions to some problems, as it is the case in Network Utility Maximization (NUM) problems [24, 25]. The formulation in (1) is an adaptation of the classical NUM to match the DBA problem in OFDMA WiMAX. Recently, Palomar and Chiang provided an exhaustive review on primal and dual decompositions applied to the classical NUM and extensions of it [26]. In particular, they proposed multilevel decomposition approaches to split the problem into different and coupled subsets of variables (e.g., link powers and transmission rates). However, the problem in primal and dual decompositions is that, in general, they converge slowly and that an adaptation step size has to be fixed by the user. So motivated, we base our work in two distinct decomposition techniques: the Mean Value Cross (MVC) decomposition [13] and the proposed novel coupled-decompositions method. In the following, we briefly review the former and describe the latter.

3.1. Mean Value Cross Decomposition. Consider the following problem formulation from [13]:

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{y}} \quad & c(\mathbf{x}) + e(\mathbf{y}) \\
\text{s.t.} \quad & \mathbf{A}_1(\mathbf{x}) + \mathbf{B}_1(\mathbf{y}) \leq \mathbf{b}_1, \\
& \mathbf{A}_2(\mathbf{x}) + \mathbf{B}_2(\mathbf{y}) \leq \mathbf{b}_2, \\
& \mathbf{x} \in \mathcal{X}, \\
& \mathbf{y} \in \mathcal{Y},
\end{aligned} \tag{4}$$

where $c : \mathbb{R}^{n_1} \to \mathbb{R}$, $e : \mathbb{R}^{n_2} \to \mathbb{R}$, $\mathbf{A}_1 : \mathbb{R}^{n_1} \to \mathbb{R}^{m_1}$, $\mathbf{B}_1 : \mathbb{R}^{n_2} \to \mathbb{R}^{m_1}$, $\mathbf{A}_2 : \mathbb{R}^{n_1} \to \mathbb{R}^{m_2}$, and $\mathbf{B}_2 : \mathbb{R}^{n_2} \to \mathbb{R}^{m_2}$ are convex functions. The sets $\mathcal{X}$ and $\mathcal{Y}$ are also convex and compact. It is further assumed that strong duality holds.

Construct now the partial Lagrangian function of the problem (4) as

$$L(\mathbf{x},\mathbf{y},\boldsymbol{\mu}) = c(\mathbf{x}) + e(\mathbf{y}) + \boldsymbol{\mu}^T (\mathbf{A}_1(\mathbf{x}) + \mathbf{B}_1(\mathbf{y}) - \mathbf{b}_1) \tag{5}$$

and minimize it over the variable $\mathbf{x}$, including the constraints that have not been taken into account in the Lagrangian definition, to obtain the function $K(\mathbf{y},\boldsymbol{\mu})$ as follows:

$$\begin{aligned}
K(\mathbf{y},\boldsymbol{\mu}) = & \min_{\mathbf{x}} \ L(\mathbf{x},\mathbf{y},\boldsymbol{\mu}) \\
& \text{s.t. } \mathbf{A}_2(\mathbf{x}) \leq \mathbf{b}_2 - \mathbf{B}_2(\mathbf{y}), \\
& \mathbf{x} \in \mathcal{X},
\end{aligned} \tag{6}$$

which is convex in $\mathbf{y}$ and concave in $\boldsymbol{\mu}$ [13].

From $K(\mathbf{y},\boldsymbol{\mu})$, the method defines the primal and the dual subproblem by fixing either the primal variable $\mathbf{y}$ or the dual variable $\boldsymbol{\mu}$. After some manipulations, the primal subproblem turns into

$$\begin{aligned}
p(\mathbf{y}) = & \min_{\mathbf{x}} \ c(\mathbf{x}) + e(\mathbf{y}) \\
& \text{s.t. } \mathbf{A}_1(\mathbf{x}) \leq \mathbf{b}_1 - \mathbf{B}_1(\mathbf{y}), \\
& \mathbf{A}_2(\mathbf{x}) \leq \mathbf{b}_2 - \mathbf{B}_2(\mathbf{y}), \\
& \mathbf{x} \in \mathcal{X}
\end{aligned} \tag{7}$$

and the dual subproblem into

$$\begin{aligned}
d(\boldsymbol{\mu}) = \min_{\mathbf{x},\mathbf{y}} \quad & c(\mathbf{x}) + e(\mathbf{y}) + \boldsymbol{\mu}^T (\mathbf{A}_1(\mathbf{x}) + \mathbf{B}_1(\mathbf{y}) - \mathbf{b}_1) \\
\text{s.t.} \quad & \mathbf{A}_2(\mathbf{x}) + \mathbf{B}_2(\mathbf{y}) \leq \mathbf{b}_2, \\
& \mathbf{x} \in \mathcal{X}, \\
& \mathbf{y} \in \mathcal{Y}.
\end{aligned} \tag{8}$$

Finally, the method is completed by passing filtered versions of the primal and dual variables between the primal and dual subproblems, as it is summarized in the following algorithm.

```
Take starting points $\boldsymbol{\mu}^0 \succcurlyeq \mathbf{0}$ and $\mathbf{y}^0 \in \mathcal{Y}$ and let $k = 1$.
Repeat
    (1) Let $\overline{\boldsymbol{\mu}}^k = (1/k)\sum_{i=0}^{k-1}\boldsymbol{\mu}^{k-1} = (1/k)\boldsymbol{\mu}^{k-1} + ((k-1)/k)\overline{\boldsymbol{\mu}}^{k-1}$ and compute $d(\overline{\boldsymbol{\mu}}^k)$ as in (8). Get $\mathbf{y}^k$ as the inner minimizer of $d(\overline{\boldsymbol{\mu}}^k)$.
    (2) Let $\overline{\boldsymbol{y}}^k = (1/k)\sum_{i=0}^{k-1}\mathbf{y}^{k-1} = (1/k)\mathbf{y}^{k-1} + ((k-1)/k)\overline{\boldsymbol{y}}^{k-1}$ and compute $p(\overline{\boldsymbol{y}}^k)$ as in (7). Get $\boldsymbol{\mu}^k$ as the inner Lagrange multiplier of $p(\overline{\boldsymbol{y}}^k)$.
    (3) $k = k + 1$.
Until $p(\overline{\boldsymbol{y}}^k) - d(\overline{\boldsymbol{\mu}}^k) < \epsilon$.
```

Further details on the MVC decomposition method can be found in [13].

3.2. Coupled-Decompositions Method. Let us consider now the following problem formulation:

$$\begin{aligned} \min_{\{\mathbf{x}_j\},\mathbf{y}} \quad & \sum_{j=1}^{J} f_j(\mathbf{x}_j) \\ \text{s.t.} \quad & \mathbf{x}_j \in \mathcal{X}_j, \quad j = 1,\ldots,J, \\ & h_j(\mathbf{x}_j) \le y_j, \quad j = 1,\ldots,J, \\ & \mathbf{1}^T\mathbf{y} \le c, \\ & \mathbf{y} \in \mathcal{Y}, \quad \mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_J, \end{aligned} \tag{9}$$

where $\mathbf{1}$ is a column vector with all J entries equal to one, and the subset $\mathcal{Y}$ is the cartesian product of J convex one-dimensional subspaces that include the minimum and maximum values of the variables $\{y_j\}$, and thus, it is convex. We consider that μ is the dual variable associated to the coupling constraint $\mathbf{1}^T\mathbf{y} \le c$. In the sequel, we briefly describe the algorithm that we propose in order to solve (9). However, the interested reader can find in [27, 28] an extended and well-reasoned version of it.

The technique intertwines the primal and dual subproblems that are obtained when classical primal and dual decompositions [22, Section 6.4] are applied to (9). In primal decomposition, the J subproblems appear when $\mathbf{y}$ is fixed. Note that under this assumption the problem is fully decoupled. Similarly, in dual decomposition we can relax the coupling constraint $\mathbf{1}^T\mathbf{y} \le c$ (constructing a partial Lagrangian of the problem with dual variable μ), and J subproblems are defined (the problem fully decouples again) for a fixed value of μ. In both classical strategies, the successive updates of $\mathbf{y}$ and μ are driven by the primal and dual master problems. In the coupled-decompositions method, the result of the primal subproblems is transformed using a redefined dual master problem, the dual projection, and plugged to the dual subproblems. Similarly, the output of the dual subproblems is transformed using the primal projection and fed to the primal subproblems. A flow diagram of the method is depicted in Figure 3. The algorithm starts with $\mu^0 = 0$ and iterates as follows: dual subproblems → primal projection → primal subproblems → dual projection → dual subproblems.

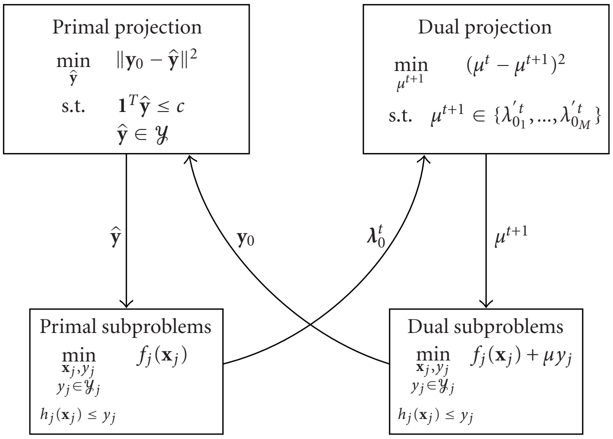

Primal projection
$\min_{\hat{\mathbf{y}}} \ \|\mathbf{y}_0 - \hat{\mathbf{y}}\|^2$
s.t. $\mathbf{1}^T\hat{\mathbf{y}} \le c$
$\hat{\mathbf{y}} \in \mathcal{Y}$
Dual projection
$\min_{\mu^{t+1}} \ (\mu^t - \mu^{t+1})^2$
s.t. $\mu^{t+1} \in \{\lambda_{0_1}'^t, \ldots, \lambda_{0_M}'^t\}$
$\hat{\mathbf{y}}$
$\mathbf{y}_0$
$\boldsymbol{\lambda}_0^t$
μ^{t+1}
Primal subproblems
$\min_{\mathbf{x}_j, y_j} \ f_j(\mathbf{x}_j)$
$y_j \in \mathcal{Y}_j$
$h_j(\mathbf{x}_j) \le y_j$
Dual subproblems
$\min_{\mathbf{x}_j, y_j} \ f_j(\mathbf{x}_j) + \mu y_j$
$y_j \in \mathcal{Y}_j$
$h_j(\mathbf{x}_j) \le y_j$

FIGURE 3: Flow diagram of the coupled-decompositions method.

Since primal and dual subproblems are extensively analyzed in literature (its formulation appears in Figure 3), let us now detail the novel parts. Notwithstanding, a complete iteration is revisited during the proof of the method. On one hand, primal projection is pretty similar to the primal master problem in primal decomposition. Assuming that $\mathbf{y}_0$ is constructed with the output of the J dual subproblems, the primal projection solves the following optimization problem:

$$\begin{aligned} \min_{\hat{\boldsymbol{y}}} \quad & \|\mathbf{y}_0 - \hat{\boldsymbol{y}}\|^2 \\ \text{s.t.} \quad & \mathbf{1}^T\hat{\boldsymbol{y}} \le c, \\ & \hat{\boldsymbol{y}} \in \mathcal{Y}, \end{aligned} \tag{10}$$

with the only particularity that the constraint $\mathbf{1}^T\hat{\boldsymbol{y}} \le c$ must be attained with equality when the last update of the Lagrange multiplier is $\mu > 0$. This is in accordance with the Karush-Kuhn-Tucker (KKT) conditions for convex problems [15, Section 5.5] (see more details in [27]). On the other hand, the dual projection takes the output values from the primal subproblems $\boldsymbol{\lambda}_0^t$ and selects the values within $\boldsymbol{\lambda}_0^t$ that have been obtained with primal variables $\hat{y}_j$ not in the boundary of $\mathcal{Y}_j$. Let us collect this subset in $\boldsymbol{\lambda}_0'^t$. The motivation is that the nonselected values do not directly impact on the value of μ (it can be seen from the KKT conditions of the problem; see more details in [27]). Thereafter, the μ update is found as

$$\mu^{t+1} = \arg\left\{ \begin{array}{ll} \min\limits_{\mu^{t+1}} & (\mu^{t+1} - \mu^t)^2 \\ \text{s.t.} & \mu^{t+1} \in \{\lambda_{0_1}'^t, \ldots, \lambda_{0_M}'^t\} \end{array} \right\}, \tag{11}$$

which updates μ with the value within $\boldsymbol{\lambda}_0'^t$ that is closer to the previous μ value.

Proof of the method: See the appendix.

4. Proposed Solution

Our solution uses a combination of both decomposition techniques. First, an MVC decomposition is applied, making it possible to split the joint problem into one flow or bandwidth allocation subproblem and one subchannel allocation subproblem. The latter depends on variables that are available at the BS, and thus, it is not necessary to explore distributed computations in order to solve it. On the contrary, the former is distributed among the BS and the SSs in order to be standard-compliant (the BS allocates aggregate bandwidth to the SSs and these decide the final allocation to flows and services). In this case, we use a two-level coupled-decompositions strategy.

First, let us consider the problem in (1) and identify rates with $\mathbf{x}$ and subchannel allocation variables with $\mathbf{y}$ in the MVC decomposition formulation in (4). Rewriting the original joint problem as

$$\begin{aligned} \max_{\{r_j^i\},\{\boldsymbol{\rho}_i\}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N_i} U_j^i(r_j^i; p_j^i) \\ & \sum_{j=1}^{N_i} r_j^i \le \boldsymbol{\rho}_i^T \mathbf{c}_i, \quad i = 1,\dots,N, \\ & \{r_j^i\} \in \mathcal{R}, \\ & \{\boldsymbol{\rho}_i\} \in \mathcal{S}, \end{aligned} \tag{12}$$

where $\mathcal{R} = \{r_j^i \mid m_j^i \le r_j^i \le d_j^i\}$ and $\mathcal{S} = \{\boldsymbol{\rho}_i \mid \Gamma\mathbf{1} \preccurlyeq \mathbf{1},\ \boldsymbol{\rho}_i \succcurlyeq \mathbf{0}\}$, we can define the primal subproblem of the MVC decomposition method as

$$\begin{aligned} \max_{\{r_j^i\}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N_i} U_j^i(r_j^i; p_j^i) \\ & \sum_{j=1}^{N_i} r_j^i \le \boldsymbol{\rho}_i^T \mathbf{c}_i, \quad i = 1,\dots,N, \\ & \{r_j^i\} \in \mathcal{R} \end{aligned} \tag{13}$$

for fixed values of $\{\boldsymbol{\rho}_i\}$ and the dual subproblem as

$$\begin{aligned} \max_{\{r_j^i\},\{\boldsymbol{\rho}_i\}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N_i} U_j^i(r_j^i; p_j^i) - \sum_{i=1}^{N}\gamma_i\left(\sum_{j=1}^{N_i} r_j^i - \boldsymbol{\rho}_i^T \mathbf{c}_i\right), \\ & \{r_j^i\} \in \mathcal{R}, \\ & \{\boldsymbol{\rho}_i\} \in \mathcal{S} \end{aligned} \tag{14}$$

for fixed values of the Lagrange multipliers $\{\gamma_i\}$ that are associated to the constraints that couple rates with subchannel allocation variables in (12). Note that the two subsets of variables are fully decoupled in (14), and thus, the maximization in $\{\boldsymbol{\rho}_i\}$ can be done independently solving the following linear program:

$$\begin{aligned} \max_{\{\boldsymbol{\rho}_i\}} \quad & \sum_{i=1}^{N}\gamma_i \cdot (\boldsymbol{\rho}_i^T \mathbf{c}_i) \\ & \{\boldsymbol{\rho}_i\} \in \mathcal{S}. \end{aligned} \tag{15}$$

The joint problem is then solved as follows.

Choose a feasible subchannel allocation $\{\boldsymbol{\rho}_i^0\}$ and let $k = 1$.

Repeat

(1) Let $\boldsymbol{\rho}_i^k = (1/k)\sum_{i=0}^{k-1}\boldsymbol{\rho}_i^{k-1}$, for all i.

(2) Solve (13) using $\{\boldsymbol{\rho}_i^k\}$ and get the dual variables $\{\gamma_i\}$.

(3) Let $\gamma_i^k = (1/k)\sum_{i=0}^{k-1}\gamma_i^{k-1}$, for all i.

(4) Solve (15) using $\{\gamma_i^k\}$ and get updated primal variables $\{\boldsymbol{\rho}_i\}$.

(5) $k = k + 1$.

Until convergence.

Since (15) is solved at the BS, the remaining issue is to find the solution of (13). In order to avoid excessive DBA-realted signaling in the subnetwork and to restrict ourselves to the standard, we propose to solve it using a two-level coupled-decompositions strategy. Note that we can rewrite (13) as

$$\begin{aligned} \max_{\{y^i\}} \quad & \sum_{i=1}^{N} U^i(y^i) \\ & \sum_{i=1}^{N} y^i \le C, \\ & y^i \le \boldsymbol{\rho}_i^T \mathbf{c}_i, \quad i = 1,\dots,N, \\ & M^i \preccurlyeq y^i \preccurlyeq D^i, \quad i = 1,\dots,N, \end{aligned} \tag{16}$$

where $M^i = \sum_{j=1}^{N_i} m_j^i$, $D^i = \sum_{j=1}^{N_i} d_j^i$, and

$$U^i(y^i) = \begin{cases} \max\limits_{\{r_j^i\}} & \sum_{j=1}^{N_i} U_j^i(r_j^i; p_j^i) \\ \text{s.t.} & \sum_{j=1}^{N_i} r_j^i \le y^i, \\ & m_j^i \le r_j^i \le d_j^i. \end{cases} \tag{17}$$

Note also that the dual Lagrange variable γ_i corresponds to the constraint $y^i \le \boldsymbol{\rho}_i^T \mathbf{c}_i$ in (16). Therefore, we apply the coupled-decompositions method to solve (16) at the upper layer (BS), and we use it again at the lower layer (at each SS) to solve (17) when it is required by the upper layer.

The iterations of the resulting two-level flow allocation algorithm and the involved signaling are summarized in the following list as well as in Figure 4.

(1) The dual variable μ^t (associated to $\sum_{i=1}^{N} y^i \le C$) is spread through the network, reaching each connection.

(2) Each connection computes the allocation given μ^t by means of solving the inner dual subproblems (the constraints in m_j^i and d_j^i can be obviated if desired without affecting convergence). The SSs and the BS get their own allocations by aggregation of the allocations below them.

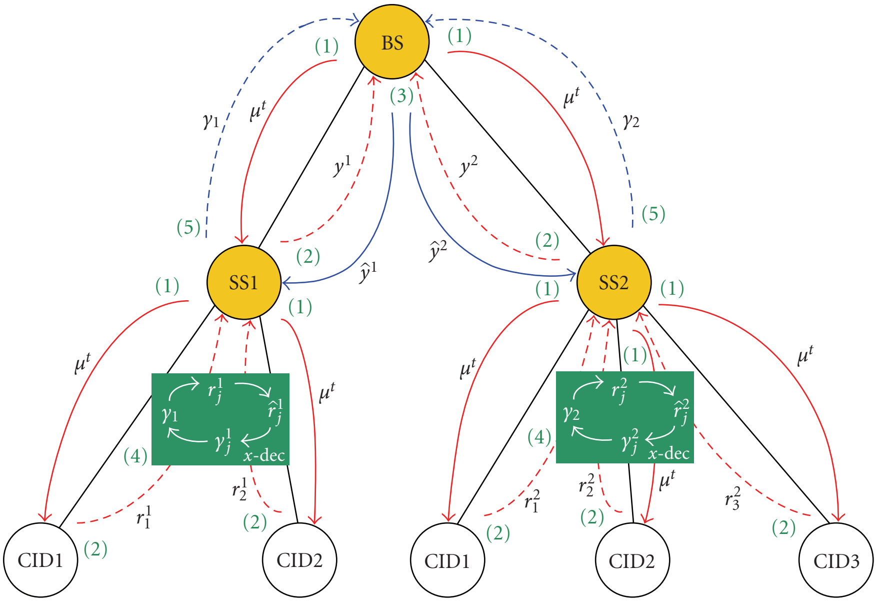


Figure 4: 2-level flow allocation algorithm.

(3) The BS corrects the previous allocations (primal projection) to attain $\sum_{i=1}^{N} y^i \leq C$ and $y^i \leq \boldsymbol{\rho}_i^T \mathbf{c}_i$, $i = 1, \ldots, N$.

(4) The corrected allocations are used by the SSs to perform inner iterations (within each SS) of the coupled-decompositions method in order to obtain new candidates γ_i.

(5) Finally, the BS updates the value of the dual variable to μ^{t+1} using the dual projection and the previous γ_i values.

Intuitively, the multilayer coupled-decompositions strategy tries to find a consensus on the price μ that has to be paid for sharing the transport capacity C of the BS. Often, primal variables are interpreted from a resource-oriented perspective whereas dual variables take the role of prices to be paid to use the resources [15, Section 5.4.4]. All CIDs participate in principle in finding such optimal value. However, the price of the connections within a particular SS may be distinct from the global price μ if, for example, its link capacity is small (hence forcing the price to locally increase). In these occasions, local prices γ_i that differ from the optimal and global consensus price μ are found.

Other works in literature [10–12] study a similar problem within generic mesh OFDM networks. In general, they search for suboptimal but affordable solutions, which are based on decoupling the joint problem into independent optimization programs that manage only a subset of the variables without looking at the others. In this work, we suggest (for the particular PMP WiMAX case) the derivation of the joint optimal rate and subchannel allocation (under fairness considerations), and we propose a distributed scheme that achieves it. Moreover, the numerical results in the next section show the practical interest of the mechanism in terms of the number of iterations (i.e., directly related to the amount of signaling). As a matter of fact, the proposed method (possibly with extensions) can be used in other scenarios to speed up the computation of optimization problems or subproblems, either in optimal or suboptimal decoupling approaches.

5. Numerical Results

Let us consider the network setup depicted in Figure 5 with 4 SSs and 9 connections (CIDs) in total. We choose logarithmic utility functions ($\alpha = 1$),

$$U_j^i(r_j^i; p_j^i) = p_j^i \log(r_j^i). \tag{18}$$

Other policies balancing the solution towards the max-sum-rate or the max-min-fair designs can be implemented by fixing other α values using the same algorithm (as discussed later). We fix all requests to 100 kbps (requests are emitted in WiMAX in terms of bytes of information but we transform them to rates taking into account the time basis) and all the minimum granted rates to 1 kbps. All connections have the same priority $p_j^i = 1$. The available number of subchannels is 7, all of them to be shared among the 4 SSs. We consider the following transport capacities (in kbps) per subchannel (10 kHz of bandwidth) and user (given one realization of flat-fading Rayleigh subchannels that have 10 dB of SNR in mean):

$$[\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4] = \begin{bmatrix} 31.49 & 18.58 & 4.07 & 15.69 \\ 34.31 & 13.19 & 29.84 & 24.55 \\ 4.62 & 37.91 & 13.37 & 34.80 \\ 20.54 & 50.62 & 38.91 & 30.92 \\ 34.32 & 22.96 & 27.38 & 48.95 \\ 39.21 & 0.01 & 32.39 & 25.97 \\ 22.10 & 23.69 & 47.14 & 3.86 \end{bmatrix}. \tag{19}$$

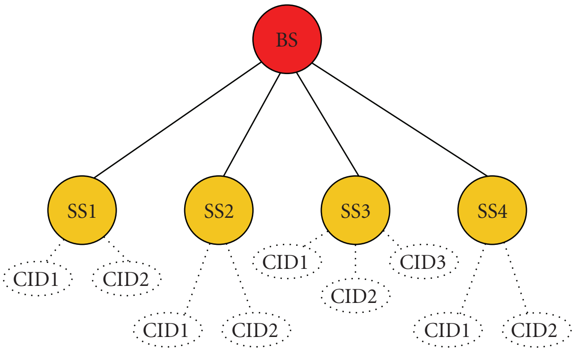


Figure 5: Setup of the network under test.

Note that depending on the scheduling length (i.e., the number of contiguous time slots in time that are allocated in a single allocation phase, which fixes the granularity of the $\boldsymbol{\rho}_i$ values) and on the channel characteristics (coherence time), it is reasonable to consider which values of $\mathbf{c}_i$ may be really achieved within each allocation phase (mid-term values seem reasonable) so that one may resort to robust designs in order to compute them. The output rate capacity of the BS is 200 kbps, and the initial subchannel allocation is $\Gamma = [\mathbf{I}_{4\times 4}, \mathbf{0}_{4\times 3}]^T$ achieving the link capacities $[c^1, c^2, c^3, c^4] = [31.49, 13.19, 13.37, 30.92]$.

Figure 6 shows the evolution of the subchannel allocation variables when we apply the proposed method, achieving new link capacities $[c^1, c^2, c^3, c^4] = [89.39, 86.83, 60.44, 49.23]$. In order to accelerate the convergence to the solution, we have used instantaneous values of $\{\gamma_i\}$ instead of the time-average that is proposed in the MVC decomposition method, averaging only the primal (allocation) variables. This solution has been derived by other authors [8] using a different approach (which validates it), and it is specially relevant in the first iterations where the $\{\gamma_i\}$ values show abrupt changes and very high values. Note that in the figure the final allocation is completely different from the initial one (only SS1 keeps using subchannel 1) but the solution still needs to be rounded to accommodate a practical scheduling implementation, which has its implications also in terms of convergence to the optimal solution because it may have sense to truncate the algorithm after some iterations and round that solution.

In Figure 7, we plot the resulting flow allocation per connection (that correspond to the CIDs ordered from left to right in Figure 5) and the final link capacity once the subchannel allocation has been obtained for the four scenarios specified in Table 1. The objective is to show how fairness considerations impact in the final allocation. The first Scenario is the same as in Figure 6, whereas Scenario 2 evaluates a different allocation scheme (with fairness parameter $\alpha = 0.1$). In the next two scenarios, we study the effect of different priorities using again a proportional fairness approach ($\alpha = 1$). The difference between Scenarios 3 and 4 is that Scenario 3 fixes the same requested rate for all the connections (100 kbps), whereas Scenario 4 has two possible requests (10 kbps and 100 kbps).

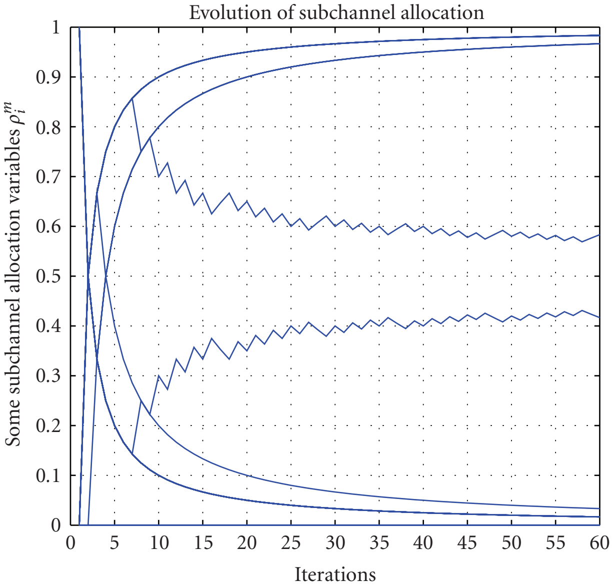


Figure 6: Evolution of some subchannel allocation variables ρ_i^m.

We notice in the results of Scenario 1 that link capacities have been adjusted (with the subchannel allocation mechanism) in order to provide a similar allocation to all connections. In Scenario 2, the allocation scheme favors the best channels so that each subchannel is assigned to the SS that experiences the maximum achievable rate at that subchannel. Therefore, SS1 gets subchannels 1, 2, and 6; SS2 gets subchannels 3 and 4; SS3 gets subchannel 7; SS4 gets subchannel 5. The corresponding link capacities are $[c^1, c^2, c^3, c^4] = [105.02, 88.54, 47.15, 48.95]$. The final rate allocation is limited by the outcoming rate at the BS (200 kbps) so that SSs 3 and 4 limit their ongoing connections to a lower rate than the connections in SSs 1 and 2, which share the remaining transport capacity. When prioritized traffic flows appear, as in Scenario 3, granted rates are balanced toward services depending on their priority values. Accordingly, it can be seen that subchannel allocation provides more link capacity to SSs 3 and 4. In Scenario 4, we further modify the requested rates with respect to Scenario 3 and the highest priority services in Scenario 3, (the ongoing connections of SS4) reach their requests. As expected, remaining resources (remember that the BS can manage no more than 200 kbps) are redistributed in order to allocate more rate to services in SS3 (with priorities equal to 3) than to services within SS1 and SS2 (with priorities equal to 1), while subchannel allocation favors the link BS-SS3 as well.

Finally, our last result analyzes the efficiency of the novel coupled-decompositions method (used to solve the flow allocation subproblem) in terms of convergence speed. For that purpose, we extend Scenario 1 to 20 SSs with 5 ongoing connections on each. The mean received SNR is 15 dB, and each ongoing connection in SSs 1–15 requests 100 kbps, whereas each connection in SSs 16–20 requests 10 kbps. The transport capacity at the BS is now increased to 1200 kbps.

Table 1: Scenario description.

Scenario number	Service priorities p_j^i	Fairness scheme α	Requested rate d_j^i	Granted rate m_j^i
1	All equal to 1	1	All equal to 100 kbps	All equal to 1 kbps
2	All equal to 1	0.1	All equal to 100 kbps	All equal to 1 kbps
3	1 for services in SS1, SS2 3 for services in SS3 5 for services in SS4	1	All equal to 100 kbps	All equal to 1 kbps
4	1 for services in SS1, SS2 3 for services in SS3 5 for services in SS4	1	100 kbps for services in SS1–SS3 10 kbps for services in SS4	All equal to 1 kbps

We plot in Figure 8 the evolution of the dual variable μ, that is, negotiated between the BS and the SSs when we use both our novel proposed method and a classic dual decomposition approach using the same 2-layer architecture. Remember that classical decomposition methods need to adjust the value of the step size of the gradient-based update. In this particular case, we have found that a setup with $\alpha(t) = 0.5/t$ at the highest level (i.e., between the BS and the SSs) and $\alpha(t) = 0.005/\sqrt{t}$ at the lowest (i.e., between SSs and connections) provides a satisfactory trade-off between convergence and speed. However, the need of a good adjustment is in practice an obstacle of the method, and it is not easy to find a step providing that good trade-off. On the contrary, one of the important advantages in the coupled-decompositions method is that any user-defined step is completely avoided. The other important advantage is in the number of iterations required. As shown in the figure, the novel technique converges in 5-6 iterations, contrary to the dual decomposition strategy (both obtain the same optimal solution), which needs more than 250 iterations. This drawback of dual decomposition appears in other works in literature, for example, in the numerical results of [10], where it is used to obtain a distributed solution that optimizes power and rate allocation within a mesh OFDM network.

6. Conclusions

In this work, we have proposed an algorithm that implements the DAMA mechanism foreseen in the IEEE 802.16 WiMAX standard. Initially, we have introduced our system model, which considers both flow and subchannel allocations in a cross-layer approach. Some PHY and MAC-layer aspects of WiMAX that are relevant to our work have been briefly reviewed as well as how to translate a series of fairness definitions into a convex optimization framework. All this has led us to formulate a network utility maximization problem.

Since the standard fixes that resources should be requested and granted in a terminal basis but we should consider several traffic flows within each SS (may be with different QoS requirements), we have proposed a distributed solution to the original convex optimization problem in order to fulfill these requirements while keeping the optimality in the allocation. Furthermore, we have explored the usage of our novel proposed coupled-decompositions algorithm and a recently proposed MVC decomposition method applied to distinct parts of the problem with the goal of achieving a more practical design than with classical primal and dual decompositions.

Results show that it is possible to find a solution to the flow allocation subproblem with very few iterations and without the manual setup of any parameter, as opposite to a classical dual decomposition. The last statement applies also to the subchannel allocation subproblem, which is able to give a good approximation to the solution within a reasonable number of iterations. Finally, we have shown with an example that our strategy is able to attain a fair distribution of resources and to support QoS by means of traffic prioritization.

Appendix

A. Proof of Convergence of the Coupled-Decompositions Method

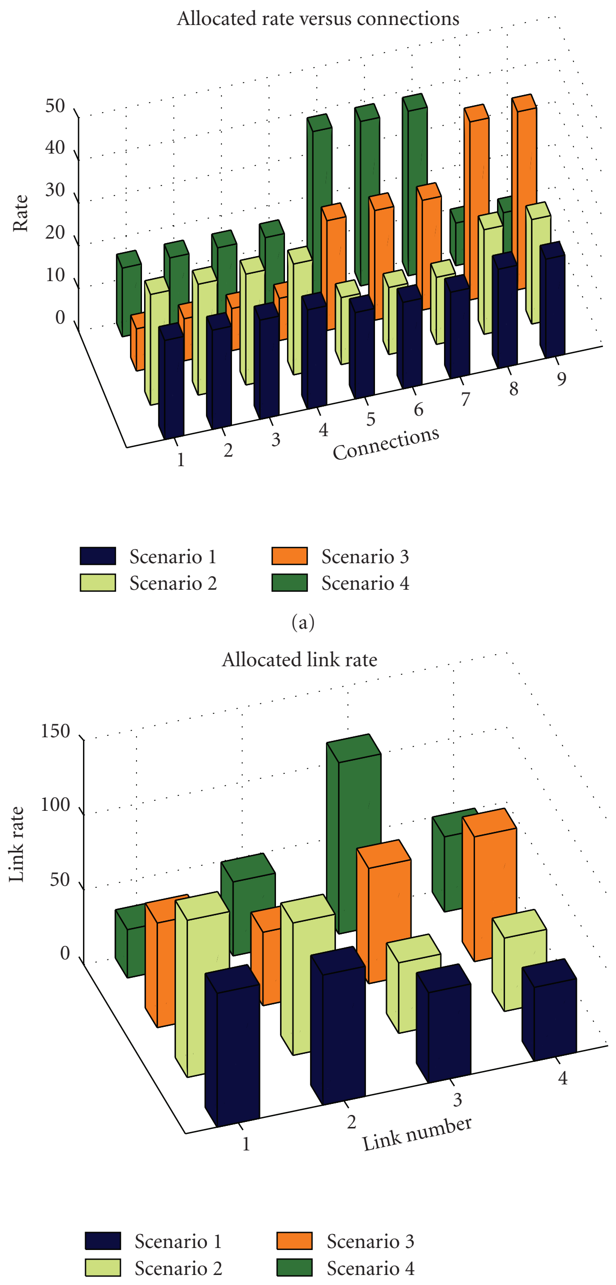
First of all, we assume that strong duality [15, Section 5.2.3] holds, which is usually verified in convex programs, so that the optimal primal variables attain the optimal dual variables when plugged into the subproblems and vice versa. In the following, the superscript t indicates iteration number although we omit it in some irrelevant occasions. Equivalently, the objective value of the problem is the same regardless it is solved directly (primal version) or by maximizing the dual function (dual version) [15, Section 5.2]. We will prove that

$$\boldsymbol{\lambda}_0^t = \mathbf{1}\mu^t \xrightarrow{t \to \infty} \boldsymbol{\lambda}^* = \mathbf{1}\mu^*, \tag{A.1}$$

where the relation $\boldsymbol{\lambda}_0^t = \mathbf{1}\mu$ is found by the application of the KKT conditions (see more details in [27]) and μ^* is the optimum value of the dual Lagrange variable. In the following, we review a complete iteration of the method.

Let us consider that $\mu^t < \mu^*$ (the proof is similar if $\mu^t > \mu^*$) and recall the result in [28, Lemma 1], where it is shown that the primal variable $\hat{y}_j$ at the jth subproblem (primal or dual) is a decreasing function of $\lambda_{0_j}^t$. This fact together with $\boldsymbol{\lambda}_0^t = \mathbf{1}\mu^t$ forces

$$\hat{y}_j(\lambda_0^t) \geq y_j^*, \quad \forall j, \tag{A.2}$$



(a)

(b)

Figure 7: Three different allocation examples.

where equality is attained only when $y_j^* \in \operatorname{bd} \mathcal{Y}_j$ (boundary of the subset) and therefore $\mathbf{1}^T \hat{\mathbf{y}} > c$.

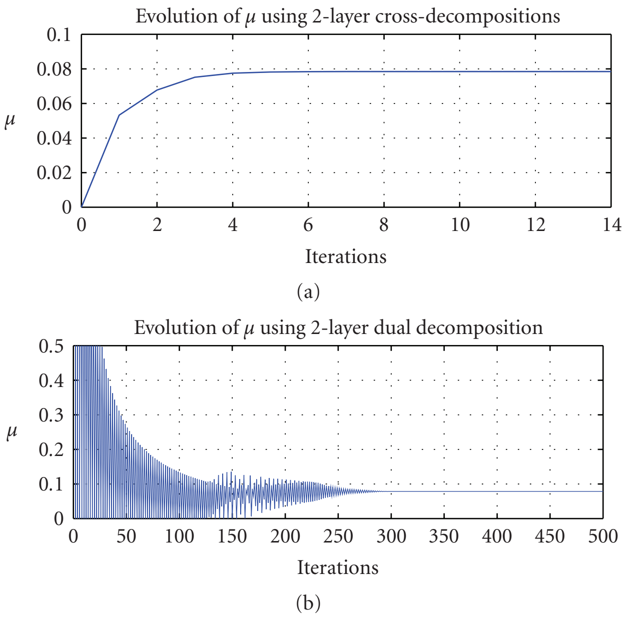
In the primal projection, it is verified that

$$\hat{y}_j = y_{0_j} - k_j, \quad k_j \geq 0, \ \forall j \tag{A.3}$$

thanks to the lemma below.

Lemma 1. *Given the optimization problem in* (10), *its optimal solution can be expressed as* $\hat{\mathbf{y}}^* = \mathbf{y}_0 - \mathbf{k}$ *with* $\mathbf{k} \succcurlyeq \mathbf{0}$.

Proof. See Section B. □



(a)

(b)

Figure 8: Evolution of μ value in the flow allocation subproblem. Comparison between a classical dual decomposition strategy and the proposed coupled-decompositions method.

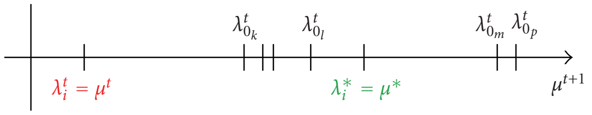


Figure 9: Example of the situation before dual projection.

Applying the relationship between the primal and dual variables of the subproblems to the previous $\hat{\mathbf{y}}^t$ value, it is fulfilled that

$$\lambda^t_{0_j} \geq \lambda^t_j, \quad \forall j. \tag{A.4}$$

Furthermore, given that $\hat{\mathbf{y}}^t$ is not the optimal value, it is verified that some of the $\lambda^t_{0_j}$ values are $\lambda^t_{0_j} \leq \lambda^*_j$ whereas the remaining ones are $\lambda^t_{0_j} \leq \lambda^*_j$, since it holds that $\mathbf{1}^T \hat{\mathbf{y}} = c$. In other words, some of the $\hat{y}_j$ values attain $\hat{y}_j \geq y^*_j$ whereas the rest verify $\hat{y}_j \leq y^*_j$. An example depicting the situation before dual projection can be found in Figure 9.

Consider now that $\boldsymbol{\lambda}'^t_0$ contains a single element. Note that a null vector is not possible since we assume that the coupling constraint is active. Then we can prove the following lemma.

Lemma 2. *Let a primal point* $\hat{\mathbf{y}}$ *attain* $\mathbf{1}^T \hat{\mathbf{y}} = c$ *and* $\hat{\mathbf{y}} \in \mathcal{Y}$. *Let also* $\boldsymbol{\lambda}'_0$ *be a vector containing the dual translation (computed by primal subproblems) of the values in* $\hat{\mathbf{y}}$ *that verify* $\hat{\mathbf{y}} \in \operatorname{int} \mathcal{Y}$ *(interior of the subset). Then, if the vector* $\boldsymbol{\lambda}'_0$ *is in fact a scalar, it is verified that*

$$\lambda'_0 \leq \lambda'^* = \mu^*, \tag{A.5}$$

where λ'^* *is the optimum value of* $\boldsymbol{\lambda}$ *for the selected position in* λ'_0 *(i.e., equal to* μ^**).*

Proof. Using Lemma 1, we can state that all the values within $\hat{\boldsymbol{y}}$ except the kth element accomplish $\hat{y}_i \in \inf \mathcal{Y}_i \ (i \neq k)$. Therefore, it holds that $\hat{y}_k < \hat{y}_k^* = y_{0_k}^* = y_k^*$. Applying the relationship between subproblems (remember that both in primal and dual subproblems, primal variables are a decreasing function of dual variables and $\hat{\boldsymbol{y}}(\boldsymbol{\lambda}_0^*) = \mathbf{y}^*$), we reach the desired result. □

Finally, we update μ^{t+1} using (11). Collecting all the results obtained up to this point, we have that

$$\mu^{t+1} > \mu^t \tag{A.6}$$

since every value in $\boldsymbol{\lambda}_0'^t$ verifies $\lambda_{0_i}'^t > \mu^t$. Furthermore, it is also true that

$$\mu^{t+1} < \mu^* \tag{A.7}$$

since the value $\lambda_{0_i}'^t$ closer to μ^t (dual projection) accomplishes $\lambda_{0_i}'^t < \lambda_i'^* = \mu^*$, which is derived from Lemma 2 and the discussion preceding it. Figure 9 provides a graphical explanation. We can finally conclude that

$$\mu^t < \mu^{t+1} < \mu^*. \tag{A.8}$$

The proof ends showing by contradiction that μ^t cannot tend to a value smaller than μ^*. Assume that there exists a value $\mu^{\triangleright}$ where successive iterations converge. Then $\mu^{\triangleright}$ is a stationary point of the method. In other words, a complete iteration of the method starting from $\mu^{\triangleright}$ returns exactly the same value. This enforces in the primal projection that $\hat{\boldsymbol{y}} = \mathbf{y}_0(\mu^{\triangleright})$, otherwise the values in $\boldsymbol{\lambda}_0'$ would increase and so the update in μ (dual projection). Given the relationship between primal and dual subproblems, we see that the previous equation is only attained if $\mu^{\triangleright} = \mu^*$ since a lower value $\mu^{\triangleright} < \mu^*$ would obtain a primal point $\mathbf{y}_0(\mu^{\triangleright})$ from dual subproblems such that $\mathbf{1}^T\mathbf{y}_0(\mu^{\triangleright}) > c$.

Before concluding this section, we want to note that it is possible to substitute the primal projection by the projection into $\mathbf{1}^T\mathbf{y} = c$ and the method still converges (it can be similarly proved). It is a more practical option since the projection can be analytically computed as [15, Section 8.1]

$$\hat{\boldsymbol{y}}^t = \mathbf{y}_0^t + \frac{(c - \mathbf{1}^T\mathbf{y}_0^t)\mathbf{1}}{J}. \tag{A.9}$$

B. Proof of Lemma 1

First, note that a point $\hat{\boldsymbol{y}} = \mathbf{y}_0 - \mathbf{k}$ with $\mathbf{k} \succcurlyeq \mathbf{0}$ is feasible since it attains both $\mathbf{1}^T\hat{\boldsymbol{y}} \le c$ and $\hat{\boldsymbol{y}} \in \mathcal{Y}$ (assuming that the intersection is not empty). Then, we have to proof that a point that does not accomplish the equation $\hat{\boldsymbol{y}} = \mathbf{y}_0 - \mathbf{k}$ for positive values in $\mathbf{k}$ cannot be optimal for problem (10).

We proof this last result by induction. Assume a certain vector $\mathbf{k}$, called $\mathbf{k}^{\triangleright}$ that attains $\mathbf{1}^T(\mathbf{y}_0 - \mathbf{k}^{\triangleright}) = c$ and $\mathbf{k}^{\triangleright} \succcurlyeq \mathbf{0}$. Construct now a new vector $\mathbf{k}^{\dagger}$ from $\mathbf{k}^{\triangleright}$ by fixing its lth element $k_l^{\dagger}$ to $-a$ with $a > 0$ and distributing the difference $|k_l^{\triangleright} - k_l^{\dagger}|$ among the rest of elements in $\mathbf{k}^{\dagger}$ so as to attain the equality coupling constraint. In other words,

$$k_i^{\dagger} = \begin{cases} -a, & i = l \\ k_i^{\triangleright} + \epsilon_i, & i \neq l,\ \epsilon_i > 0 \end{cases}, \qquad \sum_i k_i^{\dagger} = \mathbf{1}^T\mathbf{y}_0 - c. \tag{B.1}$$

Let us introduce some results from majorization theory [29] that we need to complete the proof. First, let the components of $\mathbf{x} \in \mathbb{R}^n$ be ordered in decreasing order and express it as

$$x_{[1]} \ge \cdots \ge x_{[n]}. \tag{B.2}$$

Then, it is said [29, 1.A.1] that a vector $\mathbf{y}$ majorizes a vector $\mathbf{x}$ (which we denote by $\mathbf{y} \succ^M \mathbf{x}$), $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ if

$$\begin{aligned} \sum_{i=1}^{k} x_{[i]} &\le \sum_{i=1}^{k} y_{[i]}, \quad k = 1, \ldots, n-1, \\ \sum_{i=1}^{n} x_{[i]} &= \sum_{i=1}^{n} y_{[i]}. \end{aligned} \tag{B.3}$$

From the definition above and the construction process of $\mathbf{k}^{\dagger}$, we can state that $\mathbf{k}^{\dagger} \succ^M \mathbf{k}^{\triangleright}$.

Second, a real-valued function ϕ on a set $\mathcal{A} \subseteq \mathbb{R}^n$ is called Schur-convex if [29, 3.A.1]

$$\mathbf{y} \succ^M \mathbf{x} \text{ on } \mathcal{A} \Longrightarrow \phi(\mathbf{y}) \ge \phi(\mathbf{x}). \tag{B.4}$$

And third, a function $\phi(\mathbf{x}) = \sum_i g(x_i)$, where g is convex, is Schur-convex [30, Corollary 3.1].

With those results in hand, we want to compare $\|\mathbf{y}_0 - \hat{\boldsymbol{y}}\|^2$ for $\mathbf{k} = \mathbf{k}^{\triangleright}$ and $\mathbf{k} = \mathbf{k}^{\dagger}$. Let us rewrite the quadratic norm as

$$\|\mathbf{y}_0 - \hat{\boldsymbol{y}}\|^2 = \|\mathbf{y}_0 - \mathbf{y}_0 + \mathbf{k}\|^2 = \sum_i k_i^2 \tag{B.5}$$

and consider $\phi(\mathbf{k}) = \sum k_i^2$, which is a Schur-convex function. Finally, since $\mathbf{k}^{\dagger} \succ^M \mathbf{k}^{\triangleright}$, we have

$$\|\mathbf{k}^{\dagger}\|^2 \ge \|\mathbf{k}^{\triangleright}\|^2, \tag{B.6}$$

and thus, any solution where one element within $\mathbf{k}$ is negative is not optimal (since the problem is convex and has a single solution). The proof ends by induction of this result to an arbitrary number of negative elements in $\mathbf{k}$.

Notation

$U_i(r_i; p_i, \alpha)$: Utility achieved when entity i transmits at rate r_i. The utility is parameterized by a priority p_i (entity-dependant) and a shape factor α (common to all utilities)

N: Number of SSs

N_i: Number of active connections at the ith SS

r_j^i: Rate of the jth ongoing connection at the ith SS

m_j^i: Minimum guaranteed rate to the jth ongoing connection at the ith SS

d^i_j: Requested rate of the jth ongoing connection at the ith SS

C: Maximum outgoing rate at the BS

$\boldsymbol{\rho}_i$: Subchannel allocation vector at the ith SS

$\mathbf{c}_i$: Achievable rates at the ith SS (includes all subchannels)

$c^i(\boldsymbol{\rho}_i)$: Maximum outgoing rate at the ith SS

Γ: Subchannel allocation matrix: $\Gamma = [\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_N]$

$\mathcal{R}$: Feasible rates subset: $\mathcal{R} = \{r^i_j | m^i_j \leq r^i_j \leq d^i_j\}$

$\mathcal{S}$: Feasible allocations subset: $\mathcal{S} = \{\boldsymbol{\rho}_i | \Gamma\mathbf{1} \preccurlyeq \mathbf{1}, \boldsymbol{\rho}_i \succcurlyeq \mathbf{0}\}$

Acknowledgments

This work was supported in part by the Spanish Ministry of Science and Innovation under TEC200806305 project.

References

[1] IEEE, "Air Interface for Fixed Broadband Wireless Access Systems," IEEE Standards, October 2004.

[2] IEEE, "Air Interface for Fixed and Mobile Broadband Wireless Access Systems; Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Band and Corrigendum 1," IEEE Standards, February 2006.

[3] ETSI, "Broadband Radio Access Networks (BRAN); HIPERMAN; Data Link Control (DLC) Layer," *ETSI TS 102 178*, March 2003.

[4] J. G. Andrews, A. Ghosh, and R. Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2007.

[5] B. Makarevitch, "Adaptive resource allocation for WiMAX," in *Proceedings of the 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, pp. 1–6, Athens, Greece, September 2007.

[6] H.-Y. Wei, S. Ganguly, R. Izmailov, and Z. J. Haas, "Interference-aware IEEE 802.16 WiMax mesh networks," in *Proceedings of the 61st IEEE Vehicular Technology Conference (VTC '05)*, vol. 5, pp. 3102–3106, Stockholm, Sweden, May 2005.

[7] P. Du, W. Jia, L. Huang, and W. Lu, "Centralized scheduling and channel assignment in multi-channel single-transceiver WiMax mesh network," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '07)*, pp. 1736–1741, Hong Kong, March 2007.

[8] P. Soldati, B. Johansson, and M. Johansson, "Distributed optimization of end-to-end rates and radio resources in WiMax single-carrier networks," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '06)*, pp. 1–6, San Francisco, Calif, USA, November 2006.

[9] P. Soldati and M. Johansson, "Network-wide resource optimization of wireless OFDMA mesh networks with multiple radios," in *Proceedings of IEEE International Conference on Communications (ICC '07)*, pp. 4979–4984, Glasgow, UK, June 2007.

[10] L. B. Le and E. Hossain, "Joint rate control and resource allocation in OFDMA wireless mesh networks," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '07)*, pp. 3041–3045, Hong Kong, March 2007.

[11] K.-D. Lee and V. C. M. Leung, "Fair allocation of subcarrier and power in an OFDMA wireless mesh network," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 11, pp. 2051–2060, 2006.

[12] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2726–2737, 2005.

[13] K. Holmberg and K. C. Kiwiel, "Mean value cross decomposition for nonlinear convex problems," *Optimization Methods and Software*, vol. 21, no. 3, pp. 401–417, 2006.

[14] L. Reggiani, L. G. Giordano, and L. Dossi, "Multi-user subchannel, bit and power allocation in IEEE 802.16 systems," in *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC '07)*, pp. 3120–3124, Dublin, Ireland, April 2007.

[15] L. Boyd and S. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2003.

[16] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

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

[18] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1987.

[19] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[20] A. Muthoo, *Bargaining Theory with Applications*, Cambridge University Press, Cambridge, UK, 1999.

[21] H. Yaïche, R. R. Mazumdar, and C. Rosenberg, "A game theoretic framework for bandwidth allocation and pricing in broadband networks," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 667–678, 2000.

[22] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, Mass, USA, 1999.

[23] L. S. Lasdon, *Optimization Theory for Large Systems*, Dover, New York, NY, USA, 2002.

[24] S. H. Low and D. E. Lapsley, "Optimization flow control—I: basic algorithm and convergence," *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, pp. 861–874, 1999.

[25] J.-W. Lee, M. Chiang, and A. R. Calderbank, "Network utility maximization and price-based distributed algorithms for rate-reliability tradeoff," in *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM '06)*, pp. 1–13, Barcelona, Spain, April 2006.

[26] D. P. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: framework and applications," *IEEE Transactions on Automatic Control*, vol. 52, no. 12, pp. 2254–2269, 2007.

[27] A. Morell, G. Seco-Granados, and M. A. Vázquez-Castro, "Computationally efficient cross-layer algorithm for fair dynamic bandwidth allocation," in *Proceedings of the 16th International Conference on Computer Communications and Networks (ICCCN '07)*, pp. 13–18, Honolulu, Hawaii, USA, August 2007.

[28] A. Morell, G. Seco-Granados, and J. L. Vicario, "Distributed algorithm for uplink scheduling in WiMAX networks," in *Proceedings of the 5th International Conference on Broadband*

Communications, Networks, and Systems (BROADNETS ’08), pp. 257–264, London, UK, September 2008.

[29] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, NY, USA, 1979.

[30] D. P. Palomar, *A unified framework for communications through MIMO channels*, Ph.D. dissertation, Technical University of Catalonia (UPC), Barcelona, Spain, May 2003.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 512507, 14 pages
doi:10.1155/2009/512507

Research Article

Fairness and QoS Guarantees of WiMAX OFDMA Scheduling with Fuzzy Controls

Chao-Lieh Chen,[1] Jeng-Wei Lee,[2] Chi-Yuan Wu,[2] and Yau-Hwang Kuo[2]

[1] *Department of Electronic Engineering, National Kaohsiung First University of Science and Technology, Kaohsiung 811, Taiwan*
[2] *Department of Computer Science & Information Engineering, National Cheng Kung University, Tainan 701, Taiwan*

Correspondence should be addressed to Chao-Lieh Chen, frederic@ieee.org

Received 1 July 2008; Revised 8 December 2008; Accepted 29 December 2008

Recommended by Ekram Hossain

A fairness and QoS guaranteed scheduling approach with fuzzy controls (FQFCs) is proposed for WiMAX OFDMA systems. The controllers, respectively, adjust priority and transmission opportunity (TXOP) for each WiMAX connection according to QoS requirements and service classes. The FQFC provides intra- and interclass fairness guarantees by making connections within the same class achieve equal degree of QoS while at the same time making those without QoS requirements equally share the remaining resources. Even in dynamic environments such as mobile WiMAX networks with time-variant traffic specifications, the FQFC fairly guarantees delay, throughput, and jitter, which are seldom achieved at the same time by state-of-the-art solutions.

1. Introduction

IEEE 802.16 standard (WiMAX) [1, 2] is one of the most popular standards for fixed and mobile broadband wireless access systems to provide last mile access. Due to various users with diverse QoS requirements and wireless communication technologies, the resource scheduler plays an important role to provide fairness and QoS guarantees. As summarized in [3], a resource scheduler in wireless multimedia networks needs to possess the following features: efficient link utilization, delay bound, low implementation complexity, throughput, scalability, and fairness.

For WiMAX and OFDMA systems, various scheduling algorithms have been proposed for achieving QoS guarantees. For example, Liu et al. [4] proposed a priority-based scheduler which assigns each connection a priority updated according to QoS parameters and channel state and then assigns time slots to connections according to the order of priority values. The method has low implementation complexity because the scheduler simply updates the priority of each connection per frame and allocates time slots to the connection with the highest priority. However, it does not consider fairness and jitter issues which are important metrics for real-time applications. To maintain low implementation complexity under considering fairness and jitter, we use the priority-based scheduling scheme for initial priority assignment and afterward, the proposed a fairness and QoS guaranteed scheduling approach with fuzzy controls (FQFC) mechanism takes care of the scheduling job using fuzzy control approach. Many algorithms have been proposed to deal with the fairness problem, and can be briefly divided into two categories. The first category is to reduce the resource allocation problem into an optimization problem. Based on the optimization theory, for example, [5, 6] have good performance on spectrum efficiency and system utilization. They formulate the cross-layer optimization problem to maximize the average utility of all active users subject to certain constraints. However, in addition to implementation complexity, these methods still suffer some problems. To achieve optimal spectrum efficiency, the optimization approaches may, on the other hand, fail to provide QoS guarantees. Moreover, the relation between traffic specifications and network state is uncertain. Uncertainty and dynamics in mobile environment make exact modeling of objective function and constraints impossible when performing the optimization steps. In this paper, the FQFC adopts fuzzy control technique to deal with the modeling problem. The reason we use fuzzy control is to tackle uncertainty and dynamics in wireless communication environment. Among soft computing methods, inference based on probability theory is also widely used for modeling uncertainty and dynamics. However, the controller based on probability must rely on statistical observations to perform inference. Correctness of statistical information is based on the law of large number. In case that gathering large amount of statistic information in short time is difficult, it will be infeasible. Moreover, inference based on probability usually assumes some specific probability model for result of feedback observation to follow. As shown in [7], a single model usually fails to represent the behaviors of dynamic environments such as mobile wireless networks with sudden bursts or changes.

The second category is a utility-based scheduler. A utility function is a measure of relative satisfaction from users' requirements. The schemes in [8, 9] apply utility functions to maximize the total utility of all connections. Utility-based optimization approaches did guarantee QoS of some connections but also starve others. On the other hand, some approaches such as [10, 11] propose utility-fair bandwidth adaptation schemes for multiclass traffic in wireless networks. Rather than achieving resource fairness, the bandwidth adaptation schemes make sure that all connections can obtain similar utility values to achieve the so-called utility fairness. These schemes are effective in both achieving utility fairness and increasing network resource utilization. However, the utility-fair schemes may fail to provide QoS guarantees since it does not consider the priority of the connections.

In this paper, our objective is to provide efficient control for both QoS and fairness guarantees of WiMAX OFDMA scheduling. For QoS guarantees, we address the problem of head-of-line (HOL) delay and jitter control for real time applications and throughput control for nonreal-time applications. The FQFC scheduler assigns each connection a priority and TXOP, and adjusts them according to channel quality, QoS requirements, and service classes. Due to uncertainty and dynamics of the environment, it is difficult to find out the mapping between priority and QoS requirements. For fuzzy inference, it is the simplest way to model a complex system when there is few and uncertain information available. In the field of controller design, fuzzy controller is one of the most popular approaches. Moreover, fuzzy control has been widely used in researches on communication networks such as [7, 12–17]. However, there are few articles talking about using fuzzy control for WiMAX. In this paper, the FQFC model is developed for WiMAX OFDMA systems and is proved that both fairness and QoS are guaranteed. Other fuzzy control methods [14–17] may be proved to achieve certain degree of QoS. However, fairness is seldom assured in these and state-of-the-art approaches. Then, we define two types of fairness including intraclass and interclass fairness. To achieve intraclass fairness, we set up a reference goal to each connection according to the QoS requirements, and make the connections achieve the goal by priority scheduling and TXOP allocation. If each connection can achieve its QoS requirement, intraclass fairness is guaranteed. For achieving interclass fairness, the FQFC does not allocate superfluous resources out of what required. Compared to state-of-the-art methods, connections of high-priority classes release more resources to lower priority ones. Thus, the FQFC makes the connections without QoS requirements evenly share the remaining resources. Based on the priority scheduling and TXOP allocation methods, the FQFC provides both intraclass and interclass fairness with QoS guarantees and featuring low implementation complexity.

This paper is organized as follows. In Section 2, we introduce background including network configuration, MAC QoS, PHY resource allocation, and fairness descriptions. Section 3 describes the FQFC mechanism and depicts the design of the fuzzy controllers for each service class. In Section 4, we investigate the mechanism performance of QoS and fairness through simulations. Finally, we conclude the paper in Section 5.

2. Background

2.1. Network Configuration. WiMAX specifies two communication modes which form different topologies—point-to-multipoint (PMP) and mesh modes. In PMP mode, a base station (BS) centrally allocates downlink (from BS to SS) and uplink (from SS to BS) resources to subscriber stations (SSs). All SSs are only allowed to communicate with a BS. In mesh mode, SS can act as a router to assist its neighbor to relay data. In the 802.16 standard, this mode is optional and is not discussed in this paper. Hence, we focus on proposing a downlink scheduling algorithm to provide QoS guarantees in PMP mode.

IEEE 802.16 WiMAX PHY adopts the orthogonal frequency-division multiple access (OFDMA) technology based on OFDM modulation. The OFDMA technology allows multiple users transmitting packets at the same OFDMA symbol via different subchannels, such that wireless resources are utilized ultimately.

2.2. Scheduling Services in MAC Layer. IEEE 802.16 MAC protocol is connection-oriented; each connection is assigned a connection ID (CID) and a single scheduling service determined by a set of QoS parameters. Four scheduling services in the 802.16 standard are supported: unsolicited grant service (UGS), real-time polling service (rtPS), nonreal-time polling service (nrtPS), and best effort (BE). The UGS supports real-time constant bitrate data streams, such as voice over IP (VoIP) without silence suppression. The QoS parameters of UGS service are minimum reserved traffic rate, the tolerated jitter, maximum latency, and request/transmission policy. The rtPS supports real-time variable-rate data streams, such as MPEG video or VoIP with silence suppression. The QoS parameters of rtPS are maximum latency, request/transmission policy, minimum reserved traffic rate, and traffic priority. The nrtPS supports delay-tolerant variable-rate data streams, such as FTP. The QoS parameters of nrtPS are minimum reserved traffic rate, request/transmission policy, and traffic priority. The

BE supports best-effort data streams. The QoS parameter is request/transmission policy. In IEEE 802.16e [2], an additional service class called extended real-time polling service (ertPS) has superior efficiency than both UGS and rtPS. It supports real-time variable-rate data streams, such as VoIP with silence suppression. The QoS parameters of ertPS are minimum reserved traffic rate, maximum latency, request/transmission policy, and the tolerated jitter. Hence, considering the QoS requirements of the four class services, we calculate the reference goal as traffic specification (TSPEC) according to these QoS parameters.

2.3. Resource Allocations in PHY Layer. IEEE 802.16 OFDMA system defines two types of subcarrier permutations, distributed subcarrier permutation and adjacent subcarrier permutation. The former permutation type includes partially and fully used subcarriers (PUSC and FUSC) which are pseudo-randomly selected and grouped into subchannels, while the later includes adaptive modulation and coding (AMC), and only adjacent subcarriers are clustered to form subchannels. Dispersing noise and interference in fast changing environment, the PUSC and FUSC modes are suitable for mobile networks. For AMC mode, the BS allocates appropriate subchannels for connections with larger SNR to enhance system performance, and it is suitable for fixed or low mobility environment. To support mobile WiMAX, the FQFC scheduling and allocation are based on distributed subcarrier permutation.

In OFDMA, the basic allocation unit is a slot that composes of one subchannel along with an OFDMA symbol, such that the resource allocation becomes a two-dimensional problem. By using the distributed subcarrier permutation, all subchannels are the so-called *equally adequate* for all SSs [18], and our resource allocation is based on Raster algorithm [18], in which the frame is filled row by row, from left to right and from top to bottom, and efficiently reduces the burst numbers.

2.4. Fairness. In wireless networks, the fairness definition is not straightforward. As described in [19], a fair resource allocation usually does not produce equal connection data rate because the diverse connections also suffer from diverse channel conditions, network states, and dynamics. The dynamics result from mobility and time-variant traffic specifications (TSPECs). Moreover, WiMAX needs to provide QoS guarantees for four classes of scheduling services. Therefore, for fairness, it is necessary to consider QoS guarantees for different class connections. We define two types of fairness described as follows.

(i) Intraclass fairness: the connections within the same class achieve equal degree of QoS.

(ii) Interclass fairness: the connections with QoS requirements achieve exactly their demands, and those without QoS requirements equally share the remaining resources.

Hence, our objective is to achieve both intraclass and interclass fairness.

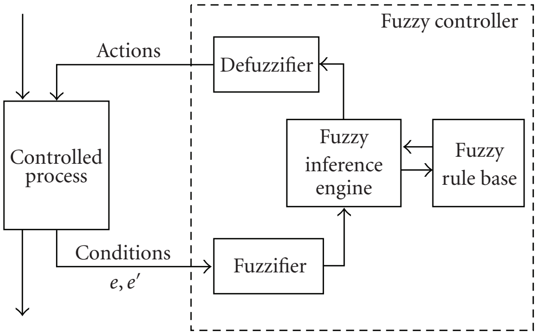


Figure 1: A general architecture of a fuzzy controller.

2.5. Fuzzy Controller. Classical controller requires modeling of the physical reality. This is significant in control problems; however, it is difficult or even impossible to construct precise mathematical models. The difficulty may result from time-variant system behaviors, dynamics, and uncertainty in mobile wireless communication environment. Fuzzy controllers perform well under these circumstances. A general fuzzy controller consists of four components: a fuzzifier, a fuzzy rule base, a fuzzy inference engine, and a defuzzifier. The interconnections among these components and the controlled process are shown in Figure 1. The fuzzifier maps crisp input into appropriate fuzzy sets to express uncertainties. The fuzzy inference engine uses the fuzzified measurements to evaluate the fuzzy implication results. Finally, the defuzzifier deals with confliction of fuzzy implications and transforms the fuzzy implication results back to the crisp output. Two conditions are usually monitored by the controller: error e and the derivative of the error e'. With e and e', the fuzzy controller issues control actions.

3. Design of the Proposed Scheduling Mechanism

In this section, we describe the scheduling mechanism for multiple connections with various QoS requirements. The FQFC scheduler assigns two variables with fuzzy inference values for each connection with CID i, that is, the priority P_i and the maximum number of packets TXOP_i that connection i can transmit in a frame duration. The FQFC scheduler first initializes the two variables based on the characteristics of connections and adjusts them, respectively, by two fuzzy controllers to adapt to the dynamics of system. As shown in Figure 2, the priority controller adjusts P_i according to channel quality, QoS requirements, and service classes. With the priority, the FQFC decides the transmission order of connections. The TXOP controller adapts TXOP_i according to transmission rate and the queue length difference between two contiguous transmissions of the MAC layer.

3.1. Controller Design for ertPS & rtPS. Unlike the UGS class having the highest priority that constant bandwidth can be

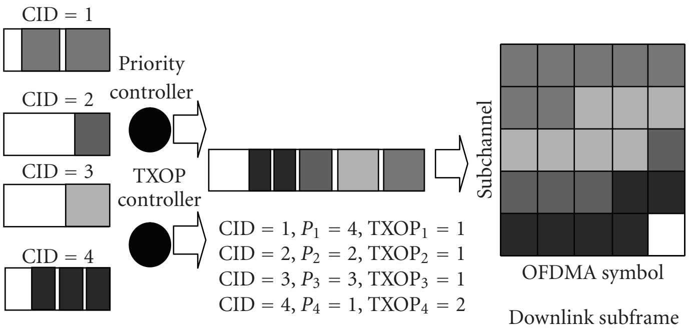

CID = 1
Priority controller
CID = 2
CID = 3
TXOP controller
CID = 4
CID = 1, P_1 = 4, TXOP_1 = 1
CID = 2, P_2 = 2, TXOP_2 = 1
CID = 3, P_3 = 3, TXOP_3 = 1
CID = 4, P_4 = 1, TXOP_4 = 2
Subchannel
OFDMA symbol
Downlink subframe

Figure 2: The proposed scheduling mechanism.

achieved by allocating fixed number of slots [1], the two service classes, rtPS and ertPS, that both support the real-time variable bit-rate data streaming require efficient and effective control to achieve QoS guarantees and fairness. A real-time connection of these two classes usually has two QoS specifications, maximum allowable latency (deadline) and jitter. The FQFC control for real-time connections comprises three steps: (1) set up goal delay and tolerable range, (2) adjust priority according to recent HOL delay, and (3) adjust TXOP according to the jitter requirement. The main idea is that the FQFC maintains the delay and jitter of each connection below the delay and jitter goals, respectively. First, we set the goal delay and the tolerable range. Figure 3 shows the control mechanism for real-time connections. Goal delay controller determines goal delay and tolerable range bounded by the lower and upper bounds. Then, the priority controller decides transmission order of connections, and the TXOP controller decides the number of transmitted packets to maintain the jitter. We describe the design of the three controllers as follows.

3.1.1. Goal Delay Controller. The purpose of the goal delay controller is to control delay and jitter within a tolerable range. If the delay exceeds the tolerable range, the FQFC increases the priority. If the delay is below the tolerable range, the FQFC decreases the priority. As shown in Figure 2, to avoid packet dropping, the goal delay is below the deadline. Since system load and transmission rate affect HOL delay obviously, we use them to decide the goal delay. Due to uncertainty and that the TSPEC changes rapidly in mobile WiMAX environment, we cannot use exact formulation to represent the goal delay. Therefore, we divide the delay space into three parts and use fuzzy sets S (small), M (medium), and L (large) to represent these three parts, respectively. Then, we decide which part that the goal delay belongs to according to the system load and transmission rate. The goal delay controller selects a goal delay and sets its upper and lower bounds to form a tolerable range for control. We denote $g_i(t)$, $g_i^{\text{up}}(t)$, and $g_i^{\text{low}}(t)$ as the goal delay of connection i in the tth frame and its upper and lower bounds, respectively. The goal delay controller uses triangular and trapezoidal membership functions as shown in Figure 4. The fuzzy input variables are system load (SL) and transmission rate (TR), and the output function is the goal delay (GD). The fuzzy sets of SL, TR, and GD are defined as follows:

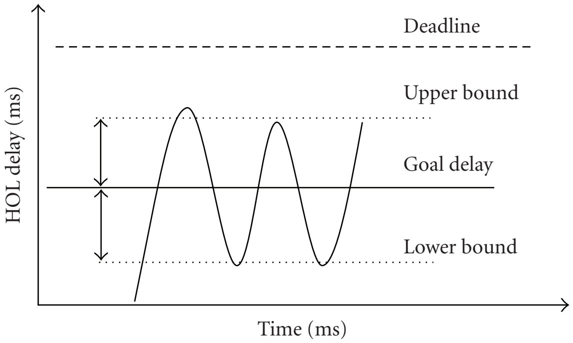


Figure 3: Control mechanism for real-time connection.

$$T(\text{SL}) = \{\text{Low}, \text{Medium}, \text{High}\} = \{\text{L}, \text{M}, \text{H}\},$$
$$T(\text{TR}) = \{\text{Fast}, \text{Medium}, \text{Slow}\} = \{\text{F}, \text{M}, \text{S}\},$$
$$T(\text{GD}) = \{\text{Small}, \text{Medium}, \text{Large}\} = \{\text{S}, \text{M}, \text{L}\}.$$

According to system load, the controller decides the goal delay $g_i^{\text{load}}(t)$ by the following fuzzy rules.

(R1) If system load is L, then $g_i^{\text{load}}(t)$ is S.

(R2) If system load is M, then $g_i^{\text{load}}(t)$ is M.

(R3) If system load is H, then $g_i^{\text{load}}(t)$ is L.

The following controller uses normalized data rate with respect to the transmission rate in the highest modulation mode to decide the goal delay $g_i^{\text{TX}}(t)$.

(R1) If transmission rate is F, then $g_i^{\text{TX}}(t)$ is S.

(R2) If transmission rate is M, then $g_i^{\text{TX}}(t)$ is M.

(R3) If transmission rate is S, then $g_i^{\text{TX}}(t)$ is L.

Using Mandamni implication and the centroid defuzzifier, we obtain the outputs, $g_i^{\text{load}}(t)$ and $g_i^{\text{TX}}(t)$. Considering system load and transmission rate, the final goal delay is

$$g_i(t) = g_i^{\text{load}}(t) \times w_1 + g_i^{\text{TX}}(t) \times w_2, \qquad (1)$$

where w_1 and w_2 are the weighting factors of system load and transmission rate, respectively.

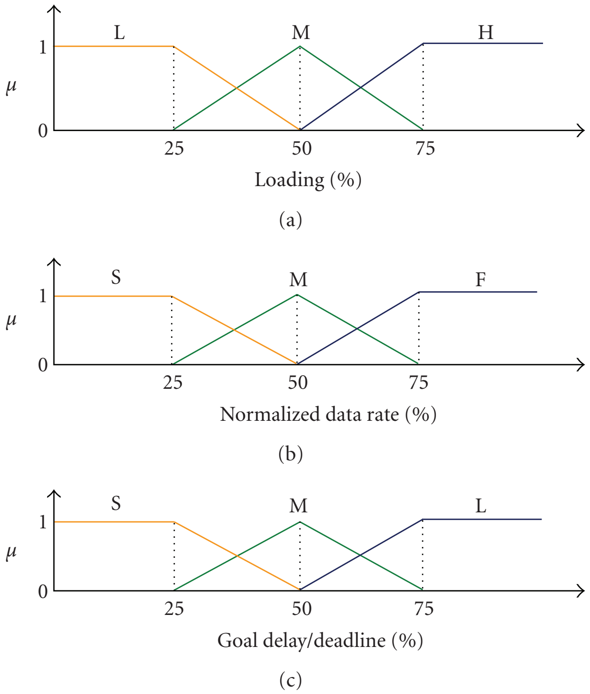


Figure 4: Membership functions of fuzzy sets for goal delay.

With goal delay $g_i(t)$ and required jitter $j_i(t)$, we define $g_i^{\text{up}}(t)$ and $g_i^{\text{low}}(t)$ as the upper and lower bounds of the tolerable range, where $g_i^{\text{up}}(t) = g_i(t) + j_i(t)/2$ and $g_i^{\text{low}}(t) = g_i(t) - j_i(t)/2$.

3.1.2. Priority Controller for Real Time Services. Figure 5 shows the control system including the priority controller, the WiMAX system plant, and the delay observer. The delay observer detects the HOL delay $d_i(t)$. Then, the priority controller compares it with the delay requirement $g_i(t)$, and adjusts priority $P_i(t)$. If $e_i(t) = d_i(t) - g_i(t)$ is around zero, the control system is stabilized around the requirement.

In our design, we denote negative, zero, and positive forces with fuzzy singletons S, M, and L. The control actions of these singletons at the conclusion parts of fuzzy rules are as follows:

$$\begin{aligned} \text{S: } & P_i(t) = P_i(t-1) - \delta_i(t), \\ \text{M: } & P_i(t) = P_i(t-1), \\ \text{L: } & P_i(t) = P_i(t-1) + \delta_i(t), \end{aligned}$$

where $\delta_i(t)$ is the priority influence of connection i in the tth frame. The priority controller must confirm that the HOL delay will not exceed the deadline. Hence, it adapts $\delta_i(t)$ according to the time duration between goal delay and the deadline. Let D_i be the deadline, ΔD_i be the guard time before the deadline, P_{rtPS} be the maximum priority of real-time connections, and t_{frame} be the frame duration. Then, we have

$$\delta_i(t) = \frac{P_{\text{rtPS}}}{(D_i - \Delta D_i - g_i(t))/t_{\text{frame}}}. \tag{2}$$

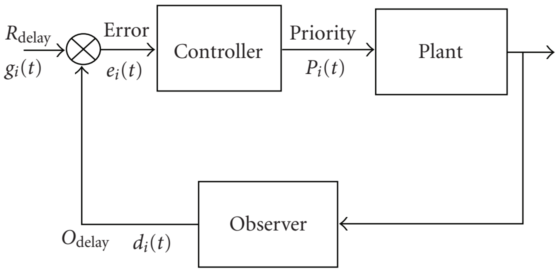


Figure 5: The block diagram of the control system for HOL delay.

As we can see in (2), when the goal delay is closer to the deadline, the adaptation force of the priority is larger. We depict the priority initialization and controlled direction as follows.

(a) *Priority Setting.* When the connection is in an initial stage or the HOL delay is below $g_i^{\text{low}}(t)$, the priority controller assigns the connection a priority according to channel quality, QoS requirement, and service classes. For a real-time connection i, the priority $P_i(t)$ in the tth frame is assigned by (3) which was proposed in [4]:

$$P_i(t) = \begin{cases} P_{\text{rtPS}} \times \dfrac{r_i(t)}{R_i^{\max}} \times \dfrac{1}{F_i(t)}, & \text{if } F_i(t) \geq 1,\ r_i(t) \neq 0, \\ P_{\text{rtPS}}, & \text{if } F_i(t) < 1,\ r_i(t) \neq 0, \\ 0, & \text{if } r_i(t) = 0, \end{cases} \tag{3}$$

where P_{rtPS} is the maximum priority of real-time connections, $R_i^{\max}$ is the data rate of connection i in the highest modulation mode, and $r_i(t)$ is the data rate of connection i in the tth frame. $r_i(t)/R_i^{\max}$ is the normalized data rate and the connection with high received SNR results in higher priority. $F_i(t)$ is the delay requirement indicator:

$$F_i(t) = D_i - \Delta D_i - d_i(t) + 1, \tag{4}$$

where D_i is the deadline, ΔD_i is the guard time before the deadline, and $d_i(t)$ denotes the HOL delay. If $F_i(t) \geqq 1$, the larger $F_i(t)$ denotes the higher satisfaction of delay requirement, which causes lower priority. If $F_i(t) < 1$, the HOL delay has been over the guard time of deadline. The connection should get resources immediately to avoid packet losses. Hence, the priority is set as P_{rtPS}. When $r_i(t)$ is zero, the connection i is under deep fading and should not be scheduled.

(b) *Priority Controller.* Let the controller action be the priority $P_i(t)$. One of the input $e_i(t)$ is the difference between the actual value of the observed HOL delay $d_i(t)$ and the desired value $g_i(t)$, that is, $e_i(t) = d_i(t) - g_i(t)$. The universe of $e_i(t)$ is $[-g_i(t), D_i(t) - g_i(t)]$. The variable $e_i(t)$ has three linguistic values N, E, and P which represent fuzzy concepts "Negative," "Equal," and "Positive," respectively. The fuzzy sets N, E, and P are characterized by the membership functions shown in Figure 6.

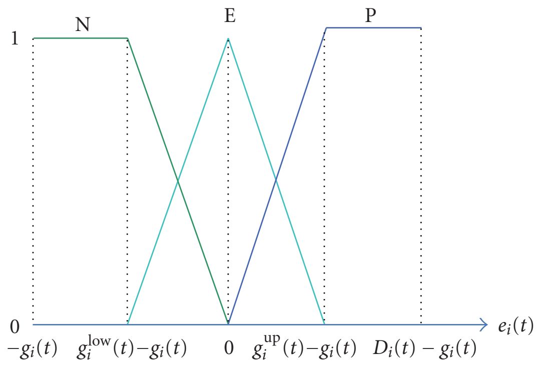
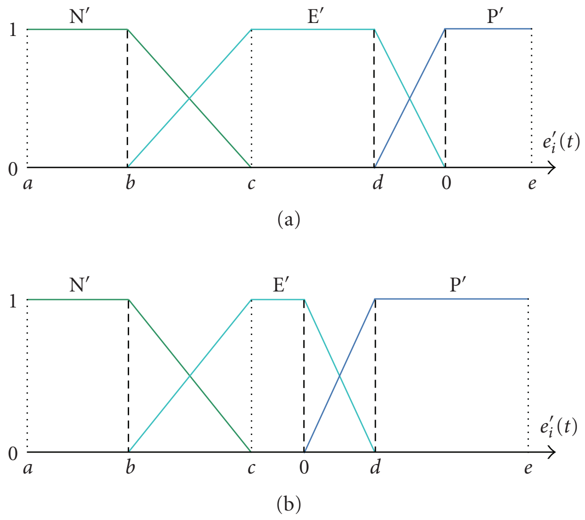


FIGURE 6: Membership functions of linguistic values for $e_i(t)$.



(a)

(b)

FIGURE 7: Membership functions of linguistic values for $e_i'(t)$.

The other input of the controller is the difference between two errors, which is defined as $e_i'(t) = e_i(t) - e_i(t-1)$. Substituting $e_i(t) = d_i(t) - g_i(t)$ to $e_i'(t)$, we obtain $e_i'(t) = d_i(t) - d_i(t-1)$. The universe of $e_i'(t)$ is $[-d_i(t-1), D_i(t) - d_i(t-1)]$. The linguistic values of $e_i'(t)$, N′, E′, and P′ also representing fuzzy concepts "Negative," "Equal," and "Positive," respectively, are characterized by the membership functions as shown in Figure 7, where $a = -d_i(t-1)$, $b = g_i^{\text{low}}(t) - d_i(t-1)$, $c = g_i(t) - d_i(t-1)$, $d = g_i^{\text{up}}(t) - d_i(t-1)$, and $e = D_i(t) - d_i(t-1)$. Sign of these values constitutes four cases as shown in Figure 7. The membership functions are time-variant and change along with the variable $d_i(t-1)$.

We consider four cases to design the fuzzy rule base as follows.

Case 1. If HOL delay is too large, that is, $d_i(t-1) > g_i^{\text{up}}(t)$, the priority should be increased with the large (L) step.

Case 2. If $g_i^{\text{up}}(t) > d_i(t-1) > g_i(t)$, maintaining priority at the median (M) level is fine.

Case 3. If $g_i(t) > d_i(t-1) > g_i^{\text{low}}(t)$, maintaining priority at the median (M) level is fine.

Case 4. If HOL delay is too small, that is, $d_i(t-1) < g_i^{\text{low}}(t)$, the priority should be decreased with negative decrement (S).

Therefore, expanding the above cases with changing rate $e_i'(t)$, we have the linguistic inference rules

(R1) If $e_i(t)$ is P and $e_i'(t)$ is P′, then $P_i(t)$ is L,
(R2) If $e_i(t)$ is P and $e_i'(t)$ is E′, then $P_i(t)$ is M,
(R3) If $e_i(t)$ is P and $e_i'(t)$ is N′, then $P_i(t)$ is M,
(R4) If $e_i(t)$ is E and $e_i'(t)$ is P′, then $P_i(t)$ is M,
(R5) If $e_i(t)$ is E and $e_i'(t)$ is E′, then $P_i(t)$ is M,
(R6) If $e_i(t)$ is E and $e_i'(t)$ is N′, then $P_i(t)$ is M,
(R7) If $e_i(t)$ is N and $e_i'(t)$ is P′, then $P_i(t)$ is M,
(R8) If $e_i(t)$ is N and $e_i'(t)$ is E′, then $P_i(t)$ is M,
(R9) If $e_i(t)$ is N and $e_i'(t)$ is N′, then $P_i(t)$ is S.

Using Mandamni implication and the centroid defuzzifier, we obtain the control action responding each HOL delay $d_i(t)$.

The priority controller makes the delay fall in the tolerable range which is below the deadline. Hence, each connection in the real-time class achieves the QoS specification, while intraclass fairness is guaranteed. When the delay is below the tolerable range, the controller decreases the priority for releasing the resources. This scheme guarantees the jitter and interclass fairness at the same time.

(c) *Priority Adaptation for Fairness.* For making the connections within the same class achieve equal degree of QoS, the priority controller adapts $P_i(t)$ by further considering the packet loss rate. All connections should receive the same packet loss rate. To compensate the packet losses in the tth frame, we define the loss rate as $\text{loss}_i(t)$ and the scaling by

$$P_i(t) = \begin{cases} P_i(t), & \text{if } \text{loss}_i(t) = 0, \\ P_i(t) \times \max\left(\dfrac{\lambda}{-\log(\text{loss}_i(t))} 1\right), & \text{if } \text{loss}_i(t) > 0, \end{cases} \tag{5}$$

where λ is a constant to normalize the loss rate according to the predefined precision. According to (5), if the connection drops packets due to out of the deadline, the priority controller allocates more resources by increasing the priority for achieving intraclass fairness. Even if all connections are in an extremely bad environment, they will suffer the same loss rate.

3.1.3. TXOP Controller. The TXOP controller initiates the TXOP based on frame duration t_{frame} and packet interval of the ith connection t_{pi} as (6)

$$\text{TXOP}_i(0) = \left\lceil \frac{t_{\text{frame}}}{t_{\text{pi}}} \right\rceil. \tag{6}$$

According to deficit round robin [20], the TXOP increases as the number of packets in a queue increases. Let $Q_i(t)$

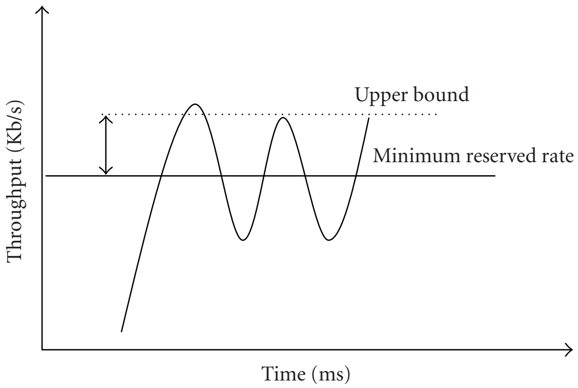


Figure 8: Control mechanism for nrtPS connection.

denote the number of packets in queue i in the tth frame. The controller stores the bounded difference $DC_i(t) = Q_i(t) \ominus Q_i(t-1) = \max(Q_i(t) - Q_i(t-1), -1)$ in the deficit counter of connection i in the tth frame. Then, we add $DC_i(t)$ to $\text{TXOP}_i(t)$as follows:

$$\text{TXOP}_i(t) = \max\left(\text{TXOP}_i(t-1) + DC_i(t), 0\right). \tag{7}$$

To avoid burst transmission, TXOP has an upper bound in

$$\text{TXOP}_i^{\text{up}}(t) = \left\lceil \frac{d_i(t) - g_i^{\text{low}}(t)}{t_{\text{pi}}} \right\rceil. \tag{8}$$

3.2. Controller Design for nrtPS. The nrtPS connection supports delay-tolerant variable-rate data streams and guarantees minimum reserved rate. The control mechanism for nrtPS connections can be divided into three steps: (1) setting up minimum reserved rate and an upper bound, (2) adjusting the priority according to average throughput of nrtPS connections and the required jitter of real-time connections, and (3) adjusting TXOP of nrtPS connections according to the required jitter. Figure 8 shows the control mechanism for nrtPS connection.

If the average throughput is lower than minimum reserved rate, the priority controller raises the priority to increase the throughput. Moreover, the controller needs to prevent large jitter from over-high priority. Besides, if the average throughput exceeds the upper bound, the controller decreases the priority to release the resource. We depict the controller design as follows.

3.2.1. Priority Controller. The priority controller in nrtPS class is easier than in the real-time class. The QoS requirement is only to guarantee the minimum reserved rate. Hence, we do not use fuzzy control and simply use the priority-based scheduler for nrtPS connections. Let $T_i(t)$ denote the average throughput of connection i in the tth frame, and let $R_i(t)$ denote the instantaneous data rate of connection i in the tth frame. The average throughput in the tth frame is usually estimated over a time constant t_c using moving average as

$$T_i(t+1) = \begin{cases} \left(1 - \dfrac{1}{t_c}\right) \times T_i(t) + \dfrac{1}{t_c} \times R_i(t), & \text{if } i = i^*, \\ \left(1 - \dfrac{1}{t_c}\right) \times T_i(t), & \text{if } i \neq i^*, \end{cases} \tag{9}$$

where i^* means connection i is scheduled in the tth frame.

For an nrtPS connection i, the priority $P_i(t)$ in the tth frame is defined as

$$P_i(t) = \begin{cases} P_{\text{nrtPS}} - \delta_i(t), & \text{if } F_i(t) \geq 1,\ r_i(t) \neq 0, \\ P_{\text{nrtPS}}, & \text{if } F_i(t) < 1,\ r_i(t) \neq 0, \\ 0, & \text{if } r_i(t) = 0, \end{cases} \tag{10}$$

where $\delta_i(t)$ is the priority decrement, P_{nrtPS} is the maximum priority of nrtPS connection, and $F_i(t)$ is the throughput requirement indicator which is the ratio of average throughput with respect to the minimum reserved rate $T_i^{\min}$

$$F_i(t) = \frac{T_i(t)}{T_i^{\min}}. \tag{11}$$

If $F_i(t) \geqq 1$, the throughput requirement is satisfied, and the controller decreases the priority to release resource. When $F_i(t) < 1$ implying that the average throughput is less than the minimum reserved rate, the connection should get more resources immediately to achieve the requirement. Hence, at this time, the priority is set to the maximum P_{nrtPS}. The priority decrement $\delta_i(t)$ is further defined as

$$\delta_i(t) = k \times \frac{\text{TXOP}_i(t) \times L_{\text{packet}}}{T_i^{\text{up}} - T_i^{\min}}, \tag{12}$$

where L_{packet} is the packet length, T_i^{up} is the upper bound of $T_i(t)$ which is the maximum sustained rate in the traffic specification, and k is a constant representing system load.

3.2.2. TXOP Setting. For nrtPS, the FQFC sets $\text{TXOP}_i(t)$ according to the throughput upper bound T_i^{up} as

$$\text{TXOP}_i(t) = \left\lceil T_i^{\text{up}} \times \frac{t_{\text{frame}}}{L_{\text{packet}}} \right\rceil. \tag{13}$$

3.2.3. TXOP Adaptation for Fairness. For intraclass fairness, all nrtPS connections should have the same throughput ratio of average throughput with respect to minimum reserved rate. Via setting the upper bound T_i^{up} in (13), we control the average throughput within the range between the minimum reserved rate and the upper bound, and we make the throughput ratio of all nrtPS connections the same. For interclass fairness, the average throughput will not exceed the upper bound. Hence, we can release more resources to the connections without QoS requirements.

3.3. Controller Design for BE

3.3.1. Priority Setting. For a BE connection i, the priority $P_i(t)$ in the tth frame is defined as

$$P_i(t) = \begin{cases} P_{\mathrm{BE}}, & \text{if } r_i(t) \neq 0, \\ 0, & \text{if } r_i(t) = 0, \end{cases} \tag{14}$$

where P_{BE} is the maximum priority of BE connection. All BE connections have the same priority. For intraclass fairness, we adopt the round robin scheduling for BE connections.

3.3.2. TXOP Setting. For fair resource allocation, the FQFC sets the $\mathrm{TXOP}_i(t)$ according to the frame duration t_{frame} and the packet interval of the ith connection t_{ip} as (14):

$$\mathrm{TXOP}_i(t) = \left\lceil \frac{t_{\mathrm{frame}}}{t_{\mathrm{ip}}} \right\rceil. \tag{15}$$

In this paper, we also perform priority adaptation. Therefore, the overhead, especially the complexity, will be slightly higher than that of the priority-only method. Since in centralized PMP mode, all traffic flows are managed by base stations which have much more powerful computing ability than SSs, the additional computation overhead will not give any sensibly negative effect. Moreover, the proposed controllers do not use any control/management packets for fairness and QoS purposes. There is no additional network overhead caused by the proposed FQFC.

4. Evaluations and Simulation Results

We first introduce intraclass and interclass fairness criteria and then according to these criteria, we evaluate the performances of the fairness.

4.1. Fairness Criteria. The descriptions of fairness indices are as follows.

4.1.1. Intraclass Fairness Index. Intraclass fairness means that the connections within the same class achieve equal QoS guarantees. Because the connections in different service classes have different QoS requirements, we define respective intraclass fairness indices for real-time, nrtPS, and BE classes.

(a) *Real-Time Connection.* A connection belonging to the real-time class requires strict maximum allowable latency (deadline) and the tolerated jitter. Packet loss occurs when packet delay is out of the deadline. Hence, we use loss rate and jitter to evaluate the intraclass fairness of real-time connections. We define a real-time indicator $I_{\mathrm{RT},i}$ as

$$I_{\mathrm{RT},i} = \begin{cases} 1, & \text{if } \mathrm{jitter}_i > \mathrm{jitter}_{\mathrm{tolerated}}, \\ \mathrm{loss}_i, & \text{if } \mathrm{jitter}_i \leq \mathrm{jitter}_{\mathrm{tolerated}}, \end{cases} \tag{16}$$

where loss_i, jitter_i, and $\mathrm{jitter}_{\mathrm{tolerated}}$ are the loss rate, jitter, and the tolerated jitter of connection i, respectively. If the jitter is larger than the tolerated jitter, the connection does not achieve QoS guarantees and we set the real-time indicator to one. Otherwise, we set the real-time indicator as the loss rate. Then, we utilize the real-time indicator to compute the real time fairness index. If the real-time indicators of all connections are closer to each other, the better intraclass fairness is achieved. We define the real-time fairness index $\mathrm{FI}_{\mathrm{RT}}$ as the standard deviation of the real-time indicators of all rtPS connections as follows:

$$\mathrm{FI}_{\mathrm{RT}} = \sqrt{\frac{1}{N_{\mathrm{RT}} - 1} \sum_{j=1}^{N_{\mathrm{RT}}} \left(I_{\mathrm{RT},j} - I_{\mathrm{RT,avg}}\right)^2}, \tag{17}$$

where N_{RT} is the number of connections in the real time class, and $I_{\mathrm{RT,avg}}$ is the average real-time indicator. Thus, a smaller value of $\mathrm{FI}_{\mathrm{RT}}$ represents better intraclass fairness of the real-time class.

(b) *nrtPS Connection.* A connection belonging to the nrtPS class requires minimum reserved rate. Hence, we use the average throughput to evaluate the intraclass class fairness of nrtPS connections. We define a nrtPS indicator $I_{\mathrm{nRT},i}$ as

$$I_{\mathrm{nRT},i} = \frac{T_i}{T_i^{\min}}, \tag{18}$$

where T_i and $T_i^{\min}$ are the average throughput and minimum reserved rate of connection i, respectively. Then, we introduce the throughput indicator to compute the nrtPS fairness index. The nrtPS fairness index $\mathrm{FI}_{\mathrm{nRT}}$ is defined as the standard deviation of the throughput indicator of connections in the same nrtPS class as follows:

$$\mathrm{FI}_{\mathrm{nRT}} = \sqrt{\frac{1}{N_{\mathrm{nRT}} - 1} \sum_{j=1}^{N_{\mathrm{nRT}}} \left(I_{\mathrm{nRT},j} - I_{\mathrm{nRT,avg}}\right)^2}, \tag{19}$$

where N_{nRT} is the number of connections in nrtPS class, and $I_{\mathrm{nRT,avg}}$ is the average nrtPS indicator. Similar to the $\mathrm{FI}_{\mathrm{RT}}$, a smaller $\mathrm{FI}_{\mathrm{nRT}}$ value represents better intraclass fairness of the nrtPS class.

(c) *BE Connection.* A connection belonging to BE requires no QoS metrics. We introduce the average throughput to compute the BE fairness index. The BE fairness index is defined as the standard deviation of the average throughput of connections in the same BE class i as follows:

$$\mathrm{FI}_{\mathrm{BE}} = \sqrt{\frac{1}{N_{\mathrm{BE}} - 1} \sum_{j=1}^{N_{\mathrm{BE}}} \left(T_j - T_{\mathrm{avg}}\right)^2}, \tag{20}$$

where N_{BE} is the number of connections in BE class, and T_{avg} is the average throughput in the BE class. Smaller $\mathrm{FI}_{\mathrm{BE}}$ represents better intraclass fairness of the BE class.

4.1.2. Interclass Fairness Index. According to the definition of interclass fairness, the interclass fairness has two folds: (1) the connections with QoS requirements achieve the demands;

(2) the connections without QoS requirements equally share the remaining resources.

For the first fold, we introduce a requirement indicator $I_{R,i}$ to show the degree of the connection close to the demands as

$$I_{R,i} = e^{-k|x_i - G_i|}, \tag{21}$$

where k is a tunable parameter which determines the tolerable range. x_i and G_i are the average state and the QoS goal of class i, respectively. In the real-time class, the average state is the mean loss rate, and its goal loss rate is zero. In the nrtPS class, the QoS parameter is the average throughput, and the goal is the minimum reserved rate. The smaller the difference between the average state and the QoS goal is, the larger requirement indicator is. When the mean allocated resources for a class are away from the requirement, no matter above or below the goal, the requirement indicator decreases. When the allocated resources reach the requirements exactly, not only the QoS is guaranteed but also the remaining resources are most preserved at the same time.

The BE class has no QoS requirement. For the second part, we introduce Jain's fairness index [21] as the BE fairness index:

$$I_{\mathrm{BE}} = \frac{\left(\sum_{i=1}^{n} T_i\right)^2}{\left(n \cdot \sum_{i=1}^{n} T_i^2\right)}, \tag{22}$$

where n is the number of connections without QoS requirements. The index equals to one indicates perfect fairness in the class without QoS requirements. Then, we utilize the requirement indicator $I_{R,i}$ and the BE fairness index I_{BE} to define the interclass fairness index as follows:

$$\begin{aligned} \mathrm{FI} &= \alpha \times \sum_{i=1}^{m} I_{R,i} w_i + \beta \times I_{\mathrm{BE}}, \\ &\sum_{i=1}^{m} w_i = 1, \\ &\alpha + \beta = 1. \end{aligned} \tag{23}$$

In (23), m is the number of classes with QoS requirements, and w_i is the weighting factor of class i, which determines the importance of the class. α and β are the weighting factors of the classes with and without QoS requirements, respectively. In contrast to the indices of intraclass fairness, a larger FI value indicates better interclass fairness.

4.2. Simulation Configuration. The parameters used in this simulation are listed in Table 1, where OFDM FFT size represents the number of subcarriers an OFDMA symbol composes. The packet length is 1024 bits, and the maximum priority of each service class is $P_{\mathrm{rtPS}} = 1.0$ and $P_{\mathrm{nrtPS}} = 0.8$, and for best effort, $P_{\mathrm{BE}} = 0.6$. The weighting factors w_i, α, β in (23) are all 0.5. Each connection uses a fixed modulation. The FQFC allocates fixed number of time slots to UGS connections. The FQFC adopts persistent resource allocation [1, 22, 23] for UGS service because it has the highest priority. We focus on the performance of real-time, nrtPS, and BE connections. Besides, in our survey, the priority-based scheduler was proposed only for WiMAX OFDM PHY [4]. FQFC outperforms many state-of-the art schedulers for WiMAX OFDM PHY. To present the improvement by FQFC, we modify the priority-based scheduler in [4] to work with WiMAX OFDMA PHY by using the Raster algorithm and regard it as the priority-only scheduler. Then, FQFC fuzzy controllers further improve the fairness and QoS performance of the priority-only scheduler. There are four simulation scenarios as follows.

(i) *Scenario* 1. We set 20 real-time connections. The QoS requirements of real-time connection are the loss rate, deadline, and required jitter. The traffic rates of connections are 8 connections in 1 Mbps, 10 connections in 500 kbps, and 2 connections in 250 kbps. This scenario is to verify the guarantees of maximum latency, the tolerated jitter, and the intraclass fairness in real-time class. It is difficult to find out the mapping between priority and QoS requirements. We prove that the FQFC can efficiently control the delay.

(ii) *Scenario* 2. We set 10 real-time connections and 10 nrtPS connections. The QoS parameter of nrtPS connection is the minimum reserved rate. The traffic rates are 2 real-time connections in 1 Mbps, 8 real-time connections in 500 kbps, 5 nrtPS connections in 1 Mbps, and 5 nrtPS connections in 500 kbps. This scenario is to verify the guarantees of minimum reserved rate and fair resource allocation of the FQFC scheme.

(iii) *Scenario* 3. We set 10 real-time connections, 10 nrtPS connections, and 10 BE connections. BE connection has no QoS requirement. The traffic rates are 1 real-time connection in 1 Mbps, 9 real-time connections in 500 kbps, 3 nrtPS connections in 750 kbps, 2 nrtPS connections in 500 kbps, 5 nrtPS connections in 1 Mbps, and 10 BE connections in 100 kbps. This scenario is to verify the fair resource allocation of FQFC.

(iv) *Scenario* 4. In this scenario, we simulate the wireless link degrades. This will cause the modulation to change. The experiment is designed to test the robustness of the FQFC whether it can efficiently track the goal delay when the channel quality degrades. The simulated network consists of 1 BS and 10 SS (numbered from 1 to 10). In the downlink, each SS with number i ($i = 1 \sim 10$) has 1 real-time, 1 nrtPS, and 1 BE connection with CID i, $10 + i$, $20 + i$, respectively. The connections from SS1 to SS5 apply with QPSK modulation, and connections from SS6 to SS7 apply with 16-QAM modulation. All the other connections initially adopt 64-QAM modulation. This is for simulating the different channel conditions.

Table 1: Simulation parameters.

Parameter	Value
System bandwidth	10 MHz
Frame duration	5 ms
OFDMA FFT size	1024
Number of subchannels	30
Number of OFDMA symbols for DL	28

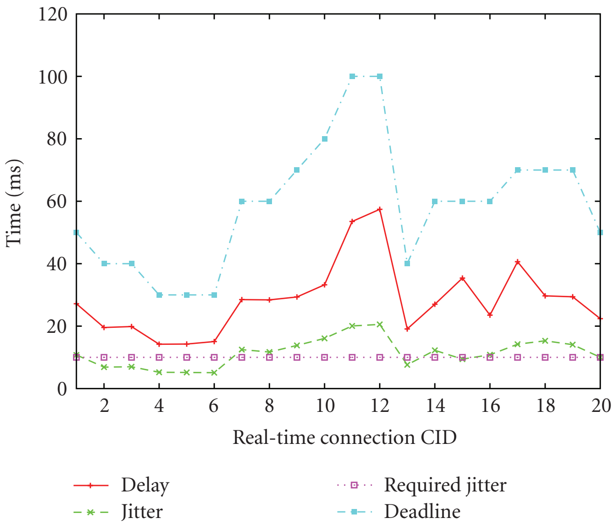


Figure 9: Delay and jitter performances of real-time connections in priority-only scheduler.

4.3. Performance Evaluation for Fairness and QoS Guarantees

4.3.1. Scenario 1: Intraclass Fairness and QoS Guarantees of Real Time Connections. We compare the FQFC with the priority-only scheduler [4]. The QoS is in terms of average delay, delay jitter, and packet loss rate. As illustrated in Figure 9, although the priority-only scheduler controls delay of connection to be below the deadline, it cannot guarantee tolerant jitter. Under the same simulation conditions, FQFC guarantees both delay and jitter requirements as shown in Figure 10. The average delay is close to the goal delay. The result also shows that it is useful by controlling the HOL delay in the tolerable range to guarantee the required jitter.

For intraclass fairness evaluation, from Figure 10, we can see that the jitter of the connections using FQFC is still smaller than the tolerated jitter in Figure 9. Figure 11 shows the delay outage probabilities of the FQFC and the priority-only scheduler. The FQFC disperses the outage probability for intraclass fairness. Moreover, as summarized in Table 2, the intraclass fairness index of the FQFC is much lower than the one of the priority-only scheduler and is almost near to zero. Hence, the FQFC guarantees the intraclass fairness for real-time connections.

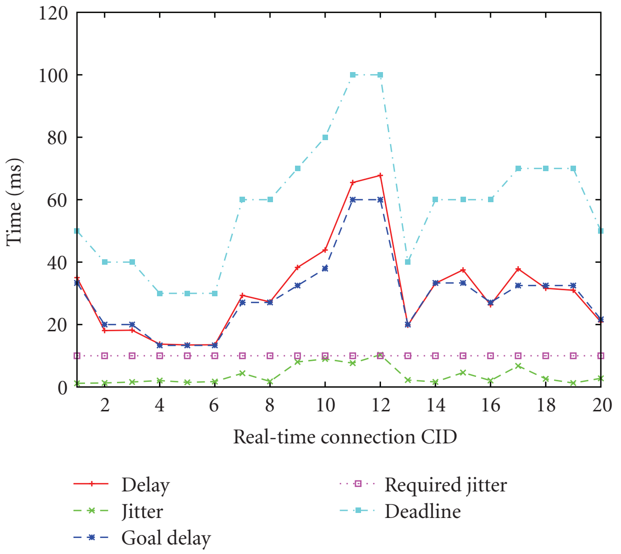


Figure 10: Delay and jitter performances of real-time connections in FQFC.

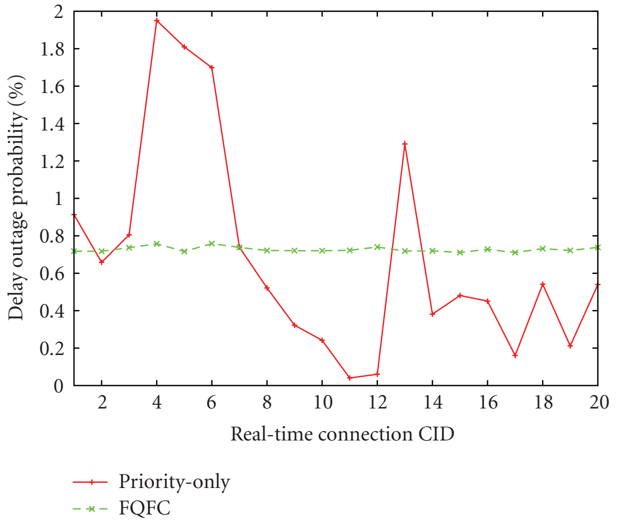


Figure 11: Average outage probability of rtPS connections.

Table 2: Intraclass fairness index.

		FQFC	Priority-only
Scenario 1	Real-time	0.000131	0.504639
Scenario 2	Real-time	0	0.489360
	nrtPS	0.012748	0.081995
Scenario 3	Real-time	0	0.502625
	nrtPS	0.003088	0.107002
	BE	6.686637	84.312975

4.3.2. Scenario 2: Intraclass Fairness and QoS Guarantees of Real-Time and nrtPS Connections. For QoS evaluation, we introduce the throughput indicator defined as the ratio of the average throughput with respect to the minimum reserved rate. In Figures 12 and 13, even adding nrtPS connections, the FQFC still guarantees the delay and jitter specifications of real-time connections. Then, we evaluate nrtPS connections

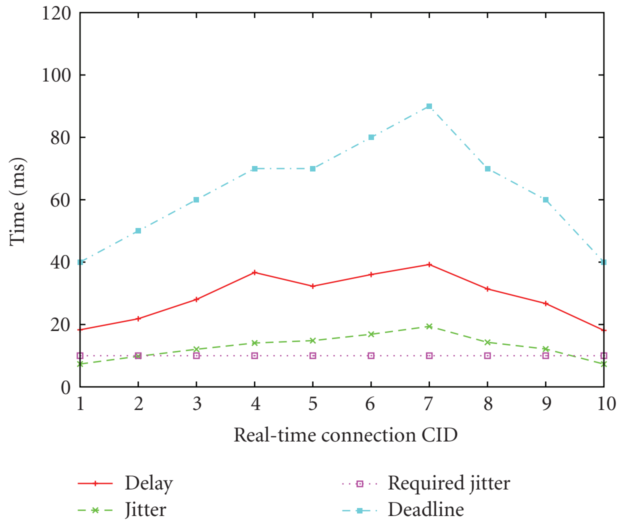


Figure 12: Delay and jitter performances of real-time connections in priority-only scheduler.

by throughput indicators. Figure 14 shows that all nrtPS connections with FQFC control keep their throughput indicators almost the same about 1.15. The result means that the FQFC guarantees minimum reserved rate. For the connections with priority-only-based scheduler [4], the throughput indicators of the last four nrtPS connections are higher than the others since priority-only-based scheduler provides more resources to the connections using high-bitrate modulation. The FQFC focuses on making the connections of the same class achieve the equal degree of QoS. As illustrated in Table 2, for nrtPS, the intraclass fairness index of the FQFC is close to zero which is much lower than the one of the priority-only scheduler. Hence, the FQFC also guarantees the intraclass fairness for the connections of the nrtPS class.

4.3.3. Scenario 3: Intra- and Interclass Fairness and QoS Guarantees of All Classes. For QoS evaluation, Figures 15, 16, and 17 show that the FQFC guarantees the delay and jitter of real-time connections as well as guarantees the throughput of nrtPS connections. Even for users with diverse QoS requirements, the FQFC still provides QoS guarantees. For BE connections, although they have no QoS requirement, the remnant resources should be fairly allocated to all BE connections. In Figure 18, BE connections under FQFC control obtain throughputs and are not starved.

For intraclass fairness evaluation of the BE class, we compare the FQFC with the priority-only scheduler regarding average throughput. Figure 18 shows that the average throughputs of all BE connections under the FQFC control are nearly the same. The priority-only scheduler provides more resource to the last four BE connections since they employ higher rate modulation. Table 2 shows that the intraclass fairness index of the FQFC is close to zero, which is much lower than the one of the priority-only scheduler. For every real-time connection, FQFC sets the goal delay below the deadline for a certain distance in terms of the tolerable jitter. Since the goal is for priority and TXOP controllers to follow, intraclass fairness is achieved when real-time connections have almost the same loss rate and jitter performances based on the intraclass fairness criteria. For nrtPS connections, the FQFC control algorithm maintains their ratios of throughput achievement over minimum reserved rate as close to 1 as possible. Again, as long as BE connections can evenly share the remained resources from real-time and nrtPS connections, intraclass fairness of BE connections is achieved. Hence, the FQFC guarantees the intraclass fairness for the BE classes.

For interclass fairness evaluation, in Figure 17, the throughput indicator of the FQFC is lower than the one of the priority-only scheduler since the FQFC always preserves

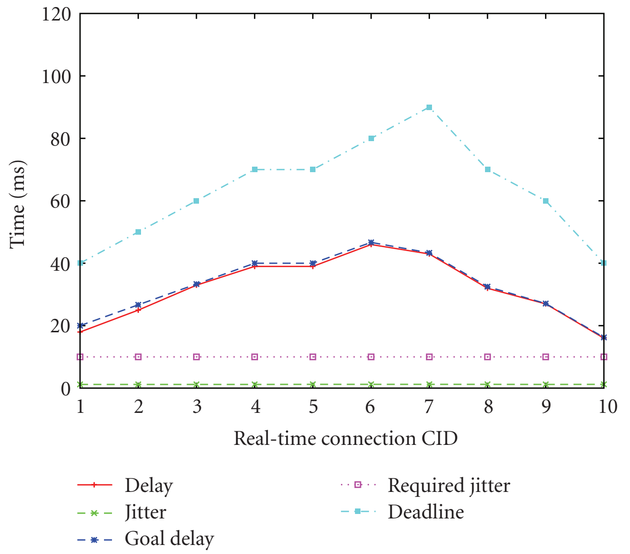


Figure 13: Delay and jitter performances of real-time connections in FQFC.

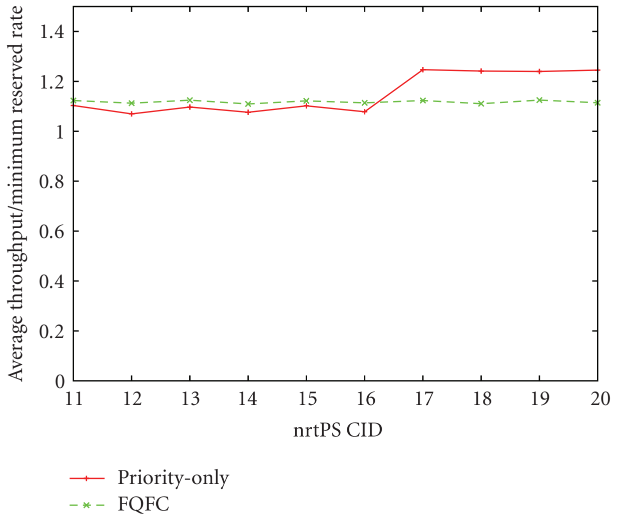


Figure 14: Average throughput/minimum reserved rate of nrtPS connections.

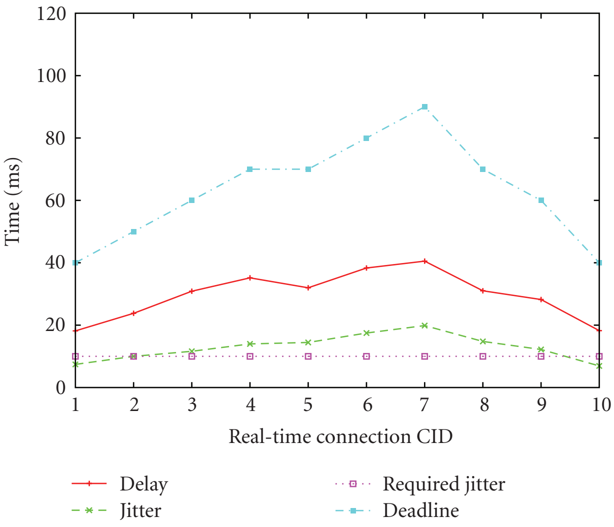


Figure 15: Delay and jitter performances of real-time connections in the priority-only scheduler.

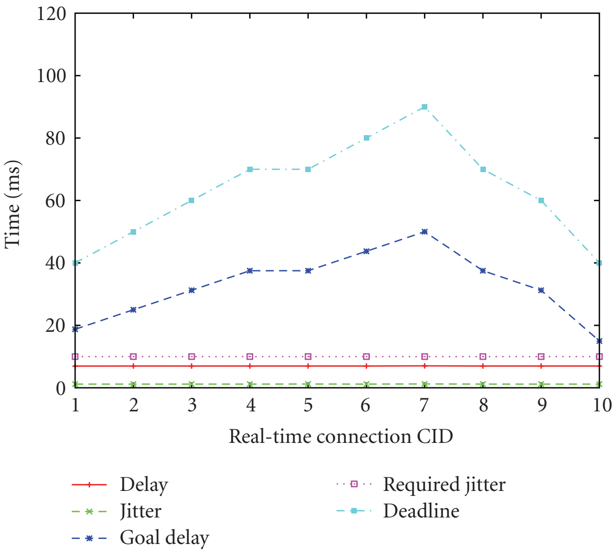


Figure 16: Delay and jitter performances of real-time connections in FQFC.

resources for lower priority classes. This causes the BE connections get more resources. For interclass fairness comparison, the FQFC outperforms priority-only scheduler as shown in Figure 18. Table 3 shows that the interclass fairness index of the FQFC is close to one. Hence, in addition to intraclass fairness, the FQFC also guarantees the interclass fairness. For priority-only scheduler, every real-time connection grabs as many channel resources as possible. Though delays can be lower than the deadlines, delay and jitter differences among connections are not maintained. For nrtPS connections, the differences of throughput ratios are not controlled in priority-only scheduler. The differences of channel resources grasped by the BE connections are, therefore, obvious.

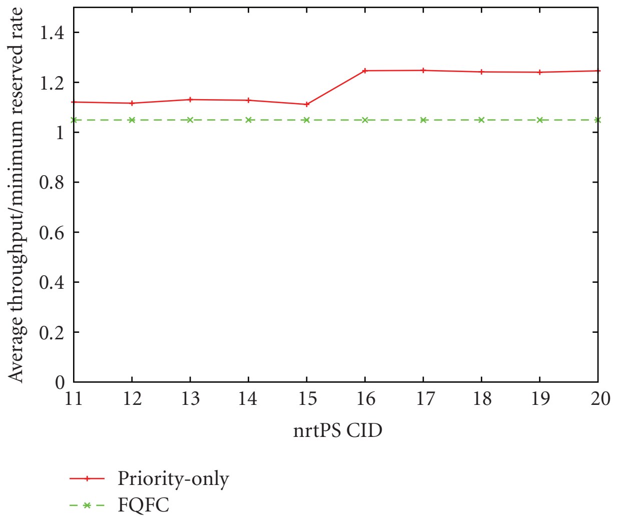


Figure 17: Average throughput/minimum reserved rate of nrtPS connections.

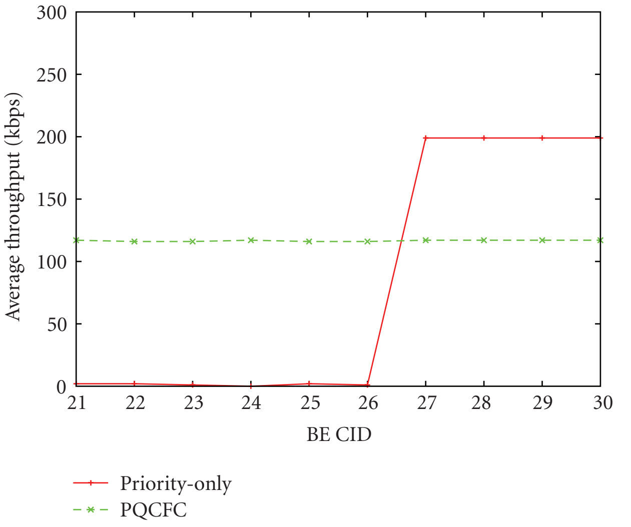


Figure 18: Average throughput of BE connections.

Table 3: Interclass fairness index.

	FQFC	Priority-only
Scenario 3	0.994669	0.677405

4.3.4. Scenario 4: Link Degradation. In this scenario, we evaluate the robustness of the FQFC against wireless link degradation. At the 4.0th second, the wireless link from BS to SS3 degrades, and the PHY layer adaptation mechanism changes the modulation over this link from 64-QAM to QPSK. At the 6.0th second, this link recovers to 64-QAM. Figures 19 and 20, respectively, show the PDU delay of real-time connection 3 and the average throughput of nrtPS connections 13 in SS3, where the link degradation occurs at the 4.0th second. Figure 20 also shows the average throughput of nrtPS connection 13 which is an external connection

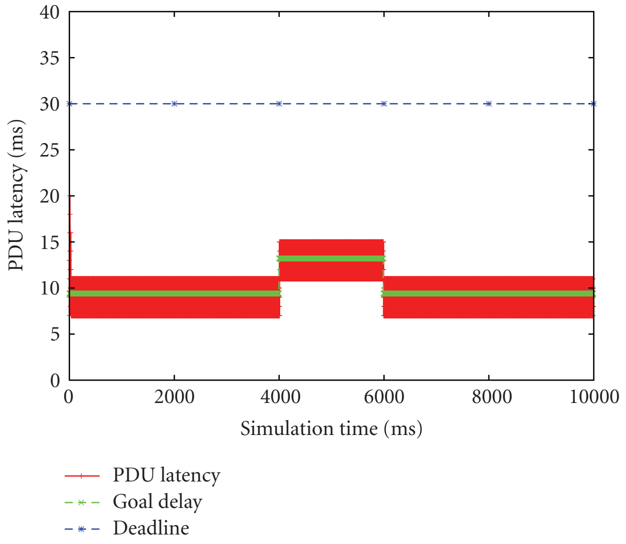


Figure 19: PDU delay of real-time connection 3.

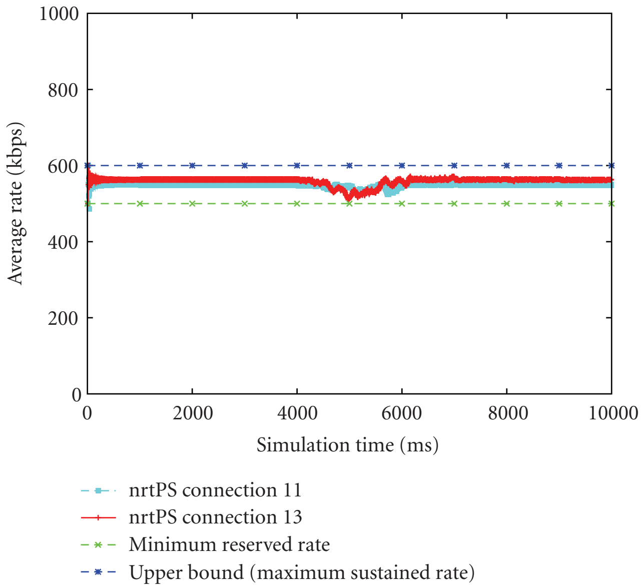


Figure 20: Average throughput of nrtPS connections 11 and 13.

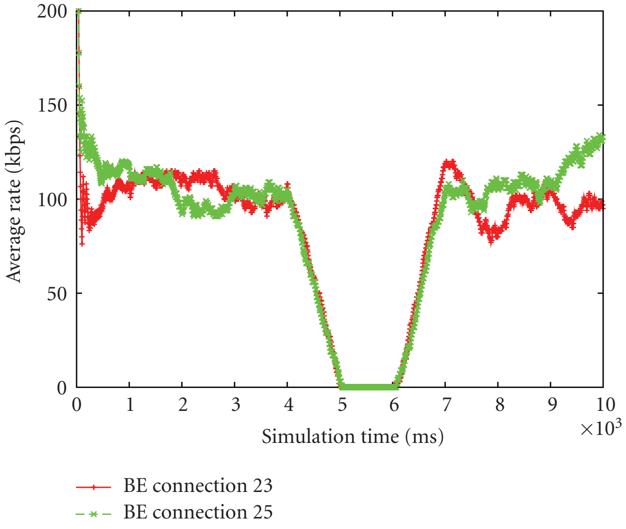


Figure 21: Average throughput of BE connections 23 and 25.

out of SS3. Figure 21 shows the average throughput of BE connections 23 and 25. The simulation shows that

(i) when the link degradation occurs, the FQFC adjusts the goal delay and the tolerable range according to the updated modulation. The FQFC continues to make the delay of real-time connection 3 fall in the tolerable range as shown in Figure 19. Hence, the FQFC can efficiently control the delay according to the goal delay and the tolerable range;

(ii) the service curves of nrtPS connections 11 and 13 in Figure 20 distinguish a throughput drop from the 4.0th second to the 6.0th second, whereas FQFC still maintains the throughput to meet the QoS requirements. The service curves of BE connections 23 and 25 in Figure 21 also distinguish a throughput drop from the 4.0th second to the 6.0th second. The resources are released to guarantee the QoS of real-time connection 7 as shown in Figure 19. For intraclass fairness in nrtPS connections and BE connections, all nrtPS connections keep almost the same resource usage ratio. For interclass fairness, nrtPS and BE connections release resources to guarantee the QoS of real-time connections. Hence, the FQFC guarantees both QoS and fairness even in case that wireless link degrades.

5. Conclusions

A fairness and QoS guaranteed scheduling approach with fuzzy controls FQFC algorithm is proposed for WiMAX OFDMA systems. Different from the utility-fairness, new fairness and QoS evaluation criteria in terms of loss rate, jitter, and throughput are proposed for different classes. The proposed FQFC scheme controls the delay, jitter, and throughput QoS parameters efficiently providing both fairness and QoS guarantees. Rather than using hard computation approaches such as utility-based optimizations, we use fuzzy controller to perform scheduling and resource allocations to resolve mapping among priority, transmission opportunity, and QoS requirements. The proposed FQFC scheme provides both intra- and interclass fairness guarantees in addition to QoS guarantees while implementation is with low complexity.

References

[1] IEEE Std. 802.16-2004 (Revision of IEEE Std. 802.16-2001), “IEEE standard for local and metropolitan area networks—part 16: air interface for fixed broadband wireless access systems,” Revision of IEEE Std. 802.16-2001, October 2004.

[2] IEEE 802.16e-2005, “IEEE Standard for Local and metropolitan area networks—part 16: air interface for fixed and mobile broadband wireless access systems amendment 2: physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum,” February 2006.

[3] H. Fattah and C. Leung, “An overview of scheduling algorithms in wireless multimedia networks,” *IEEE Wireless Communications*, vol. 9, no. 5, pp. 76–83, 2002.

[4] Q. Liu, X. Wang, and G. B. Giannakis, “A cross-layer scheduling algorithm with QoS support in wireless networks,” *IEEE Transactions on Vehicular Technology*, vol. 55, no. 3, pp. 839–847, 2006.

[5] G. Song and Y. Li, “Cross-layer optimization for OFDM wireless networks—part I: theoretical framework,” *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, pp. 614–624, 2005.

[6] M. Ergen, S. Coleri, and P. Varaiya, “QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems,” *IEEE Transactions on Broadcasting*, vol. 49, no. 4, pp. 362–370, 2003.

[7] C.-L. Chen, J.-W. Lee, S.-Y. Chen, and Y.-H. Kuo, “Hierarchical cross-layer fuzzy control for compromise of multiple objectives in wireless mobile networks,” in *Proceedings of the International Conference on Mobile Technology, Applications, and Systems (Mobility ’08)*, pp. 1–7, Yilan, Taiwan, September 2008.

[8] D. Zheng and J. Zhang, “A two-phase utility maximization framework for wireless medium access control,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 12, pp. 4299–4307, 2007.

[9] K.-H. Liu, L. Cai, and X. Shen, “Multiclass utility-based scheduling for UWB networks,” *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1176–1187, 2008.

[10] N. Lu, J. Bigham, and N. Nasser, “An intra-class and inter-class utility-fair bandwidth adaptation algorithm for multi-class traffic in wireless networks,” in *Proceedings of the Asia-Pacific Conference on Communications (APCC ’06)*, pp. 1–5, Busan, Korea, August-September 2006.

[11] N. Lu and J. Bigham, “On utility-fair bandwidth adaptation for multi-class traffic QoS provisioning in wireless networks,” *Computer Networks*, vol. 51, no. 10, pp. 2554–2564, 2007.

[12] C.-L. Chen, “IEEE 802.11e EDCA QoS provisioning with dynamic fuzzy control and cross-layer interface,” in *Proceedings of the 16th International Conference on Computer Communications and Networks (ICCCN ’07)*, pp. 766–771, Honolulu, Hawaii, USA, August 2007.

[13] C.-L. Chen, “Morphisms from IEEE 802.11 DCF specifications to its EDCA QoS practice with cross-layer interface,” in *Proceedings of the 13th International Conference on Parallel and Distributed Systems (ICPADS ’07)*, vol. 2, pp. 1–8, Hsinchu, Taiwan, December 2007.

[14] C. Douligeris and G. Develekos, “Neuro-fuzzy control in ATM networks,” *IEEE Communications Magazine*, vol. 35, no. 5, pp. 154–162, 1997.

[15] I. W. Habib, “Applications of neurocomputing in traffic management of ATM networks,” *Proceedings of the IEEE*, vol. 84, no. 10, pp. 1430–1441, 1996.

[16] K. Yang, J. Zhang, and H.-H. Chen, “A flexible QoS-aware service gateway for heterogeneous wireless networks,” *IEEE Network*, vol. 21, no. 2, pp. 6–12, 2007.

[17] H. B. Kazemian and L. Meng, “Neuro-fuzzy control for MPEG video transmission over bluetooth,” *IEEE Transactions on Systems, Man and Cybernetics, Part C*, vol. 36, no. 6, pp. 761–771, 2006.

[18] Y. Ben-Shimol, I. Kitroser, and Y. Dinitz, “Two-dimensional mapping for wireless OFDMA systems,” *IEEE Transactions on Broadcasting*, vol. 52, no. 3, pp. 388–396, 2006.

[19] C. Wengerter, J. Ohlhorst, and A. G. E. von Elbwart, “Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA,” in *Proceedings of the 61st IEEE Vehicular Technology Conference (VTC ’05)*, vol. 3, pp. 1903–1907, Stockholm, Sweden, May-June 2005.

[20] M. Shreedhar and G. Varghese, “Efficient fair queuing using deficit round-robin,” *IEEE/ACM Transactions on Networking*, vol. 4, no. 3, pp. 375–385, 1996.

[21] R. Jain, D. Chiu, and W. Hawe, “A quantitative measure of fairness and discrimination for resource allocation in shared computer systems,” DEC Research Report TR-301, Digital Equipment, Littleton, Mass, USA, September 1984.

[22] J. Freitag and N. L. S. da Fonseca, “Uplink scheduling with quality of service in IEEE 802.16 networks,” in *Proceedings of the 50th Annual IEEE Global Telecommunications Conference (GLOBECOM ’07)*, pp. 2503–2508, Washington, DC, USA, November 2007.

[23] M.-H. Fong, R. Novak, S. McBeath, and R. Srinivasan, “Improved VoIP capacity in mobile WiMAX systems using persistent resource allocation,” *IEEE Communications Magazine*, vol. 46, no. 10, pp. 50–57, 2008.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 425367, 8 pages
doi:10.1155/2009/425367

Research Article

CDIT-Based Constrained Resource Allocation for Mobile WiMAX Systems

Felix Brah, Jerome Louveaux, and Luc Vandendorpe

Ecole Polytechnique de Louvain, Université Catholique de Louvain, Place du Levant 2, 1348 Louvain-la-Neuve, Belgium

Correspondence should be addressed to Felix Brah, felix.brah@uclouvain.be

Received 1 July 2008; Accepted 31 October 2008

Recommended by Ekram Hossain

Adaptive resource allocation has been shown to provide substantial performance gain in OFDMA-based wireless systems, such as WiMAX, when full channel state information (CSI) is available at the transmitter. However, in some fading environments (e.g., fast fading), there may not be a feedback link sufficiently fast to convey the CSI to the transmitter. In this paper, we consider resource allocation strategies for downlink multiuser mobile WiMAX systems, where the transmitter has only the channel distribution information (CDI), but no knowledge of the instantaneous channel realization. We address the problem of subchannel assignment and power allocation, to maximize the ergodic weighted-sum rate under long-term fairness, minimum data rate requirement and power budget constraints. This problem is formulated as a nonlinear stochastic constrained optimization problem. We provide an analytical solution based on the Lagrange dual decomposition framework. The proposed method has a complexity of $\mathcal{O}(KM)$ for K users and M subchannels. Simulation results are provided to compare the performance of this method with other allocation schemes and to illustrate the tradeoff between maximized weighted-sum rate and the constraints.

1. Introduction

The mobile version of the Worldwide Interoperability for Microwave Access (mobile WiMAX) is one of the solutions in the competition for wireless broadband applications in challenging mobile environments [1, 2]. The mobile WiMAX air interface is based on orthogonal frequency division multiple access (OFDMA) for improved performances in multipath environments. One of the future aspects of OFDMA is the subchannelization which allows to group a total number of subcarriers into subsets of subcarriers called subchannels [3]. The major advantage of subchannelization is the provision of frequency diversity. A byproduct of the subchannelization is that the need for knowledge of radio channel quality is reduced from per-subcarrier to per-subchannel resolution and resources are allocated on per-subchannel basis. There are three types of subchannelizations, namely, adaptive modulation and coding (AMC), partially used subchannelization (PUSC), and fully used subchannelization (FUSC). With AMC, the subchannels are composed of contiguous groups of subcarriers. With both PUSC and FUSC, the subchannels are composed of distributed subcarriers. For PUSC, the set of used subcarriers, that is, data and pilots, is first partitioned into subchannels, and then pilot subcarriers are allocated within each subchannel. For FUSC, the pilot tones are common for all subchannels and are allocated first and then the remaining subcarriers are divided into data subchannels. In general, AMC is well suited for stationary, portable, and low mobility applications, whereas PUSC and FUSC are the best alternatives for mobile applications. We employ FUSC in this work. This method uses all the subchannels and employs full-channel diversity by distributing the allocated subcarriers to subchannels using a permutation mechanism. Thanks to the frequency diversity provided by the FUSC, the performance degradation due to fast fading in mobile environments is minimized.

Mobile WiMAX aimed at delivering broadband mobile services ranging from real-time interactive gaming, VoIP, and streaming media to nonreal-time web browsing and simple file transfers. Users have channels of different quality. With classical best effort transmission, unfair resource allocation can lead to starvation of some users in bad channel conditions. Therefore, the achievement of fairness among users while satisfying users' minimum rate requirements is an important issue.

Most of the previous works on OFDMA resource allocation have considered only the case where instantaneous channel state (CSI) is available at the transmitter and various algorithms based on instantaneous CSI have been developed [4–14]. In [4], adaptive subcarriers assignment to minimize the total transmit power is investigated. The authors presented a heuristic algorithm, the so-called Hungarian algorithm, based on constructive assignment and iterative improvement. Following the Hungarian approach, [5] proposed an iterative algorithm for power minimization and bit loading. The algorithm is considered as suboptimal for adaptive modulation. To reduce the computational complexity, [6] proposed low complexity and computationally efficient bandwidth and power allocation algorithms to solve the problem of minimizing the total power consumption under bit error rate and transmission rate constraints. In [7], the performance of bandwidth-constrained power minimization and power minimization schemes in terms of outage probability and packet error rate under user data rate satisfaction are compared. It is shown that, in severe shadowing environment with both frequency selective and flat fading, the former scheme outperforms the later. Fairness issues in a wireline multiaccess channel have been taken into account in [8, 9]. The authors introduce the concept of balanced capacity to characterize the multiuser channel performance with total power constraints in [8] and they extend the concept to individual power constraints in [9]. This concept of balanced capacity is closely related to the one presented in [10] where a low complexity suboptimal algorithm that maximizes the sum capacity while maintaining proportional fairness among the users data rate is developed. In [11], suboptimal resource grids and power allocation algorithms to maximize the total throughput under user's data rate requirement are presented. Rate-power allocation algorithms for expected mutual information maximization based on partial channel knowledge have been developed in [13]. In [14], the authors investigated the impact of imperfect channel information on OFDMA-based systems under fairness and minimum rate constraints. Instantaneous resource allocation strategies are suitable for quasistatic or slow fading environments. However, when the channel variations are fast, the transmitter may not be able to adapt to the instantaneous channel realization. Hence, CSI-aware resource allocation is not suitable for environments with high mobility.

When the channel state can be accurately tracked at the receiver, the statistical channel model at the transmitter can be based on channel distribution information feedback from the receiver. We refer to knowledge of the channel distribution at the transmitter as CDIT. Power allocation for ergodic capacity maximization in relay networks based on CDIT under high SNR regime has been studied in [15].

This paper addresses CDIT-based resource allocation strategies for mobile WiMAX in all SNR regimes. The goal is to adaptively assign subchannels and distribute the total power to users with the objective to maximize the ergodic weighted-sum rate under tunable long-term fairness, minimum data rate requirements, and a total power constraint. This constrained optimization problem is formulated as an infinite dimensional stochastic problem. To efficiently solve the problem, we propose an analytical method based on the Lagrange dual decomposition framework. The remainder of this paper is organized as follows. In Section 2, the system model considered is described and the ergodic weighted-sum rate is derived. The problem of multiuser resource allocation based on CDIT is formulated in Section 3 and a solution guideline is given. In Section 4, some simulation results are presented. Finally, conclusions are drawn in Section 5.

2. System Model

Throughout the paper, we consider a single cell downlink WiMAX communication from a base station (BS) to K mobile user terminals, over a realistic frequency-selective fast fading channel with total bandwidth B. The BS splits up the downlink bandwith into different subchannels. Then the data to be transmitted to different mobile user terminals are amalgamated using downlink FUSC. After the downlink subchannelization, the resulting frequency domain OFDMA symbols are converted into time domain OFDMA symbols using inverse FFT. Then a cyclic prefix is added to each symbol to provide immunity against multipath propagation. Finally the signal undergoes frequency upconversion before it is transmitted from the base station to the user terminals. We assume that the user terminal has perfect CSI to perform coherent detection, but there is no fast feedback link to perfect the CSI to the base station. Hence, the base station has only channel distribution information (CDI), but no knowledge of the instantaneous channel realizations. Assuming that the receiver employs a maximum ratio combiner (MRC), the effective SNR of user k at mth subchannel is given by

$$\gamma_{k,m} = \frac{1}{\Gamma_k N \sigma_n^2} \sum_{n=0}^{N-1} g_{k,m}(n). \tag{1}$$

In (1), N is the number of distributed subcarriers per-subchannel, $g_{k,m}(n)$ is the channel gain of user k at subcarrier n of mth subchannel, which is the product of the distance attenuation and the fast fading gain, σ_n^2 is the noise variance, Γ_k is referred to as SNR gap related to the required bit error rate of user k (BER_k) and is approximated as $\Gamma_k \cong -\ln(5\,\mathrm{BER}_k/1.5)$ for QAM modulations [10]. We assume a Rayleigh channel model. Hence $\gamma_{k,m}$ is a *central chi-squared* (χ^2) distributed random variable with two degrees of freedom and with probability density function (pdf)

$$T_{\gamma_{k,m}}(\gamma_{k,m}) = \frac{1}{\overline{\gamma}_{k,m}} e^{-\gamma_{k,m}/\overline{\gamma}_{k,m}}, \tag{2}$$

where $\overline{\gamma}_{k,m}$ is the mean of the $\gamma_{k,m}$ distribution.

Each user is adaptively assigned a number of different subchannels to send and receive data. An indicator $\rho_{k,m}$ is used to represent whether the mth subchannel is assigned to user k. Note that in a single cell OFDMA system, each subchannel can be assigned to at most one user at any time, that is, $\sum_{k=1}^{K}\rho_{k,m} \in \{0,1\}$ for all m. Due to the consideration for the reduction of the signaling overhead in WIMAX, the power is equally distributed across subcarriers within each subchannel. We assume the duration of the transmission codewords is long enough to undergo all channel realizations. We further assume perfect CDIT, thereby allowing to take the expectation over the distribution. The ergodic weighted-sum rate of the multiuser system is defined as

$$U_\gamma = E_\gamma\left\{\sum_{k=1}^{K}\frac{1}{\overline{R}_k^{\alpha_f}}\sum_{m=1}^{M}\rho_{k,m}\log_2(1+p_{k,m}\gamma_{k,m})\right\}, \tag{3}$$

where $\gamma = [\gamma_1^T,\ldots,\gamma_K^T]^T$ with $\gamma_k = [\gamma_{k,1},\ldots,\gamma_{k,M}]$, $p_{k,m}$ denotes the power allocated to the user k on subchannel m, $E_\gamma\{\cdot\}$ represents the statistical expectation with respect to γ, $\overline{R}_k$ is user k's average data rate so far at the allocation time, and α_f is a tunable fairness parameter. Setting α_f to 1 results in the *proportional fair* allocation. For $\alpha_f = 0$, this results in *maximum throughput* allocation. The average user rates $\overline{R}_k$ are updated according to

$$\overline{R}_k(t+1) = \left(1-\frac{1}{\tau_c}\right)\overline{R}_k(t) + \frac{1}{\tau_c}r_k(t), \tag{4}$$

where $r_k(t)$ is the rate allocated to user k at time t and τ_c is the parameter that controls the latency of the system. This way, we consider both current rate as well as rates given to the users in the past, what is suitable for long-term fairness evaluation.

3. CDIT-Based Constrained Resource Allocation

3.1. Formulation of the Problem. The issue is how to adaptively assign the M subchannels to the K users and distribute the total power budget P_{tot} in order to maximize the ergodic weighted-sum rate (3) while satisfying user's minimum rate and system fairness requirements under a total power constraint. Mathematically, this constrained optimization problem is formulated as

$$f^* = \max_{\rho_{k,m},p_{k,m}} E_\gamma\left\{\sum_{k=1}^{K}\frac{1}{\overline{R}_k^{\alpha_f}}\sum_{m=1}^{M}\rho_{k,m}\log_2(1+p_{k,m}\gamma_{k,m})\right\},$$

subject to

$$E_\gamma\left\{\sum_{m=1}^{M}\rho_{k,m}\log_2(1+p_{k,m}\gamma_{k,m})\right\} \geq R_k, \tag{5}$$

$$E_\gamma\left\{\sum_{k=1}^{K}\sum_{m=1}^{M}\rho_{k,m}p_{k,m}\right\} \leq P_{\text{tot}}.$$

The first constraint in (5) is for the user's specific minimum data rate demand. We assume that appropriate admission control is performed such that the minimum data rates R_k are feasible. The second constraint is system limitation on transmits powers.

Note that the optimization problem (5) involves both continuous variables $p_{k,m}$ and boolean variable $\rho_{k,m}$. Such an optimization problem is neither convex nor concave with respect to $(\rho_{k,m}, p_{k,m})$.

3.2. Solution Based on Lagrange Dual Decomposition. We can solve problem (5) using the Lagrange dual decomposition framework. Following the approach in [12], we relax $\rho_{k,m}$ to be $\rho_{k,m} \in [0,1]$. Then $\rho_{k,m}$ can be regarded as time-sharing factor. Thanks to the linearity property of the expectation, the Lagrangian function of the primal problem (5) can be expressed as

$$\begin{aligned} L_\gamma(p_{k,m},\rho_{k,m},\lambda_k,\mu) &= \sum_{k=1}^{K}\sum_{m=1}^{M}E_{\gamma_{k,m}}\left\{\frac{\rho_{k,m}}{\overline{R}_k^{\alpha_f}}\log_2(1+p_{k,m}\gamma_{k,m})\right\} \\ &\quad + \sum_{k=1}^{K}\lambda_k\sum_{m=1}^{M}E_{\gamma_{k,m}}\{\rho_{k,m}\log_2(1+p_{k,m}\gamma_{k,m})\} \\ &\quad - \mu\sum_{k=1}^{K}\sum_{m=1}^{M}\rho_{k,m}p_{k,m} - \sum_{k=1}^{K}\lambda_k R_k + \mu P_{\text{tot}}, \end{aligned} \tag{6}$$

where λ_k and μ are Lagrangian multipliers. Let $p_{k,m}^*$ and $\rho_{k,m}^*$ denote the optimal solution of (6). We first investigate the problem for fixed values of λ_k and μ. By Karush-Kuhn-Tucker (KKT) first optimality condition [16], $\rho_{k,m}^*$ and $p_{k,m}^*$ should satisfy the following:

$$\left.\frac{\partial L_\gamma}{\partial p_{k,m}}\right|_{p_{k,m}=p_{k,m}^*} \begin{cases} <0, & p_{k,m}^* = 0, \\ =0, & p_{k,m}^* > 0, \end{cases} \tag{7}$$

$$\left.\frac{\partial L_\gamma}{\partial \rho_{k,m}}\right|_{\rho_{k,m}=\rho_{k,m}^*} \begin{cases} <0, & \rho_{k,m}^* = 0, \\ =0, & 0<\rho_{k,m}^* < 1, \\ >0, & \rho_{k,m}^* = 1. \end{cases} \tag{8}$$

For a nonzero power allocation and $\rho_{k,m}^* \in (0,1)$, we obtain from (7) and (8)

$$\rho_{k,m}^* E_{\gamma_{k,m}}\left\{\frac{\overline{R}_k^{(-\alpha_f)}+\lambda_k}{\ln 2}\cdot\frac{\gamma_{k,m}}{1+p_{k,m}^*\gamma_{k,m}} - \mu\right\} = 0, \tag{9}$$

$$E_{\gamma_{k,m}}\left\{\frac{\overline{R}_k^{(-\alpha_f)}+\lambda_k}{\ln 2}\cdot\ln(1+p_{k,m}^*\gamma_{k,m}) - \mu p_{k,m}^*\right\} = 0. \tag{10}$$

We deduce from (9) that $p_{k,m}^*$ has to satisfy the following condition:

$$\rho_{k,m}^* E_{\gamma_{k,m}}\left\{\frac{\gamma_{k,m}}{1+p_{k,m}^*\gamma_{k,m}} - \frac{\mu\ln 2}{A_k}\right\} = 0, \tag{11}$$

where $A_k = \overline{R}_k^{(-\alpha_f)} + \lambda_k$.

When $\rho_{k,m}^* = 0$, the value of $p_{k,m}$ is undefined, and any value can be taken without any influence on the objective function or on the constraints. On the other hand, for any other positive value, $\rho_{k,m}$ vanishes out of the expression and we get

$$E_{\gamma_{k,m}}\left\{\frac{\gamma_{k,m}}{1+p_{k,m}^{*}\gamma_{k,m}}-\frac{\mu\ln 2}{A_k}\right\}=0. \tag{12}$$

We can use the pdf of the SNR distribution (2) to transform (12) into

$$\int_0^{\infty}\left(\frac{\gamma_{k,m}}{1+p_{k,m}^{*}\gamma_{k,m}}-\frac{\mu\ln 2}{A_k}\right)\cdot\frac{1}{\overline{\gamma}_{k,m}}e^{-\gamma_{k,m}/\overline{\gamma}_{k,m}}d\gamma_{k,m}=0, \tag{13}$$

which is equivalent to

$$\begin{aligned}&\frac{\mu\ln 2}{A_k}p_{k,m}^{*2}\overline{\gamma}_{k,m}-p_{k,m}^{*}\overline{\gamma}_{k,m}\\&\qquad+\exp\left(\frac{1}{p_{k,m}^{*}\overline{\gamma}_{k,m}}\right)E_1\left(\frac{1}{p_{k,m}^{*}\overline{\gamma}_{k,m}}\right)=0,\end{aligned} \tag{14}$$

where

$$E_1(x)=\int_1^{\infty}\frac{e^{-tx}}{t}dt \tag{15}$$

is the exponential integral function of x [17].

Equation (10) is equivalent to

$$E_{\gamma_{k,m}}\left\{\frac{A_k}{\mu\ln 2}\ln\left(1+p_{k,m}^{*}\gamma_{k,m}\right)-p_{k,m}^{*}\right\}=0. \tag{16}$$

Using the pdf (2), (16) can be transformed into

$$\begin{aligned}&\int_0^{\infty}\left(\frac{A_k}{\mu\ln 2}\ln\left(1+p_{k,m}^{*}\gamma_{k,m}\right)-p_{k,m}^{*}\right)\\&\qquad\cdot\frac{1}{\overline{\gamma}_{k,m}}e^{-\gamma_{k,m}/\overline{\gamma}_{k,m}}d\gamma_{k,m}=0,\end{aligned} \tag{17}$$

which is finally equivalent to

$$\frac{A_k}{\mu\ln 2}\exp\left(\frac{1}{p_{k,m}^{*}\overline{\gamma}_{k,m}}\right)E_1\left(\frac{1}{p_{k,m}^{*}\overline{\gamma}_{k,m}}\right)-p_{k,m}^{*}=0. \tag{18}$$

From (14) and (18), we derive

$$p_{k,m}^{*}=\left[\frac{A_k}{\mu\ln 2}-\frac{1}{\overline{\gamma}_{k,m}}\right]^{+}, \tag{19}$$

where $[x]^{+}=\max(0,x)$. The expression (19) is in the form of *multilevel water-filling* power allocation with cut-off subchannel SNR $(\mu\ln 2)/A_k$ below which we do not transmit any power, and above which we transmit more power when $\overline{\gamma}_{k,m}$ is higher. The important difference is that, in contrast to the CSIT-based allocation where the $p_{k,m}^{*}$depends on the instantaneous channels realizations, the optimal allocation here is dependent on the mean of the channel distribution, and thus needs to be computed only when the statistics of the channel has changed.

We have from (8) and (18) that

$$\rho_{k,m}^{*}=\begin{cases}1, & \text{if } G_{k,m}\left(p_{k,m}^{*}\right)>0,\\ \in(0,1), & \text{if } G_{k,m}\left(p_{k,m}^{*}\right)=0,\\ 0, & \text{if } G_{k,m}\left(p_{k,m}^{*}\right)<0,\end{cases} \tag{20}$$

where

$$G_{k,m}\left(p_{k,m}^{*}\right)=\frac{A_k}{\mu\ln 2}\exp\left(\frac{1}{p_{k,m}^{*}\overline{\gamma}_{k,m}}\right)E_1\left(\frac{1}{p_{k,m}^{*}\overline{\gamma}_{k,m}}\right)-p_{k,m}^{*}. \tag{21}$$

Due to the exclusive subchannel assignment constraint in OFDMA, we can conclude that for each subchannel m, if $G_{k,m}$ are all different, then only the user with the largest $G_{k,m}$ can use that subchannel. In other words,

$$\rho_{k_m^{*},m}^{*}=1,\quad \rho_{k,m}^{*}=0,\quad \forall k\neq k_m^{*}, \tag{22}$$

where

$$k_m^{*}=\arg\max_k G_{k,m}\left(p_{k,m}^{*}\right). \tag{23}$$

Substituting (19) into the Lagrange function (6) and thanks to the exclusive subchannel assignment constraint, we obtain the following per-subchannel dual problem:

$$g^{*}=\min_{\lambda_k,\mu}L_{\gamma}^{*}(\lambda_k,\mu), \tag{24}$$

where $L_{\gamma}^{*}(\lambda_k,\mu)$ is the dual function given by

$$\begin{aligned}L_{\gamma}^{*}(\lambda_k,\mu)={}&E_{\gamma_{k,m}}\left\{\overline{R}_k^{(-\alpha_f)}\log_2\left(\frac{\overline{R}_k^{(-\alpha_f)}+\lambda_k}{\mu\ln 2}\overline{\gamma}_{k,m}\right)\right\}\\&+\lambda_k E_{\gamma_{k,m}}\left\{\log_2\left(\frac{\overline{R}_k^{(-\alpha_f)}+\lambda_k}{\mu\ln 2}\overline{\gamma}_{k,m}\right)\right\}\\&-\lambda_k R_k-\mu\left(\frac{\overline{R}_k^{(-\alpha_f)}+\lambda_k}{\mu\ln 2}-\frac{1}{\overline{\gamma}_{k,m}}\right)+\mu P_{\text{tot}}.\end{aligned} \tag{25}$$

Next we turn to the optimization of the dual function (25) over μ and λ_k. First we consider the optimization over λ_k for μ fixed to find. We differentiate (25) with respect to λ_k and set the derivative to 0 to obtain

$$\left.\frac{\partial L_{\gamma}^{*}}{\partial\lambda_k}\right|_{\lambda_k=\lambda_k^{*}}=\log_2\left(\frac{\overline{R}_k^{(-\alpha_f)}+\lambda_k}{\mu\ln 2}\overline{\gamma}_{k,m}\right)-R_k=0. \tag{26}$$

The optimum λ_k^{*} is derived from (26) as follows:

$$\lambda_k^{*}(\mu)=2^{R_k}\frac{\mu\ln 2}{\overline{\gamma}_{k,m}}-\frac{1}{\overline{R}_k^{(\alpha_f)}}. \tag{27}$$

If some of the individual rate constraints are exceeded, the corresponding λ_k is equal to zero.

Substituting (27) into (25) we obtain

$$L_\gamma^*(\mu) = \max_{\lambda_k} L_\gamma^*(\lambda_k, \mu) = \frac{R_k}{\overline{R}_k^{(\alpha_f)}} - \frac{2^{R_k}\mu}{\overline{\gamma}_{k,m}} + \frac{\mu}{\overline{\gamma}_{k,m}} + \mu P_{\text{tot}}. \tag{28}$$

We next consider the optimization of $L_\gamma^*(\mu)$ over μ. The function $L_\gamma^*(\mu)$ can be shown to be a convex function of μ, which can then be minimized via a one-dimensional search with geometric convergence. The optimal values μ^* correspond to the ones that satisfy the total power constraint (with equality).

We can conclude that, if $G_{k,m}$ are all different, then a given subchannel m is exclusively assigned to the user $\widetilde{k}_m^*$ satisfying

$$\widetilde{k}_m^* = \arg\max_k G_{k,m}(\widetilde{p}_{k,m}^*), \tag{29}$$

where $\widetilde{p}_{k,m}^*$ is the optimal power allocation given by

$$\begin{aligned}\widetilde{p}_{k,m}^* &= \left[\frac{\overline{R}_k^{(-\alpha_f)} + \lambda_k^*}{\mu^* \ln 2} - \frac{1}{\overline{\gamma}_{k,m}}\right]^+, \quad \text{if } k = \widetilde{k}_m^*,\\ &= 0, \quad \text{if } k \neq \widetilde{k}_m^*.\end{aligned} \tag{30}$$

3.3. Relative Duality Gap. The relative duality (optimality) gap which indicates how far we are from the optimal value can be expressed as

$$d^* = \frac{g^* - f^*}{f^*} \geq 0, \tag{31}$$

where $f^* > 0$ and $g^* > 0$ given in (5) and (24) are the optimal values of the primal and dual problems. The inequality follows from the positivity of f^* and the weak duality theorem [18].

Without the minimum rate constraints in (5), problem (5) becomes a standard convex optimization problem, and then the duality gap is zero. Due to the nonlinearity of the minimum rate constraints, the convexity of problem (5) does not hold. However, the nonconvex optimization problem (5) for the investigated OFDMA-based WIMAX system fulfills the *time-sharing* condition as defined in [19]. Then when the power constraint is met tightly, that is, with equality, the duality gap is zero, and thus solving the dual problem (24) also solves the primal problem (5).

3.4. Instantaneous Resource Allocation Based on CSIT. In order to assert the relevance of our approach, it was decided to compare it to the instantaneous allocation based on partial CSIT and to the instantaneous allocation based on perfect CSIT.

3.4.1. Resource Allocation Based on Partial CSIT. Assuming that partial channel state information is available at the transmitter in the form of an estimate of the SNR, it has been shown that resources can be optimally allocated based on this partial CSIT (see, e.g., [13, 14]). Let $\gamma_{k,m}$ and $\widehat{\gamma}_{k,m}$ denote the real and the estimated subchannel SNR. For Rayleigh fading channels, $\gamma_{k,m}$ conditioned on $\widehat{\gamma}_{k,m}$ is a *noncentral chi-squared* distributed random variable with two degrees of freedom [13, 14]. Its probability density function (pdf) can be approximated to a *Gamma* function as

$$T_{\gamma_{k,m}}(\gamma_{k,m} \mid \widehat{\gamma}_{k,m}) \approx \frac{\beta^\alpha}{\Gamma(\alpha)} \gamma_{k,m}^{(\alpha-1)} e^{-\beta\gamma_{k,m}}. \tag{32}$$

In expression (32), $\alpha = (\widehat{\gamma}_{k,m}\gamma_{e/n}^{-1} + 1)^2/(2\widehat{\gamma}_{k,m}\gamma_{e/n}^{-1}+1)$ and $\beta = \alpha/(\widehat{\gamma}_{k,m}+\gamma_{e/n})$ are the shape parameter and the rate parameter of the *Gamma* pdf, respectively, where $\gamma_{e/n}$ is the ratio of the subchannel estimation error variance to the background noise variance. $\Gamma(x)$ is the *Gamma* function of x.

Under the partial CSIT assumption, the optimization goal is to maximize the expected weighted-sum rate instead of the ergodic weighted-sum rate. In [14], the problem has been formulated as

$$\max_{p_{k,m},\rho_{k,m}} E_\gamma\left\{\sum_{k=1}^{K}\sum_{m=1}^{M}\frac{\rho_{k,m}}{\overline{R}_k^{\alpha_f}}\log_2(1 + p_{k,m}\gamma_{k,m}) \mid \widehat{\gamma}_{k,m}\right\}, \tag{33}$$

subject to

$$\begin{gathered}E_\gamma\left\{\sum_{m=1}^{M}\rho_{k,m}\log_2(1 + p_{k,m}\gamma_{k,m}) \mid \widehat{\gamma}_{k,m}\right\} \geq R_k,\\ \sum_{k=1}^{K}\sum_{m=1}^{M} p_{k,m} \leq P_{\text{tot}}.\end{gathered} \tag{34}$$

Using the pdf (32) and applying the KKT optimality conditions, it has been shown in [14] that the optimal power allocation $p_{k,m}^*$ is the solution of

$$\frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \frac{\gamma_{k,m}^\alpha}{1 + p_{k,m}^*\gamma_{k,m}} e^{-\beta\gamma_{k,m}} d\gamma_{k,m} - \frac{\mu \ln 2}{A_k} = 0. \tag{35}$$

Also by KKT optimality conditions, it has been shown in [14] that a given subchannel m is exclusively assigned to the user k_m^* satisfying

$$k_m^* = \arg\max_k A_k G_k(p_{k,m}^*), \tag{36}$$

where

$$\begin{aligned}G_k(p_{k,m}) = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty &\gamma_{k,m}^{(\alpha-1)}\log_2(1 + p_{k,m}\gamma_{k,m})\\ &\times e^{-\beta\gamma_{k,m}} d\gamma_{k,m} - \mu p_{k,m}.\end{aligned} \tag{37}$$

3.4.2. Resource Allocation Based on Perfect CSIT. Under the *unrealistic* perfect CSIT assumption, instead of maximizing the ergodic or the expected weighted-sum rate, the optimization goal is to maximize the instantaneous weighted-sum rate

$$\max_{p_{k,m},\rho_{k,m}} \sum_{k=1}^{K}\frac{1}{\overline{R}_k^{\alpha_f}}\sum_{m=1}^{M}\rho_{k,m}\log_2(1 + p_{k,m}\gamma_{k,m}), \tag{38}$$

subject to

$$\begin{gathered}\sum_{m=1}^{M}\rho_{k,m}\log_2(1+p_{k,m}\gamma_{k,m}) \geq R_k,\\ \sum_{k=1}^{K}\sum_{m=1}^{M} p_{k,m} \leq P_{\text{tot}}.\end{gathered} \tag{39}$$

From the KKT optimality conditions, the optimal power allocation, solution of (39) is given by

$$\tilde{p}^*_{k,m} = \left[\frac{A_k}{\mu\ln 2} - \frac{1}{\gamma_{k,m}}\right]^+. \tag{40}$$

This is a *multilevel water-filling* power allocation with cut-off subchannel SNR $(\mu\ln 2)/A_k$. The difference between (40) and (19) is that the power allocation in (40) depends on the instantaneous subchannel SNR $\gamma_{k,m}$ while the one in (19) depends on the mean of the SNR distribution $\overline{\gamma}_{k,m}$.

We also deduce from KKT optimality conditions [14] that a given subchannel m is exclusively assigned to the user $\tilde{k}^*_m$ ($\rho_{\tilde{k}^*_m,m} = 1$, $\rho_{k,m} = 0$ for $k \neq \tilde{k}^*_m$) satisfying

$$\tilde{k}^*_m = \max_k A_k(\log_2(1+p^*_{k,m}\hat{\gamma}_{k,m}) - \mu p^*_{k,m}). \tag{41}$$

3.5. Feedback Reduction and Complexity Analysis. First, thanks to the subchannelization, the need for knowledge of radio channel quality in mobile WiMAX is reduced from per-subcarrier to per-subchannel resolution and resources are allocated on per-subchannel basis. Second, under CDIT-based allocation, instead of feeding back the instantaneous channel coefficients to the transmitter, the users simply feed back the mean of the subchannel SNR distribution. Putting these two facts together, the amount of feedback required for the resource allocation reduces significantly.

Using a dual decomposition framework, the optimization problem has been reduced to a per-subchannel optimization, and the computational complexity has been significantly decreased.

Since the optimal λ^*_k and μ^* depend on the mean of the subchannel SNR distribution and not on the instantaneous values, they need to be computed only when the statistic of the channel has changed. We need to run the line search to compute for μ^*. This is followed by the computation of KM values of multipliers λ^*_k (27) and power allocation values $\tilde{p}^*_{k,m}$ (30). We assume the line search to require I_μ iterations. The computation of λ^*_k and $\tilde{p}^*_{k,m}$ requires $\mathcal{O}(KM)$ operations. The overall complexity order for the CDIT-based resource allocation is thus $\mathcal{O}(KMI_\mu)$. Since I_μ is just constant independent of K and M, the complexity is $\mathcal{O}(KM)$. Once μ^*, λ^*_k, and $\tilde{p}^*_{k,m}$ have been determined, we do not need to update them as long as the statistics of the fading channel remains the same.

Both expressions (19) and (40) are in the form of *multilevel water-filling* power allocation with cut-off subchannel SNR $(\mu\ln 2)/A_k$. Thus, the complexity in term of water filling is the same. The main difference between (19) and (40) is the amount of feedback required to perform resource allocation. Recall that, for the CSIT-based scheme, the allocation is performed after each OFDM symbol period. Let N_s be the number of OFDM symbol periods after which the CDIT-based resource allocation is performed. Then a rough estimation tells us that the complexity of the CDIT-based allocation is reduced by $1/N_s$ compared to the perfect CSIT scheme.

3.6. Tradeoff Analysis. In the tradeoff analysis, we vary the constraints, and see the effect on the maximized weighted-sum rate. We define a relaxation coefficient η_q for the minimum rate constraint and replace the minimum rate requirement R_k by $(1-\eta_q)R_k$ to form a perturbed problem. When $\eta_q = 0$, this reduces to the original problem (5). By increasing η_q, the minimum rate constraints are relaxed. We can also vary the fairness parameter α_f. Setting α_f to 0 results in the *maximum throughput* allocation. For $\alpha_f = 1$, this results in the *proportional fair* allocation. The relaxation of the constraints leads in general to an improvement of the optimal objective. The tradeoff curves are found by solving the perturbed problem for many values of η_q and α_f.

4. Simulation Results

To illustrate the performance of the proposed resource allocation method, we perform simulations for a three-users mobile WiMAX system with bandwidth divided into $M = 8$ subchannels. The subchannels are formed using the FUSC which is suitable for mobile applications. The FFT size of the OFDMA is 512 points. The performance is evaluated in multipath channel environments modeled as a tapped delay line with six taps as specified in the ITU M.1225 Vehicular A channel model [20]. We consider a scenario where user 2 is every time closer to the base station than users 1 and 3 and the relative mean SNR difference between user 2 and users 1 and 3 is -5 dB and -3 dB, respectively, while the minimum data rate demand of user 3 is higher than the one of users 2 and 1 ($R_3 > R_2 > R_1$). The target bit error rate is set to 10^{-3} (without channel coding). The performances are evaluated using simulations over 10 000 instances of independent channel realizations. For all the simulations, the total power is set to $P_{\text{tot}} = KM$.

In Figure 1, the performance of the proposed adaptive resource allocation is compared to those of optimal resource allocation based on perfect CSIT, resource allocation based on partial CSIT and a uniform power allocation. The result shows that the proposed adaptive resource allocation brings significant gain over resource allocation based on partial CSIT with higher estimation error. We can observe that when $\gamma_{e/n}$ is small, that means the effect of the estimation error is less dominant than the one of the background noise, the optimization under partial CSI is closed to the one under perfect CDIT. For very low estimation errors, the partial CSIT-based scheme outperforms the perfect CDIT scheme. The weighted-sum rate degrades quickly as the estimation error grows, especially for high SNRs. The highest weighted-sum rate is obtained with perfect CSIT but the

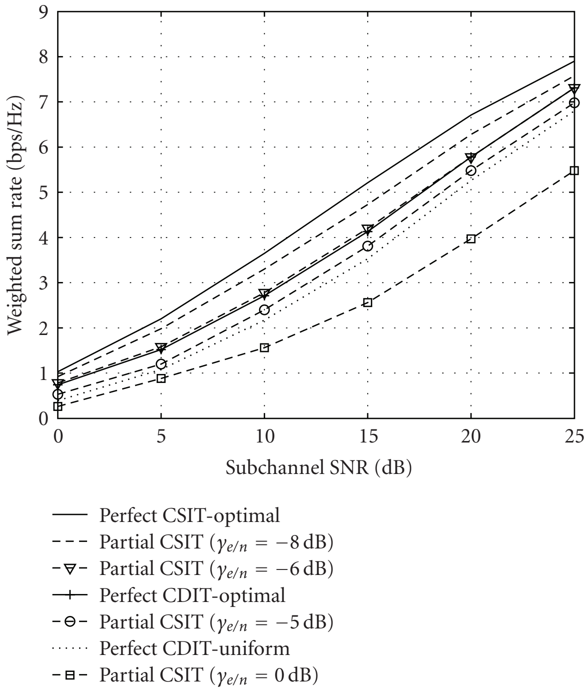


FIGURE 1: Maximized weighted-sum rate versus mean SNR for different resource allocation schemes and fairness parameter $\alpha_f = 0$.

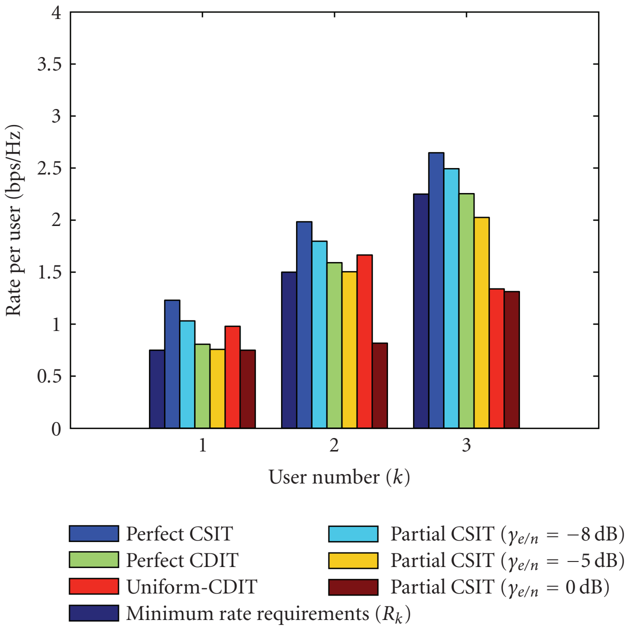


FIGURE 2: Users rates for different resource allocation schemes, mean SNR of 15 dB and fairness parameter $\alpha_f = 0$.

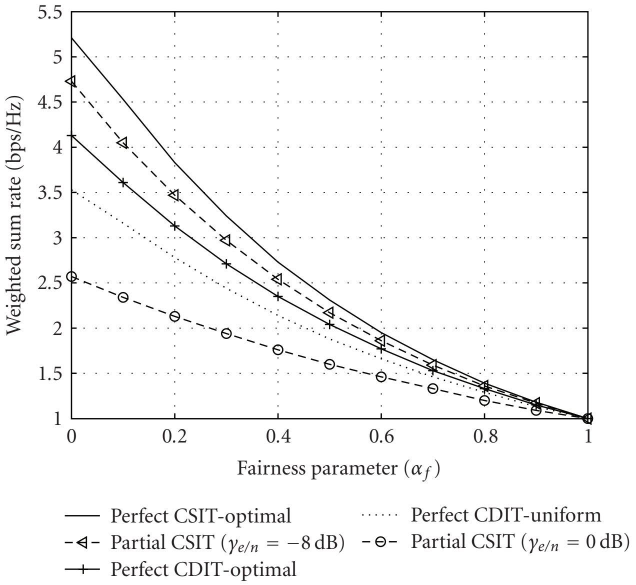


FIGURE 3: Tradeoff between maximized weighted-sum rate and fairness requirement for different resource allocation schemes and mean SNR of 15 dB.

difference in terms of performance is not so significant compared to the difference of complexity between CDIT-based and CSIT-based allocation schemes. The proposed method outperforms the uniform power allocation.

Figure 2 shows the user's rate for different allocation schemes when the users minimum data rate demands are constrained to $R_3 = 2R_2 = 3R_1$ and the fairness parameter α_f is set to 0. We observe that under optimal allocation based on perfect CDIT, the need of all users in terms of data rate is satisfied. This is neither the case under allocation based on partial CSIT with high estimation error nor under uniform allocation where the high data rate demand of user 3 is not satisfied.

Figure 3 illustrates the tradeoff between the maximized weighted-sum rate and the fairness constraint when the minimum rate demand is relaxed to $(1 - \eta_q)R_k$ with $0 \leq \eta_q \leq 0.8$. The average user rates are updated according to (4) with $\tau_c = 20$. The maximum weighted-sum rate is achieved when $\alpha_f = 0$ which is very unfair. We can see that, as the fairness constraint is enforced, the weighted-sum rate decreases. For $\alpha_f = 1$, the allocation is strictly fair but inefficient in terms of sum rate. Looking at the solution obtained for different values of α_f, the system designer may then make a choice about the configuration he considers to be the most appropriate.

The tradeoff between reduced complexity and performance degradation of the proposed CDIT-based resource allocation in comparison with the perfect CSIT allocation is shown in Figure 4. Adapting the transmission strategy to the short-term channel statistics, that is, reducing N_s increases the performance but also increases the complexity. However, if the transmission is adapted to the long-term channel statistics, that is, for larger N_s, the complexity decreases significantly but with a penalty on the performance. For a CDIT-based allocation with a distribution taken over 16 OFDM symbol periods, the complexity is reduced by 93.75% while the performance degradation in terms of weighted-sum rate is less than 15%.

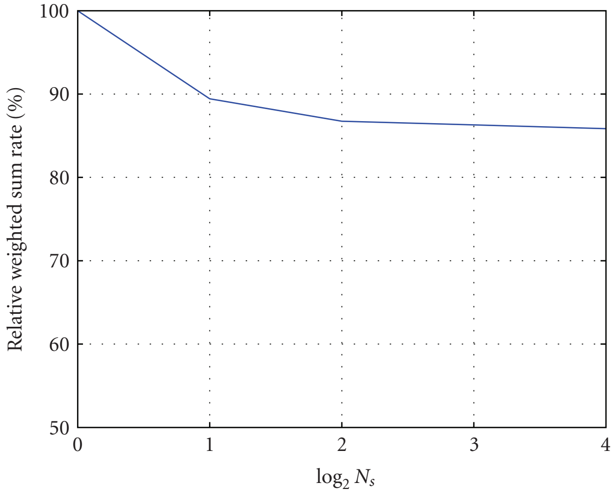


Figure 4: Tradeoff between reduced complexity and performance degradation of the CDIT-based allocation compared with the perfect CSIT.

5. Conclusion

In this paper, we have presented a resource allocation method that maximizes the ergodic weighted-sum rate of a multiuser mobile WiMAX while satisfying user's specific minimum rate demand and system fairness requirement for a given power budget. Though this is originally a nonlinear optimization problem, the problem can be reformulated as a Lagrangian dual problem. From this, a method has been proposed to efficiently solve the problem. The proposed method can find the optimal solution with significant lower computational complexity than the optimal CSIT-based allocation schemes. In fading environments, even with CDIT only, adaptive resource allocation strategies provide performance gain for OFDMA systems. Since user mobility is the principal driving force for mobile WiMAX, CDIT-based resource allocation strategies are of particular interest. These methods can be applied to other mobile OFDMA-based wireless systems such as Long Term Evolution (LTE) or High-Speed Downlink Packet Access (HSDPA).

Acknowledgment

This work was supported in part by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (Contract no. 216715).

References

[1] WiMAX Forum, "Mobile WiMAX—part I: a technical overview and performance evaluation," March 2006.

[2] WiMAX Forum, "A comparative analysis of mobile WiMAX deployment alternatives in the access networks," May 2007.

[3] S. Pietrzyk, *OFDMA for Broadband Wireless Access*, Artech House, London, UK, 2006.

[4] C. Y. Wong, C. Y. Tsui, R. S. Cheng, and K. B. Letaief, "A real-time sub-carrier allocation scheme for multiple access downlink OFDM transmission," in *Proceedings of the 50th IEEE Vehicular Technology Conference (VTC '99)*, vol. 2, pp. 1124–1128, Amsterdam, The Netherlands, September 1999.

[5] M. Ergen, S. Coleri, and P. Varaiya, "Qos aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Transactions on Broadcasting*, vol. 49, no. 4, pp. 362–370, 2003.

[6] D. Kivanc, G. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Transactions on Wireless Communications*, vol. 2, no. 6, pp. 1150–1158, 2003.

[7] N. Damji and T. Le-Ngoc, "Adaptive downlink resource allocation strategies for real-time data services in OFDM cellular systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2006, Article ID 17526, 11 pages, 2006.

[8] T. Sartenaer, L. Vandendorpe, and J. Louveaux, "Balanced capacity of wireline multiuser channels," *IEEE Transactions on Communications*, vol. 53, no. 12, pp. 2029–2042, 2005.

[9] T. Sartenaer, J. Louveaux, and L. Vandendorpe, "Balanced capacity of wireline multiple access channels with individual power constraints," *IEEE Transactions on Communications*, vol. 56, no. 6, pp. 925–936, 2008.

[10] I. C. Wong, Z. Shen, B. L. Evans, and J. G. Andrews, "A low complexity algorithm for proportional resource allocation in OFDMA systems," in *Proceedings of IEEE Workshop on Signal Processing Systems (SiPS '04)*, pp. 1–6, Austin, Tex, USA, October 2004.

[11] X. Zhang, E. Zhou, R. Zhu, S. Liu, and W. Wang, "Adaptive multiuser radio resource allocation for OFDMA systems," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '05)*, vol. 6, pp. 3846–3850, St. Louis, Mo, USA, November 2005.

[12] Z. Han, Z. Ji, and K. J. R. Liu, "Fair multiuser channel allocation for OFDMA networks using Nash bargaining solutions and coalitions," *IEEE Transactions on Communications*, vol. 53, no. 8, pp. 1366–1376, 2005.

[13] Y. Yao and G. B. Giannakis, "Rate-maximizing power allocation in OFDM based on partial channel knowledge," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 1073–1083, 2005.

[14] F. Brah, L. Vandendorpe, and J. Louveaux, "Constrained resource allocation in OFDMA downlink systems with partial CSIT," in *Proceedings of IEEE International Conference on Communications (ICC '08)*, pp. 4144–4148, Beijing, China, May 2008.

[15] C. T. K. Ng and A. J. Goldsmith, "Capacity and power allocation for transmitter and receiver cooperation in fading channels," in *Proceedings of IEEE International Conference on Communications (ICC '06)*, vol. 8, pp. 3741–3746, Istanbul, Turkey, July 2006.

[16] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[17] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, NY, USA, 1964.

[18] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Boston, Mass, USA, 2nd edition, 1999.

[19] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Transactions on Communications*, vol. 54, no. 7, pp. 1310–1322, 2006.

[20] Recommendation ITU-R M.1225, "Guidelines for evaluation of radio transmission technologies for IMT-2000," 1997.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 212783, 10 pages
doi:10.1155/2009/212783

Research Article

Cross-Layer Resource Scheduling for Video Traffic in the Downlink of OFDMA-Based Wireless 4G Networks

Feroz A. Bokhari,[1] Halim Yanikomeroglu,[1] William K. Wong,[2] and Mahmudur Rahman[1]

[1] *Broadband Communications and Wireless Systems Centre, Department of System and Computer Engineering, Carleton University, Ottawa, ON, Canada K1S 5B6*

[2] *Terrestrial Wireless Systems Branch, Communication Research Centre of Canada, 3701 Carling Avenue, P.O. Box 11490 Station H, Ottawa, ON, Canada K2H 8S2*

Correspondence should be addressed to Mahmudur Rahman, mmrahman@sce.carleton.ca

Received 27 June 2008; Accepted 30 December 2008

Recommended by Zhu Han

Designing scheduling algorithms at the medium access control (MAC) layer relies on a variety of parameters including quality of service (QoS) requirements, resource allocation mechanisms, and link qualities from the corresponding layers. In this paper, we present an efficient cross-layer scheduling scheme, namely, Adaptive Token Bank Fair Queuing (ATBFQ) algorithm, which is designed for packet scheduling and resource allocation in the downlink of OFDMA-based wireless 4G networks. This algorithm focuses on the mechanisms of efficiency and fairness in multiuser frequency-selective fading environments. We propose an adaptive method for ATBFQ parameter selection which integrates packet scheduling with resource mapping. The performance of the proposed scheme is compared to that of the round-robin (RR) and the score-based (SB) schedulers. It is observed from simulation results that the proposed scheme with adaptive parameter selection provides enhanced performance in terms of queuing delay, packet dropping rate, and cell-edge user performance, while the total sector throughput remains comparable. We further analyze and compare achieved fairness of the schemes in terms of different fairness indices available in literature.

1. Introduction

The approaching fourth-generation (4G) wireless communication systems, such as the Third-Generation Partnership Project's Long Term Evolution (3GPP LTE) [1] and the IEEE 802.16 standards family (e.g., [2]), are projected to provide a wide variety of new multimedia services, ranging from high quality voice to other high-data-rate wireless applications. Another notable 4G wireless effort is the WINNER project, which aims to develop an innovative concept in radio access in order to achieve high flexibility and scalability with respect to data rates and radio environments [3]. Concepts developed in the WINNER project are applicable to evolving 4G standards due to common system considerations such as orthogonal frequency-division multiple access- (OFDMA-) based air interface, and support of relays and multiple-antenna configurations.

Unlike wireline networks, wireless resources are scarce. The data-rate capacity that a radio-frequency channel can support is limited by Shannon's capacity law. Moreover, due to the time-varying nature of wireless channel, radio resource management, especially packet scheduling and resource allocation, is crucial for wireless networks. Traditionally, the research on packet scheduling has emphasized QoS and fairness issues, and opportunistic scheduling algorithms have focused on exploiting the time-varying nature of the wireless channels in order to maximize throughput. This segregation between packet scheduling and radio resource allocation is inefficient. As fairness and throughput are reciprocally related, an intelligent compromise is necessary to obtain the required QoS while exploiting the time-varying characteristics of the wireless channel. Therefore, it is important to merge the packet scheduling and the resource allocation to design a cross-layer scheduling scheme [4].

A number of scheduling schemes in the literature analyze physical- (PHY-) and MAC-related design issues by assuming that all users are backlogged, that is, all users in the system have nonempty buffers. However, it is shown in [5] that this assumption is not always correct, since the number of packets in the buffers can vary significantly, and there is a relatively high probability that the buffers are empty. For example, in time-slotted networks, the packets in the queues are aggregated into time slots. Consequently, empty queues and partially filled time slots will affect the system performance. Furthermore, these non-queue-aware scheduling algorithms lack the capability to provide required fairness among user terminals (UTs). Hence, it becomes necessary to consider queue states in scheduling and resource allocation [6].

In recent years, some schemes have considered integrating packet scheduling and radio resource scheduling into queue and channel aware scheduling algorithms. In [7], a weighted fair queuing (WFQ) scheduling scheme is proposed, where the largest share of the radio resources is given to the users with the best instantaneous channel conditions in a code division multiplexing (CDM-) based network. Another example of a queue- and channel-aware scheduling algorithm is the modified-largest weighted delay first (M-LWDF) algorithm, where priorities are given to the users with maximum queuing delays weighted by their instantaneous and average rates [8]. The associated decision metrics in these schemes are based on the combination of the delay and instantaneous channel rates. Finding an optimal metric based on these parameters is difficult due to varying requirements for different service classes.

In this paper, we present a scheduler which comprises packet scheduling and resource mapping taking both queue and channel states into account. In the first level of scheduling (packet scheduling), users to be served are selected based on the token bank fair queuing (TBFQ) algorithm, considering fairness and delay constraints among users. Although TBFQ was originally proposed for single-carrier time-division multiple access (TDMA) systems [9], it has been modified in this study by introducing additional parameters that adaptively interact with the second level of scheduling (resource mapping). These parameters take into account the network loading and the user channel conditions. Based on these parameters, the second-level scheduler assigns resources to the selected users in an adaptive manner that exploits the frequency selectivity of the OFDMA air interface. The modified combined scheduling scheme is called ATBFQ.

The performance of ATBFQ is studied in the context of the WINNER wide-area downlink scenario and is compared to that of the SB scheduling algorithm (which was the baseline scheduling scheme in WINNER) [10] and the RR scheme by extensive simulations. The rest of this paper is organized as follows. In Section 2, the ATBFQ algorithm is described in detail, along with its parameter selection. Methods of fairness assessment are addressed in Section 3. The system model and the simulation parameters are presented in Section 4. Simulation results are provided in Section 5, followed by conclusions in Section 6.

2. ATBFQ Scheduling Algorithm

2.1. Original TBFQ Algorithm. The TBFQ algorithm was initially developed for wireless packet scheduling in the downlink of TDMA systems [9, 11], and was later modified for wireless multimedia services using uplink as well. Its concept was based on the leaky-bucket mechanism which polices flows and conforms them to a certain traffic profile.

A traffic flow belonging to user i is characterized by the following parameters:

λ_i: packet arrival rate,

r_i: token generation rate,

p_i: token pool size,

E_i: counter that keeps track of the number of tokens borrowed from or given to the token bank by flow i.

Each L-byte packet consumes L tokens. For each flow i, E_i is a counter that keeps track of the number of tokens borrowed from or given to the token bank. As tokens are generated at rate r_i, the tokens overflowing from the token pool (of size p_i bytes) are added to the token bank, and E_i is incremented by the same amount. When the token pool is depleted and there are still packets to be served, tokens are withdrawn from the bank by flow i, and E_i is decreased by the same amount. Thus, during periods when the incoming traffic rate of flow i is less than its token generation rate, the token pool always has enough tokens to serve arriving packets, and E_i increases and becomes positive and increasing. On the other hand, during periods when the incoming traffic rate of flow i is greater than its token generation rate, the token pool is emptied at a faster rate than it can be refilled with tokens. In this case, the connection may borrow tokens from the bank. The priority of a connection in borrowing tokens from the bank is determined by the priority index (P_i), given by

$$P_i = \frac{E_i}{r_i}. \tag{1}$$

By prioritizing in this manner, we ensure that flows belonging to UTs that are suffering from severe interference, and shadowing conditions in particular, will have a higher priority index, since they will contribute to the bank more often.

2.2. ATBFQ Algorithm. In this study, the TBFQ algorithm, originally proposed for single carrier TDMA systems, is improved by introducing adaptive parameter selection and extended to suit the WINNER multicarrier OFDMA systems [12]. The motivation behind this modification was to incorporate the design and performance requirements of the scheduler in 4G networks into the original scheme. In such networks, the utilization of the resources and hence the performance of the network can be enhanced by making use of the multiuser diversity provided by the multiple access scheme being used. Also, such networks support users with high mobility. Therefore, in order to make use of the

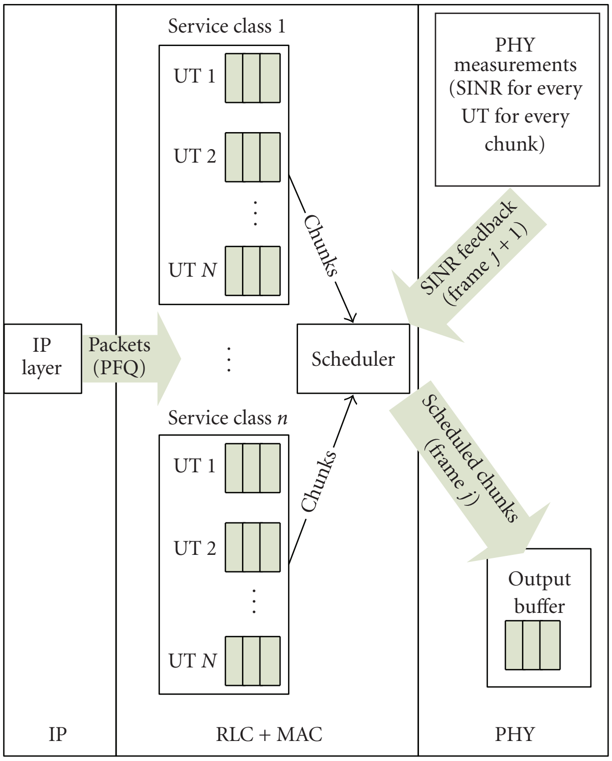


Figure 1: Overview of the proposed cross-layer scheduling operation.

channel feedback, faster scheduling (at a much smaller time scale) is required. Another requirement is the ability to maintain fairness and provide a minimum acceptable QoS performance to all users.

The basic time-frequency resource unit in OFDMA is denoted as a *chunk*. It consists of a rectangular time-frequency area that comprises a number of subsequent OFDM symbols and a number of adjacent subcarriers. Packets from the traffic flows are exclusively mapped on to these chunks based on QoS requirements obtained from the higher radio link control (RLC) layer along with the channel feedback received from the physical layer. The channel feedback comprises signal-to-interference plus noise ratio (SINR) which is measured in the downlink portion of the frame j at the UTs, as shown in Figure 1. This feedback is then provided to the BS in the uplink duration of the frame $j+1$ and can be utilized for scheduling purposes at the MAC layer in the downlink of the next frame, $j+2$. The frame duration, as mentioned in WINNER [13], is 0.6912 milliseconds. The feedback is valid for two frame durations, which is less than the coherence time for mobile speeds of up to 100 km/hr.

Like TBFQ, the ATBFQ scheduling principle is based on the leaky-bucket mechanism. Each traffic flow i is characterized by a packet arrival rate λ_i, token generation rate r_i, token pool size p_i, and a counter E_i to keep track of the number of tokens borrowed from or given to the token bank. Each L-byte packet consumes L tokens. As tokens are generated at rate r_i, the tokens overflowing from the token pool are added to the token bank, and E_i is incremented by the same amount. When the token pool is depleted and there are still packets to be served, tokens are withdrawn from the bank by flow i, and E_i is decremented by the same amount. A debt limit d_i is set as a threshold to limit the amount a UT can borrow from the bank. It also acts as a measure to prevent malicious UTs (transmitting at unusually high transmission rates) from borrowing extensively. The packets are then queued in subqueues in a per-flow queuing (PFQ) manner such that each subqueue belongs to a particular flow, as shown in Figure 1.

The operation of the ATBFQ scheduler is shown by the flowchart shown in Figure 2. This can be summarized by the following steps, which are executed each time the scheduler is invoked at the beginning of the frame.

Step 1. At the scheduler, information is retrieved from the higher layer about all active users using the *getActiveUsers()* function. An active user is defined as a backlogged queue which has packets waiting to be served.

Step 2. Based on this list of active users, a priority is calculated according to the index given by (1). The *highestBorrowPriority()* function is called to calculate this for all active users N_{act}. This function then returns the user i with the highest priority given by

$$i^*(t_k) = \underset{1\le i\le N_{\text{act}}}{\arg\max}(P_i). \tag{2}$$

Step 3. Using the *borrowbudget()* function, a certain budget is calculated for the priority user i^* which depends on the token counter E_i^*, and the debt limit d_i^*, and is given by $E_i^* - d_i^*$. E_i^* keeps track of how much the user has borrowed or given to the bank. The debt limit d_i^* keeps track of how much a user can further borrow from the bank in order to accommodate the burstiness of the traffic over the long term.

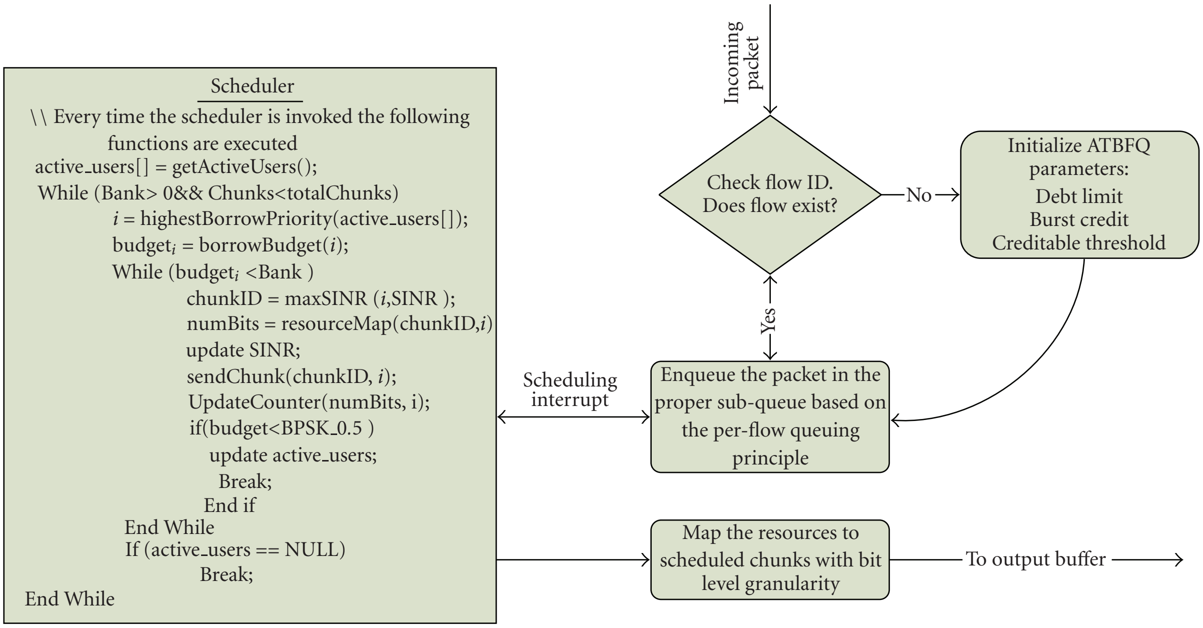
Step 4. If the calculated budget is less than the bank size, resources are allocated to the user i using the *maxSINR()* function. This is the second level of scheduling, and deals with allocation of chunk resources to the selected user i. This allocation is based on the maximum SINR principle, where the chunk j with the best SINR is given to the selected user [14] and can be expressed by

$$j^*(t_k) = \underset{1\le j\le N_{\text{chunks}}}{\arg\max}\left(\gamma_{ij}(t_k)\right), \tag{3}$$

where γ_{ij} is the SINR of the selected user i in chunk j. This is the most opportunistic of all scheduling algorithms for time-slotted networks. This means that the adaptive modulation and coding (AMC) policy maximally exploits the frequency diversity of the time-frequency resource, where a chunk is allocated to only one user and a user can have multiple chunks in a scheduling instant.

Step 5. The *resourceMap()* function determines the amount of bits that can be mapped to the chunk depending on the AMC mode used.

Step 6. Each time a chunk resource is allocated, the *updateCounter()* function is called. This function updates the bank, the counter E_i, and the allocated budget.



```
Scheduler
\\ Every time the scheduler is invoked the following
        functions are executed
active_users[] = getActiveUsers();
While (Bank> 0&& Chunks<totalChunks)
        i = highestBorrowPriority(active_users[]);
        budget_i = borrowBudget(i);
        While (budget_i <Bank )
                chunkID = maxSINR (i,SINR );
                numBits = resourceMap(chunkID,i)
                update SINR;
                sendChunk(chunkID, i);
                UpdateCounter(numBits, i);
                if(budget<BPSK_0.5 )
                    update active_users;
                    Break;
                End if
          End While
          If (active_users == NULL)
                Break;
End While
```

Incoming packet
Check flow ID. Does flow exist?
No
Initialize ATBFQ parameters:
Debt limit
Burst credit
Creditable threshold
Yes
Scheduling interrupt
Enqueue the packet in the proper sub-queue based on the per-flow queuing principle
Map the resources to scheduled chunks with bit level granularity
To output buffer

Figure 2: Flowchart of scheduling operation.

The selected user i gets to transmit as long as (1) its queue remains backlogged and (2) the allocated budget is less than the total bank size and more than the number of bits that can be supported with the lowest AMC mode (binary phase-shift keying (BPSK) rate-1/2, considered in this study). If either of these conditions is not satisfied, the user is classified as nonactive. A new priority is calculated on the updated active users, and Steps 1–6 are repeated. This procedure is repeated until there are no chunk resources available or there are no active users.

2.3. ATBFQ Parameter Selection. The performance of the ATBFQ scheduler depends on its parameters that define the debt limit, the burst credit (BC), and the token generation rate. The token generation rate is critical to the extent to which the burstiness of the UT traffic can be accommodated. A UT in its burst mode transmits more data in a short interval of time than its actual statistics, and hence, requires more resources in order to maintain a certain QoS level. The debt limit is set to −5 MB. The token generation rate should be large enough to handle instantaneous bursty traffic. In simulations, this generation rate has been considered three times larger than the average packet arrival rate.

The burst credit for flow i (BC_i) determines the amount of bits selected user i^* can receive in a frame. While this quantity was a fixed value in TBFQ, it is adaptive in ATBFQ. In a cellular network, the user loading level in terms of active users per sector is highly dynamic, due to the ON and OFF characteristics of the bursty traffic. It is observed through simulations that for low-loading cases, a higher value for BC_i performs better, as shown in Table 1. On the other hand, for high-loading conditions, a lower value for BC_i is desired as it exploits multiuser diversity, as shown in Table 2. It is further seen that for both low- and high-loading conditions, BC_i should be adapted per user basis in order to obtain high spectral efficiency. For UT i, this adaptive value can be formulated as

$$\mathrm{BC}_i = \frac{\eta_i(\text{bits/sec/Hz}) \times M(\text{Hz} \cdot \text{ sec}) \times N_{\text{chunks}}}{N_{\text{act}}}, \qquad (4)$$

where η_i is the past spectral efficiency, N_{chunks} is the number of available chunks, M is the amount of time-frequency resources in a chunk, and N_{act} is the number of active UTs in that particular scheduling frame. η_i is a moving average which is updated each time by averaging over the past 100 transmissions of user i.

3. Fairness Study

Opportunistic scheduling algorithms aim to provide high throughput for UTs having good channel conditions (closer to the BS), and consequently, experience a degraded performance. In such cases, the overall throughput of the system is maximized but the fairness amongst UTs is greatly affected. Therefore, it is essential to design a performance metric that is an appropriate indicator of the fairness. One such index is the *Jain's fairness index* proposed in [15]. This fairness index is bounded between zero and unity, and has been widely used [16, 17]. If a system allocates resources to n contending UTs such that the ith user receives an allocation x_i, then this fairness index $f_I(x)$ is given by

$$f_I(x) = \frac{\left[\sum_{i=1}^{n} x_i\right]^2}{n\sum_{i=1}^{n} x_i^2}, \qquad (5)$$

where $x_i \geq 0$. This index measures the equality of UT allocation x. If x_is are equal for all UTs, then the fairness index is 1 and the system is 100% fair, and vice versa. In this

Table 1: Burst credit for ATBFQ for low loading (8 users).

Burst credit (BC)	Queuing delay (sec)	Packets dropped (per frame)	Throughput (Byte per frame)	Spectral efficiency (bits/sec/Hz)
BC = 1000	0.025	4.36	815.4	2.37
BC = 5000	0.017	0.76	1473.3	2.05
BC = 10000	0.015	0.42	1546.6	1.98
Adaptive BC	0.012	0.30	1551.1	2.34

Table 2: Burst credit for ATBFQ for high loading (20 users).

Burst credit (BC)	Queuing delay (sec)	Packets dropped (per frame)	Throughput (Byte per frame)	Spectral efficiency (bits/sec/Hz)
BC = 1000	0.044	3.19	2299.4	2.09
BC = 5000	0.036	3.98	2094.0	1.88
BC = 10000	0.033	4.00	2090.4	1.87
Adaptive BC	0.038	2.01	2497.1	2.29

paper, the allocation metric "x" is defined as the ratio of UT throughput and queue size, and is given by

$$x_i = \frac{TP_i^{(t_1,t_2)}}{Q_i^{(t_1,t_2)}}, \tag{6}$$

where $TP_i^{(t_1,t_2)}$ is the transmitted throughput in bits for UT i during the time interval $[t_1, t_2]$ and $Q_i^{(t_1,t_2)}$ is the total number of packets arriving in the queue for UT i during (t_1, t_2). In simulations, (t_1, t_2) is chosen to be equal to 16 frame time durations.

In (6), the throughput is normalized to avoid ambiguity since the throughput alone as a metric does not provide an insight into the overall fairness.

Another method of fairness assessment, proposed in WiMAX standard [18], is determined by the normalized cumulative distributive function (CDF) of throughput per UT. The normalized UT throughput with respect to the average throughput, $\widetilde{T}_i$ for UT i, is expressed by

$$\widetilde{T}_i = \frac{T_i}{(1/n)\sum_{j=1}^{n} T_j}, \tag{7}$$

where T_i is the instantaneous throughput of UT i in a particular frame, and n is the total number of UTs. As stated in [18], the CDF of this normalized throughput should lie to the right of the coordinates (0.1, 0.1), (0.2, 0.2), and (0.5, 0.5).

The results using both of these fairness assessment methods are discussed in detail in Section 5.

4. System Model and Simulation Parameters

ATBFQ is studied in the wide-area downlink scenario. To reduce the simulation complexity, the bandwidth is reduced to 15 MHz from the original 45 MHz. The chunk dimension is given as 8 subcarriers by 12 OFDM symbols or 312.5 kHz × 345.6 microseconds. The frame duration is defined as 691.2 microseconds, that is, there are a total of 96 chunks per frame.

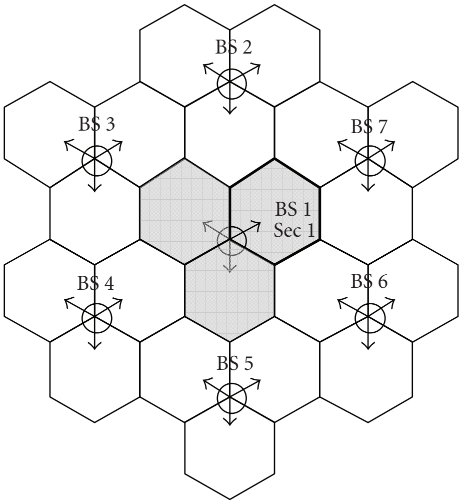


Figure 3: Network layout.

The network layout under investigation is shown in Figure 3. Each cell in the network has three sectors. A frequency reuse factor of 1 in each sector (all resources are used in each sector) is assumed. The UTs are uniformly placed in the central sector.

Time- and frequency-correlated Rayleigh channel samples obtained from power delay profile for the WINNER wide area scenario are used to generate the channel fading. The user speed is defined as 70 km/hr, and the intersite distance is 1 km. The following exponential path-loss model has been used [19]

$$\text{PL} = 38.4 + 35.0 \log_{10}(d)[\text{dB}], \tag{8}$$

where PL is the path loss in dB, and d is the transmitter-receiver separation in meters.

The average thermal noise power is calculated with a noise figure of 7 dB. We have considered independent lognormal random variables with a standard deviation of

Table 3: Lookup table for AMC modes and corresponding chunk throughput.

AMC mode	SINR (dB)	Chunk throughput (bits)
BPSK 1/2	0.2311 ≥ SINR > −1.7	48
BPSK 2/3	1.231 ≥ SINR > 0.231	72
QPSK 1/2	3.245 ≥ SINR > 1.231	96
QPSK 2/3	4.242 ≥ SINR > 3.245	128
QPSK 3/4	6.686 ≥ SINR > 4.242	144
16QAM 1/2	9.079 ≥ SINR > 6.686	192
16QAM 2/3	10.33 ≥ SINR > 9.079	256
16QAM 3/4	14.08 ≥ SINR > 10.33	288
64QAM 2/3	15.6 ≥ SINR > 14.08	384
64QAM 3/4	SINR > 15.6	432

8 dB for shadowing. Sector transmit power is assumed to be 46 dBm, and chunks are assigned fixed equal powers.

The interference is modeled by considering the effect of intercell interference and intracell interference on the sector of interest in the central cell (denoted as sector 1 in BS 1). For this purpose, the interference from the first tier is taken into account. In this case, for a link of interest in sector 1 in BS 1, the interference will comprise 18 (6 BS × 3 sectors) intercell and 2 intracell links.

The SINR obtained for chunk j of user i can be expressed by

$$\mathrm{SINR}_{i,j} = \frac{P^{1,1}_{\mathrm{signal}\,i,j}}{\left(P_{\mathrm{inter}\,i,j} + P_{\mathrm{intra}\,i,j}\right) + P_{\mathrm{noise}\,i,j}}, \tag{9}$$

where $P^{1,1}_{\mathrm{signal}\,i,j}$ denotes the desired signal power in sector 1 in BS 1, and $P_{\mathrm{noise},i}$ is the noise power. For the given layout in Figure 3, intracell interference $P_{\mathrm{intra},i,j}$, and intercell interference $P_{\mathrm{inter},i,j}$ are given by the following expressions:

$$\begin{aligned} P_{\mathrm{intra}\,i,j} &= \sum_{s=2}^{3} I_j^{b=1,s} X_I, \\ P_{\mathrm{inter}\,i,j} &= \sum_{b=2}^{7} \sum_{s=1}^{3} I_j^{b,s} X_I, \end{aligned} \tag{10}$$

where $I_j^{b,s}$ is the interference power for chunk j from sector s in BS b. X_I has a binary value defined by

$$X_I = \begin{cases} 1, & x \leq \mathrm{AF}, \\ 0, & x > \mathrm{AF}, \end{cases} \tag{11}$$

where x is a uniform random variable defined over $[0,1]$, and AF (activity factor) is defined as a probability for a particular interfering link to be active. For example, AF of 1 denotes a high level of interference where all the links are active interferers (100% interference).

Adaptive modulation with block low-density parity-check (B-LDPC) code is used. Thresholds for transmission schemes are determined assuming a block length of 1704 bits and 10% block error rate (BLER) as shown in Table 3 [13]. A chunk using quadrature phase-shift queueing (QPSK) rate-1/2 can carry 96 information bits. This is based on the initial transmissions, that is, hybrid automatic repeat request (HARQ) retransmissions are not considered. Real-time video streaming traffic is used in this study. Two interrupted renewal process (IRP) sources are superimposed to model user's video traffic in the downlink transmission as indicated in [20]. The average packet rate of one UT is 1263.8 packets per second. The resulting downlink data rate for each user is 1.92 Mbps.

The performance of the ATBFQ algorithm is compared to that of the RR and the SB algorithms. The SB algorithm was proposed in [10], and was modified to the WINNER multicarrier OFDMA system for this work. It is a variation of the proportional fair (PF) algorithm which is the most widely adopted opportunistic scheduling algorithm [21]. The SB scheduler selects user i in slot k with the best score, where the score is calculated based on the current rank of the user's SINR among its past values in the current window $\{\gamma_i(t_k), \gamma_i(t_{k-1}), \dots, \gamma_i(t_{k-W+1})\}$, where $\gamma_i(t_k)$ is the SINR value of a user at time instant k, and W is the window size. The corresponding score for the user i is given by

$$s_i(t_k) = 1 + \sum_{l=1}^{W-1} 1_{\{r_i(t_k) < r_i(t_{k-l})\}} + \sum_{l=1}^{W-1} 1_{\{r_i(t_k) = r_i(t_{k-l})\}} X_l, \tag{12}$$

where X_l are i.i.d. random variables on $\{0,1\}$ with $P_r(x = 0) = P_r(x = 1) = 0.5$.

Packets are scheduled on a frame-by-frame basis at the start of every frame. Any packet that arrives at current frame time will have to wait at least until the start of the next frame. As video streaming has the most stringent delay requirement, packets are dropped if they experience a delay in excess of 190 milliseconds. The simulation parameters are summarized in Table 4; most of them are taken from the WINNER baseline simulation assumptions [13].

5. Simulation Results

The performance results are classified into four categories: (1) average user statistics, (2) performance of the cell-edge users, (3) effect of varying user loading and interference conditions, and (4) fairness analysis. Furthermore, the results are compared to the SB and RR algorithms. The window size plays an important role in the performance of the SB algorithm [10]. The performance of ATBFQ has been studied with different window sizes in the WINNER context [22, 23].

5.1. User Performance. Figure 4 shows the CDF of the packets dropped per frame for low and high loading, respectively. These curves indicate the opportunistic nature of SB, since it tends to favor the users with good channel conditions. Consequently, a higher drop rate, even at low loading, is observed for SB.

The CDF of average user throughput per sector (measured in bytes per frame) for 8 and 20 user loading cases is shown in Figure 5. ATBFQ performs better for the

Table 4: Summary of simulation parameters.

Parameter	Used value/model
Scenario	Wide area DL (frequency adaptive)
Channel model	WINNER C2 channel
Shadowing	Independent lognormal random variables (standard deviation 8 dB)
Sector Tx antenna	120° directional with WINNER baseline antenna pattern
UT receive antenna	Omnidirectional
Intersite distance	1000 m
Signal bandwidth	15 MHz (i.e., 48 chunks which is 1/3rd of the baseline assumptions)
Mobility	70 km/hr
Sector Tx power	46 dBm
Scheduler	Adaptive Token Bank Fair Queuing, score based, and round-robin
Interference model	brute force method (central cell is considered with interference from the 1st-tier)
Antenna configuration	Single-in-single-out (SISO)
Coding	B-LDPCC
AMC modes	BPSK (rate 1/2 and 2/3), QPSK (rate 1/2, 2/3, and 3/4), 16QAM (rate 1/2, 2/3, and 3/4), and 64QAM (rate 2/3 and 3/4)
AMC thresholds	With FEC block of 1728 bits and 10% BLER
Frame duration	0.6912 ms (scheduling interval)
Traffic model	1.9 Mbps 2IRP model for MPEG video
Packet size	188 Bytes
Packet drop criterion	Delay ≥ 0.19 sec
Simulation time	60 sec
Simulation tools	MATLAB and OPNET

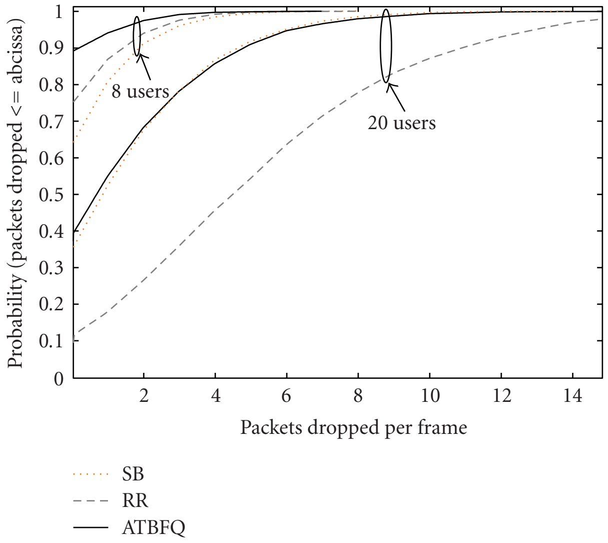


Figure 4: CDF of packets dropped per user per frame.

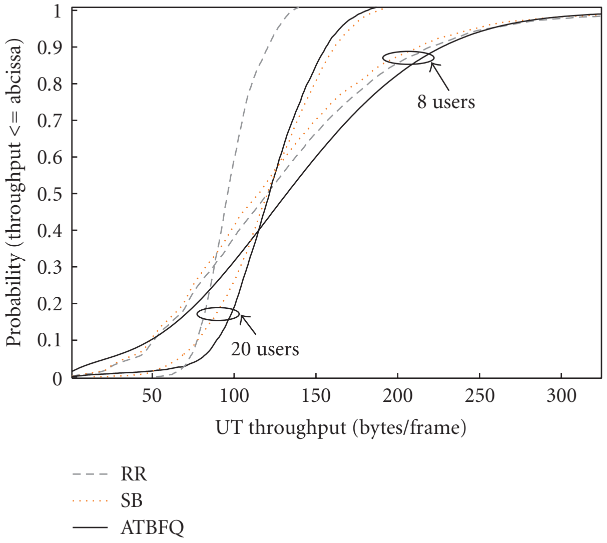


Figure 5: CDF of user throughput.

lower loading case, whereas SB achieves marginally higher throughput at higher loading. For the high loading case, it is observed that the CDF curve for ATBFQ has a steeper slope indicating better fairness, since users are served with similar throughput. Note that this is not true for SB. As ATBFQ attempts to maintain fairness, it tries to serve cell-edge users with poor channel conditions as compared to those located closer to the BS. Therefore, ATBFQ also utilizes more chunks. On the other hand, SB aims to maximize the throughput while being fair in the opportunistic sense.

5.2. Cell-Edge User Performance. Figure 6 shows the packet transmit ratio (defined as the transmitted packet per total packets generated) versus distance from BS for 20 users per sector. It can be observed that as the distance increases, the packet transmit ratio for SB decreases, that is, the number of

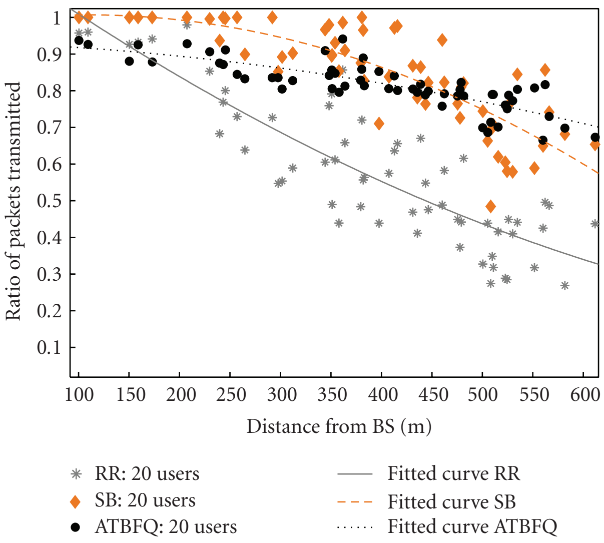


Figure 6: Ratio of packets dropped versus distance form BS.

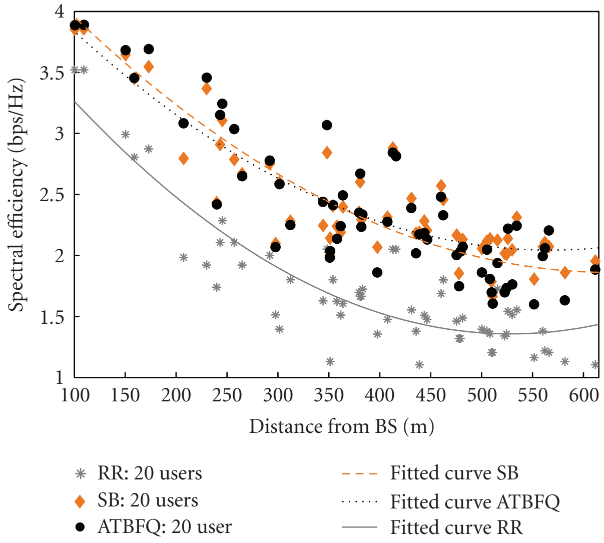


Figure 7: Average user spectral efficiency versus distance form BS.

dropped packets increases. This can be further visualized by the quadratic-fitted curves for both algorithms, which show their respective trends with the varying distance. As SB tries to maximize the throughput, the cell-edge users are affected, and suffer packet losses. ATBFQ, on the other hand, is fair in nature and shows enhanced performance for the cell-edge users. If a cell-edge user is suffering from poor channel conditions, ATBFQ gives it priority to transmit in the next scheduling interval. By assigning priorities in such a manner, ATBFQ considerably improves the spectral efficiency for the cell-edge users, as shown in Figure 7.

5.3. Varying User Loading and Interference Conditions. Performance indicators such as average dropped packets, average UT throughput, and average UT queuing delay have been considered in evaluating ATBFQ by comparison with the reference SB and RR schemes.

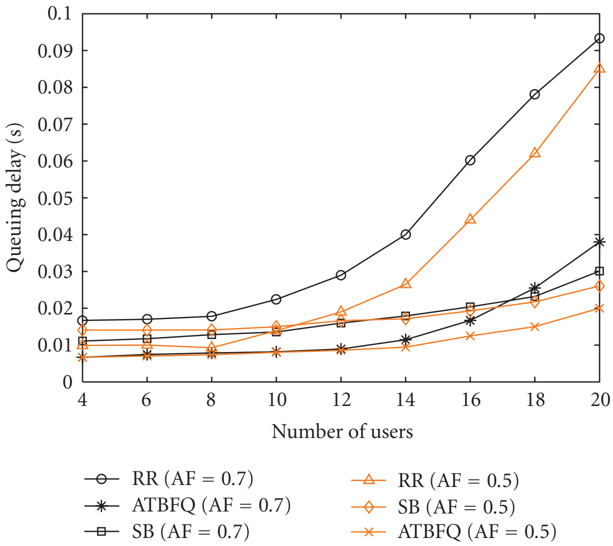


Figure 8: Average UT queuing delay versus number of UTs.

Figures 8, 9, and 10 show the performance results for average UT queuing delay, average packets dropped per frame, and the total sector throughput, respectively, in varying loading conditions for ATBFQ, SB, and RR. The curves are plotted for two different AFs of 0.5 and 0.7 to model moderate and high interference situations, respectively. ATBFQ outperforms the reference SB and RR algorithms in terms of the above-mentioned performance parameters for all loading conditions when the AF is 0.5. In this case, the UTs experience better channel conditions resulting from low interference. Hence, fewer chunks are utilized to transmit data as compared to the number of chunks utilized for a higher AF. Consequently, RR performs better than SB at lower loading levels.

For low-to-medium loading with an AF of 0.7, it is observed again that ATBFQ outperforms the reference schemes in terms of all observed parameters. This trend changes as network loading increases to 20 UTs per sector. In this case, SB outperforms ATBFQ and RR in terms of average UT queuing delay, average packets dropped per frame, and the total sector throughput, respectively. This is due to the fact that SB is opportunistic in nature, whereas ATBFQ is fairness aware. As the number of UTs increases, SB takes advantage of the multiuser diversity to achieve higher throughput.

5.4. Fairness Analysis. The CDF of the Jain's fairness index given by (5) is shown in Figure 11. These curves represent network loading of 20 UTs per sector with an AF of 0.7. It is observed that ATBFQ achieves better fairness compared to SB and RR. Figure 12 shows the CDF plot of the normalized throughput given by (7) for 20 UTs per sector with an AF of 0.7. By normalizing the throughput, the performance of the cell edge users represented by the tail of the throughput CDF curve is enhanced. It is again observed that a higher

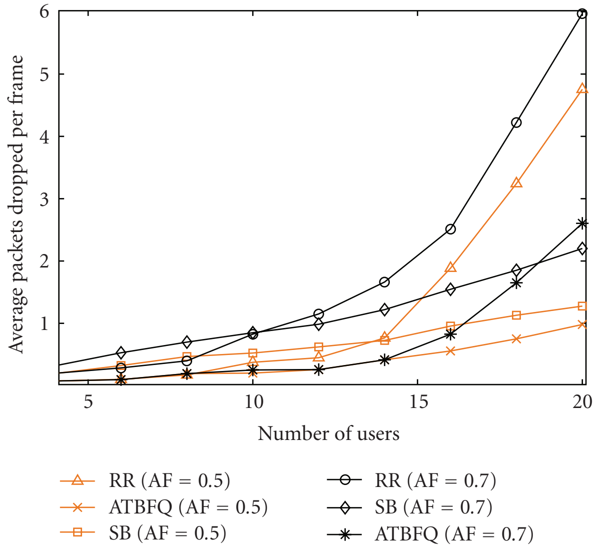


Figure 9: Average UT packets dropped per frame versus number of UTs.

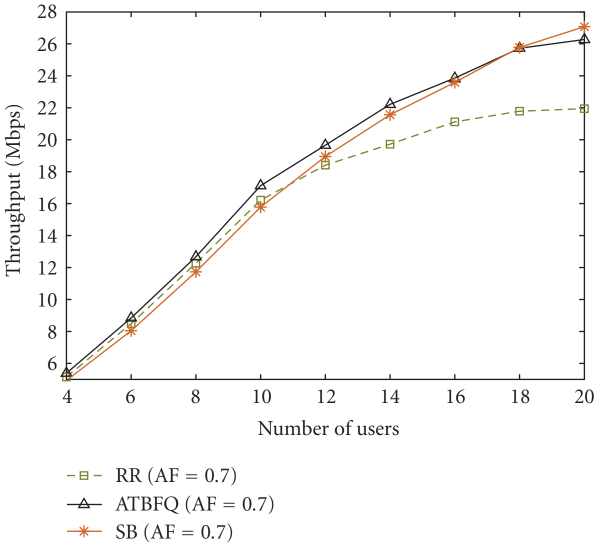


Figure 10: Sector throughput.

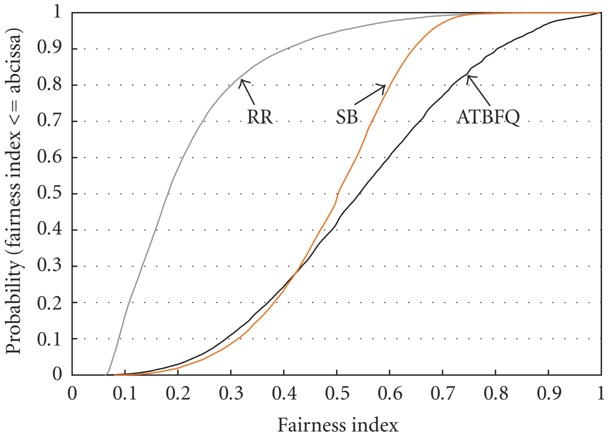


Figure 11: CDF of fairness index.

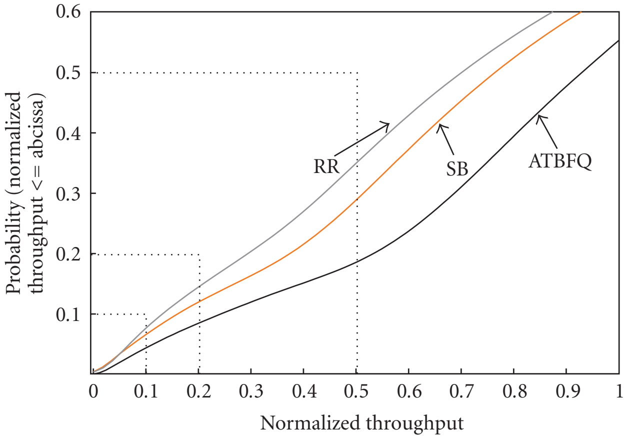


Figure 12: CDF of normalized throughput (zoomed in).

normalized throughput is achieved for ATBFQ compared to that in SB, and the curve lies to the right of the above-mentioned coordinates.

6. Conclusion

In this paper, the performance of the ATBFQ scheduling algorithm with adaptive parameter selection is investigated in the context of the 4G WINNER wide-area downlink scenario. It is a queue- and channel-aware scheduling algorithm which attempts to maintain fairness among all users. Performance of ATBFQ is presented with reference to the SB and RR schedulers. Being an opportunistic scheduler belonging to the proportional fair class, SB aims to maximize throughput by making use of multiuser diversity while trying to maintain fairness. However, this comes at a certain cost, since the cell edge users in this scheme, suffering from poor channel conditions, are more severely affected. Also, due to the bursty nature of the traffic, such users experience higher queueing delays, resulting in a higher number of packet dropping.

Contrary to SB, ATBFQ is a credit-based scheme which aims to accommodate the burstiness of the users by assigning them more resources in the short term, provided that long-term fairness is maintained. For lower to medium loading, ATBFQ provides higher throughput, lower queuing delay, and a lower number of packets dropped as compared to SB and RR. At high loading, ATBFQ still outperforms SB and RR with regard to the queuing delay and packet dropping, however, with a slight degradation in the sector throughput. This is because ATBFQ attempts to satisfy users with poor channel conditions by assigning more resources, even with a lower chunk spectral efficiency. An overall improvement of the performance of cell-edge users is observed in terms of spectral efficiency and packet-dropping ratio for ATBFQ as compared to SB and RR.

The observed throughput, queuing delay, and packet dropping rate clearly indicate the superiority of the ATBFQ algorithm. This apparent improvement in the fairness performance of the ATBFQ algorithm based on these performance parameters is further validated by evaluating the fairness indices available in the literature.

Acknowledgments

The authors would like to express their gratitude to Mr. Jiangxin Hu for his technical support and Dr. Abdulkareem Adinoyi for providing his valuable comments on the manuscript. They also thank OPNET Technologies, Inc. for providing software license to carry out the simulations of this research. This work was a part of the Wireless World Initiative New Radio (WINNER) project, http://www.ist-winner.org/, with the support of the Natural Sciences and Engineering Research Council (NSERC) of Canada. Preliminary results of this work have been presented in IEEE VTC2008-Spring and IEEE VTC2008-Fall conferences.

References

[1] Overall Description: Stage 2 (Release 8), "3GPP Std. 3GPP E-UTRA and E-UTRAN Technical Specification TS 36.300 V8.4.0," March 2008, http://www.3gpp.org/ftp/Specs/html-info/36300.htm.

[2] IEEE 802.16 Std. 802.16j/D5, "Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems—Multihop Relay Specification," June 2008, http://www.ieee802.org/16.

[3] "Project Presentation," WINNER Deliverable D8.1, March 2004, http://www.ist-winner.org/deliverables_older.html.

[4] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 3, pp. 839–847, 2006.

[5] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 1, pp. 321–331, San Francisco, Calif, USA, March-April 2003.

[6] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, 2003.

[7] A. Stamoulis, N. D. Sidiropoulos, and G. B. Giannakis, "Time-varying fair queueing scheduling for multicode CDMA based on dynamic programming," *IEEE Transactions on Wireless Communications*, vol. 3, no. 2, pp. 512–523, 2004.

[8] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–154, 2001.

[9] W. K. Wong and V. C. M. Leung, "Scheduling for integrated services in next generation packet broadcast networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '99)*, vol. 3, pp. 1278–1282, New Orleans, La, USA, September 1999.

[10] T. Bonald, "A score-based opportunistic scheduler for fading radio channels," in *Proceedings of the 5th European Wireless Conference (EW '04)*, Barcelona, Spain, February 2004.

[11] W. K. Wong, H. Y. Tang, and V. C. M. Leung, "Token bank fair queuing: a new scheduling algorithm for wireless multimedia services," *International Journal of Communication Systems*, vol. 17, no. 6, pp. 591–614, 2004.

[12] "Final Report on Identified RI Key Technologies, System Concept, and their Assessment," WINNER I Deliverable D2.10, November 2005, http://www.ist-winner.org/deliverables_older.html.

[13] "Test Scenarios and Calibration Cases Issue 2," WINNER II Deliverable D6.13.7, December 2006, http://www.ist-winner.org/ deliverables.html.

[14] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proceedings of IEEE International Conference on Communications (ICC '95)*, vol. 1, pp. 331–335, Seattle, Wash, USA, June 1995.

[15] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Tech. Rep. DEC-TR-301, Digital Equipment Corporation, Maynard, Mass, USA, September 1984.

[16] H. Sirisena, A. Haider, M. Hassan, and K. Fawlikowski, "Transient fairness of optimized end-to-end window control," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '03)*, vol. 7, pp. 3979–3983, San Francisco, Calif, USA, December 2003.

[17] G. Berger-Sabbate, A. Duda, O. Gaudoin, M. Heusse, and F. Rousseau, "Fairness and its impact on delay in 802.11 networks," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '04)*, vol. 5, pp. 2967–2973, Dallas, Tex, USA, November-December 2004.

[18] IEEE 802.16 Std., "IEEE 802.16m Evaluation Methodology Document," September 2007, http://www.ieee802.org/16.

[19] "Final report on link level and system level channel models," WINNER I Deliverable D5.4, November 2005, http://www.ist-winner.org/deliverables_older.html.

[20] "Traffic model for 802.16 TG3 MAC/PHY simulations," IEEE 802.16 Work-in-progress document 802.16.3c-01/30r1, March 2001, http://www.ieee802.org/16.

[21] E. F. Chaponniere, P. J. Black, J. M. Holtzman, and D. N. C. Tse, "Transmitter directed code division multiple access system using path diversity to equitably maximize throughput," US Patent no. 6449490, September 2002.

[22] "Inteference Avoidance Concepts," WINNER II Deliverable D4.7.2, June 2007, http://www.ist-winner.org/deliverables.html.

[23] F. A. Bokhari, W. K. Wong, and H. Yanikomeroglu, "Adaptive token bank fair queuing scheduling in the downlink of 4G wireless multicarrier networks," in *Proceedings of the 67th IEEE Vehicular Technology Conference (VTC '08)*, pp. 1995–2000, Marina Bay, Singapore, May 2008.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 462396, 14 pages
doi:10.1155/2009/462396

Research Article

Busy Bursts for Trading off Throughput and Fairness in Cellular OFDMA-TDD

Birendra Ghimire,[1] Gunther Auer,[2] and Harald Haas[1, 3]

[1] *Institute for Digital Communications, Joint Research Institute for Signal and Image Processing, The University of Edinburgh, EH9 3JL, UK*
[2] *DOCOMO Euro-Labs, Landsberger Straße 312, 80687 Munich, Germany*
[3] *School of Engineering and Science, Jacobs University Bremen, 28759 Bremen, Germany*

Correspondence should be addressed to Harald Haas, h.haas@ed.ac.uk

Received 1 July 2008; Accepted 8 December 2008

Recommended by Mohamed Hossam Ahmed

Decentralised interference management for orthogonal frequency division multiple access (OFDMA) operating in time division duplex (TDD) cellular systems is addressed. Interference aware allocation of time-frequency slots is accomplished by letting receivers transmit a busy burst (BB) in a time-multiplexed minislot, upon successful reception of data. Exploiting TDD channel reciprocity, an exclusion region around a victim receiver is established, whose size is determined by a threshold parameter, known at the entire network. By adjusting this threshold parameter, the amount of cochannel interference (CCI) caused to active receivers in neighbouring cells is dynamically controlled. It is demonstrated that by tuning the interference threshold parameter, system throughput can be traded off for improving user throughput at the cell boundary, which in turn enhances fairness. Moreover, a variable BB power is proposed that allows an individual link to signal the maximum CCI it can tolerate whilst satisfying a certain quality-of-service constraint. The variable BB power variant not only alleviates the need to optimise the interference threshold parameter, but also achieves a favourable tradeoff between system throughput and fairness. Finally, link adaptation conveyed by BB signalling is proposed, where the transmission format is matched to the instantaneous channel conditions.

1. Introduction

Orthogonal frequency division multiplexing (OFDM) has been selected as a radio access technology for a number of wireless communication systems, for instance, the wireless local area network (WLAN) standard IEEE 802.11 [1], the European terrestrial video broadcasting standard DVB-T [2], and for beyond 3rd generation (B3G) mobile communication systems [3]. In OFDMA, the available resources are partitioned into time-frequency slots, also referred to as *chunks*, which can be flexibly distributed among a number of users who share the wireless medium. Provided that channel knowledge is available at the transmitter, resources can be assigned to users with favourable channel conditions, giving rise to multiuser diversity [4].

Interference management is one of the major challenges for cellular wireless systems, as transmissions in a given cell cause cochannel interference (CCI) to the neighbouring cells. Full-frequency reuse where the transmitters are allowed an unrestricted access to all resources causes high CCI, which particularly impacts the cell-edge users [5–7]. Although CCI can be mitigated by traditional frequency planning, this potentially results in a loss in bandwidth efficiency due to insufficient spatial reuse of radio resources. Fractional frequency reuse (FFR) [4–6, 8] addresses this issue by realising that in the cellular networks CCI predominantly affects users near the cell boundary. FFR typically involves a subband with full-frequency reuse that is exempt from any slot assignment restrictions. The allocation of the remaining subbands is coordinated among neighbouring cells, in a way that the users in the given cell are denied access to subbands assigned to the cell-edge users in the adjacent cells. To this end, in [5] a user is classified as a cell-edge user based on the path loss to the desired base station (BS). This approach ignores the fact that the channel attenuation of the desired and the interfering signals is uncorrelated, and therefore fails to exploit interference diversity. Moreover, frequency planning results in a hard spatial reuse of the available resources. As a result, it cannot cater for the dynamic traffic and load across different sites. Furthermore, in systems where BSs are dynamically added in an uncoordinated manner, such as home base stations [9], reconfigurable frequency reuse planning may prove to be increasingly cumbersome.

The busy-signal concept [10–16] has been identified as an enabler for decentralised and interference aware slot assignment. Receiver feedback informs a potential transmitter about the instantaneous CCI it causes to the "victim" receivers, which enables the transmitter to take appropriate steps to avoid interference, such as deferring its own transmission to another chunk. Early works [10, 11] rely on dedicated frequency-multiplexed channels that carry narrowband busy tones for channel reservation. As the protocol requires the transceivers to listen to the out-of-band busy tones whilst transmitting, complex RF units are required due to additional filters and duplexers involved. This drawback is avoided in [12–14], where time-multiplexed busy bursts (BBs) serve as a reservation indicator for a reservation-based medium access control (MAC) protocol. By transmitting an in-band BB in an associated minislot following a successful transmission, two important goals are accomplished [13, 14]. First, the transmitter of its own link is informed whether or not the data is successfully received. Second, at the same time potential transmitters of the competing links are notified about ongoing transmissions, so that these transmitters can take appropriate steps to avoid interference. Therefore, both slot reservation and channel sensing tasks are accomplished. Furthermore, *interference diversity* is exploited, in the way that competing links may spatially reuse the same slot, given the interfering links are sufficiently attenuated.

None of the busy tone-based MAC protocols [11–14] allow for dynamic resource allocation where multiple users share a set of parallel frequency slots of a broadband frequency-selective radio channel, such as the 100 MHz channel of the WINNER (Wireless world Initiative New Radio, www.ist-winner.org) TDD mode [17].

By extending the BB concept to OFDMA [15, 16], the channel reciprocity of TDD [18] is exploited for decentralised interference management such that the chunks can be dynamically assigned on a short-term basis thereby ensuring a soft spatial reuse of chunks among cells. This concept termed BB-OFDMA works in a completely decentralised fashion and is therefore applicable to self-organising networks, which may consist of cellular as well as *ad hoc* network topologies.

The attainable system throughput of BB-OFDMA is sensitive to the selected interference threshold [15, 16]. In this paper, it is demonstrated how the interference threshold can be tuned to tradeoff system throughput to enhance cell-edge user throughput, thereby enhancing fairness. Moreover, by using a variable BB power that takes into account the quality of the intended link, a favourable tradeoff between system throughput and fairness is achieved. A variable BB power exhibits the further advantage that the sensitivity of the selected interference threshold on the performance is mitigated. Finally, BB-OFDMA with variable BB power is the basis for a novel receiver-driven link adaptation algorithm. System-level simulations demonstrate a significant improvement both in terms of fairness and total system throughput of BB-OFDMA, compared to the system with full-frequency reuse, where attempts to avoid interference are not made.

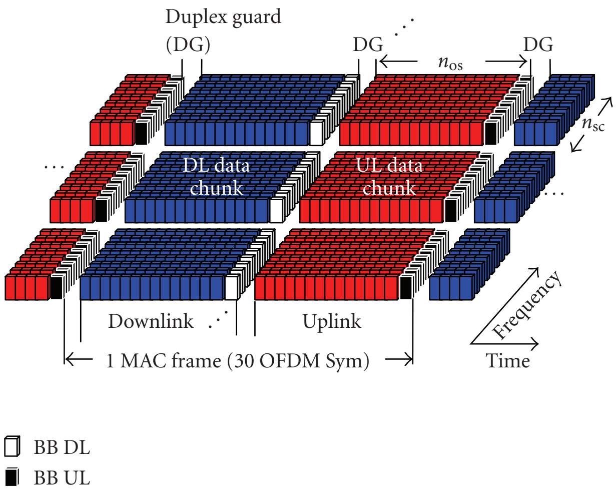


Figure 1: Frame structure for OFDMA-TDD with BB signalling.

The remainder of the paper is arranged as follows. Section 2 describes the air interface of WINNER-TDD. The allocation of radio resources among the competing user population is discussed in Section 3. Section 4 introduces the BB signalling mechanism and its variants as well as the proposed link adaptation algorithm. The considered Manhattan grid deployment scenario and the system level simulator are introduced in Section 5, and the simulation results are discussed in Section 6. Finally, the conclusions are drawn in Section 7.

2. System Model

A time-frequency slotted OFDMA-TDD air interface based on the WINNER-TDD mode [8] is implemented, as illustrated in Figure 1. A chunk comprises of n_{sc} subcarriers and n_{os} OFDM symbols and represents a resource unit that can be allocated to one out of U users located in cell q. Successive downlink (DL) and uplink (UL) slots, each of which contains N_{C} chunks, constitute a frame. A chunk with frequency index $1 \leq n \leq N_{\mathrm{C}}$ at frame k is denoted by (n, k). The transmit power of user ν at chunk (n, k) is denoted by $T^{\mathrm{d}}_{\nu,q}[n, k]$.

The transmitted signal of chunk (n, k) propagates through a mobile radio channel. The corresponding channel gain $G_{\nu,q}[n, k]$ comprises radio effects such as distance-dependent path loss, log-normal shadowing as well as channel variations due to frequency-selective fading and user mobility [19]. While channel variations of $G_{\nu,q}[n, k]$ between adjacent chunks in time and frequency are taken into account, fluctuations within a chunk are neglected. This approximation is justified as long as the chunk dimensions are significantly smaller than the coherence time and frequency [20].

The received signal power of user ν can be expressed as

$$\widetilde{R}^{\mathrm{d}}_{\nu,q}[n,k] = R^{\mathrm{d}}_{\nu,q}[n,k] + I^{\mathrm{d}}_{\nu,q}[n,k] + N, \tag{1}$$

where N is the thermal noise power. Both the received signal powers of the intended and the interfering links, denoted by $R^{\mathrm{d}}_{\nu,q}[n,k] = T^{\mathrm{d}}_{\nu,q}[n,k]G_{\nu,q}[n,k]$ and $I^{\mathrm{d}}_{\nu,q}[n,k]$, may vary significantly between different chunks, as elaborated in more detail in Section 4. The achieved signal-to-interference-plus-noise ratio (SINR) at chunk (n,k) is in the form

$$\gamma_{\nu,q}[n,k] = \frac{R^{\mathrm{d}}_{\nu,q}[n,k]}{I^{\mathrm{d}}_{\nu,q}[n,k] + N}. \tag{2}$$

3. Multiuser Resource Allocation

Provided that only one user per cell transmits on a given chunk, the base station (BS) may flexibly assign chunks to users, such that the intracell interference is avoided. However, as chunks may be simultaneously accessed by adjacent cells, CCI is encountered. Multiuser resource allocation is carried out by a score-based scheduler [21] variant, which distributes the $1 \le n \le N_{\mathrm{C}}$ chunks among $1 \le \nu \le U$ users served by the BS in cell q. Assuming that the channel gains $G_{\nu,q}[n,k]$ are available at BS_q, the score for user ν at chunk (n,k) is computed as

$$s_{\nu,q}[n,k] = 1 + \sum_{\ell=1}^{N_{\mathrm{C}}} \Upsilon_{\{G_{\nu,q}[n,k] \le G_{\nu,q}[\ell,k]\}} + \epsilon_{\nu,q}[n,k], \tag{3}$$

where the Boolean operator $\Upsilon_x \in \{0,1\}$ is set to 1 or 0 when the condition x is true or false, respectively. The parameter $\epsilon_{\nu,q}[n,k] \in \{0,\infty\}$ indicates whether or not user ν is granted access to chunk (n,k). For interference aware and reservation-based MAC protocols such as BB-OFDMA (see Section 4.4), setting $\epsilon_{\nu,q}[n,k] \to \infty$ ensures that user ν in cell q is denied access to chunk (n,k). This effectively avoids radiation of CCI from cell q to any neighbouring cells that use the same chunk (n,k).

Score based multiuser scheduling with reservation assigns chunk (n,k) to user ν if either a reservation indicator was set in the previous frame, $\beta_q[n,k-1] = \nu$, or the score given by (3) is minimised

$$a_q[n,k] = \begin{cases} \arg\min\limits_{\nu} s_{\nu,q}[n,k], & \beta_q[n,k-1] = 0, \\ \beta_{\nu,q}[n,k-1], & \text{otherwise.} \end{cases} \tag{4}$$

In case $\epsilon_{\nu,q}[n,k] \to \infty$ for all users, cell q leaves chunk (n,k) unassigned in (4). The user ν that is assigned chunk (n,k) transmits data to its intended receiver. The set of chunks $n \in \{1,\ldots,N_C\}$ at time k, for which $a_q[n,k] = \nu$ are denoted by $\mathcal{A}_{\nu,q}$. Allocated chunks $a_q[n,k] = \nu$ whose achieved SINR $\gamma_{\nu,q}[n,k]$ exceeds the target Γ, such that

$$b_q[n,k] = \begin{cases} \nu, & a_q[n,k] = \nu \text{ and } \gamma_{\nu,q}[n,k] \ge \Gamma, \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

represent the set of successfully allocated chunks of user ν, denoted by $\mathcal{B}_{\nu,q} \subseteq \mathcal{A}_{\nu,q}$ [15].

For BB-OFDMA chunks with $b_q[n,k] \neq 0$ are reserved and protected from interference at the next frame $k+1$ by setting the reservation indicator to $\beta_q[n,k] = b_q[n,k]$ in (4). When the SINR target is not met, $\gamma_{\nu,q}[n,k] < \Gamma$, the reservation indicator is reset to $\beta_q[n,k] = b_q[n,k] = 0$. These chunks $\mathcal{A}_{\nu,q} \setminus \mathcal{B}_{\nu,q}$ are released in a way that user ν is prohibited access in the next slot $k+1$ by setting $\epsilon_{\nu,q}[n,k+1] \to \infty$. Subsequently, chunk $(n,k+1)$ is assigned to other users by (4).

In a cellular OFDMA system without interference protection, there is no restriction for accessing any chunks, so $\epsilon_{\nu,q}[n,k] = 0\ \forall n,k$ in (3) for all users in the cell. Moreover, no reservation indicator is set, $\beta_q[n,k] = 0\ \forall n,k$ in (4), irrespective of $b_q[n,k]$ in (5).

4. Busy Burst Signalling

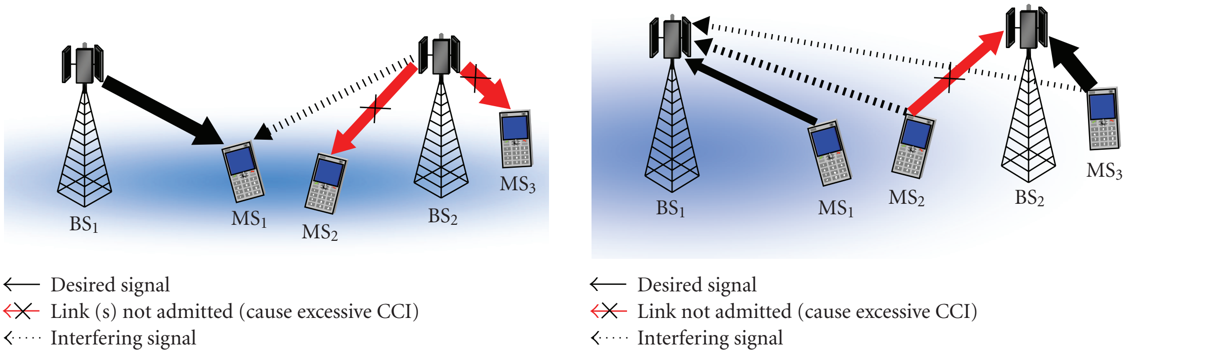
Interference management using busy burst (BB) signalling [13, 14] establishes an exclusion region around active receivers. An exclusion region defines an area around an active receiver in cell q, where potential transmitters in adjacent cells $p \neq q$ must not transmit, so that excessive CCI by close-by interferers is mitigated. In the context of OFDMA, the exclusion regions are to be established individually for each chunk (n,k) [15]. In BB-OFDMA, an MAC frame is divided into data slots and BB minislots as illustrated in Figure 1. The BS transmits data in slot "Data DL." Provided that the SINR target for an allocated chunk (n,k) is met, the intended mobile station (MS) transmits a BB in the associated minislot "BB UL" at uplink chunk (n,k). This reserves chunk n of "Data DL" for the next frame $k+1$. Likewise, for uplink data transmitted by the MS in slot "Data UL," the BB is transmitted by the intended BS in the downlink minislot "BB DL." In summary, BB-OFDMA is described by the following protocol.

(1) All potential transmitters must sense the BB associated to the data chunk (n,k) prior to transmission.

(2) Transmitters are prohibited to access chunks where a BB is detected above a given threshold.

The resulting BB signalling overhead amounts to 6.7%, as 2 OFDM symbols out of 30 OFDM symbols per frame are used for BB signalling. However, instead of dismissing BB signalling as overhead, the BB minislots may be utilised to convey the feedback and control information. Hence, BB signalling may serve as an alternative control channel.

To exemplify the principle of BB-enabled interference avoidance in cellular system, a typical downlink and uplink interference scenario is illustrated in Figure 2. In the downlink shown in Figure 2(a), MS_1 has transmitted a BB after successful reception from BS_1. As BS_2 detects a strong BB from MS_1, BS_2 cannot spatially reuse this chunk with BS_1. In the uplink shown in Figure 2(b), BS_1 has transmitted a BB after successful reception from MS_1. While MS_2 is denied access to this chunk, as it detects a strong BB from BS_1, MS_3 is located outside the exclusion region of BS_1, and may therefore simultaneously access this chunk with MS_1.



(a) Downlink

(b) Uplink

Figure 2: BB signalling applied to cellular system. The arrows depict the direction of desired and interfering signals and their relative strength is indicated by their width. The strength of BB signal is indicated by the darkness of the shade around the vulnerable receiver.

4.1. *Two Competing Links.*

To mathematically describe BB-enabled interference avoidance, let $\mathbf{x} = (\nu, q)$ define a transmitter or receiver (either BS or MS) of user ν within cell q. With this notation, the channel gain of the intended link at chunk (n,k) becomes $G_{\mathbf{x}}[n,k] = G_{\nu,q}[n,k]$. The channel gain of an interfering link, between transmitter $\mathbf{y} = (\mu, p)$ of user μ located in an adjacent cell $p \neq q$ and receiver $\mathbf{x}$, is denoted by $G_{\mathbf{yx}}[n,k]$. In case two links compete for resources, the CCI between transmitter $\mathbf{y}$ and receiver $\mathbf{x}$ amounts to $I^{\mathrm{d}}_{\mathbf{x}}[n,k] = G_{\mathbf{yx}}[n,k]T^{\mathrm{d}}_{\mathbf{y}}[n,k]$. (The term $I^{\mathrm{d}}_{\mathbf{x}}[n,k]$ is equivalent to the CCI $I^{\mathrm{d}}_{\nu,q}[n,k]$ as defined in (1). While the notation $I^{\mathrm{d}}_{\mathbf{x}}[n,k]$ is preferred for intercellular interference management, the latter is used for intracell resource allocation. The same rule applies for related quantities that denote transmitted and received signal powers.) Likewise, $T^{\mathrm{b}}_{\mathbf{x}}[n,k]$ and $I^{\mathrm{b}}_{\mathbf{y}}[n,k] = G_{\mathbf{xy}}[n,k]T^{\mathrm{b}}_{\mathbf{x}}[n,k]$ are the transmit power of the BB transmitter $\mathbf{x}$ (data receiver) and the interfering BB power received at data transmitter $\mathbf{y}$ (BB receiver), respectively.

Exploiting TDD channel reciprocity [18], transmitter $\mathbf{y}$ can ascertain $I^{\mathrm{d}}_{\mathbf{x}}[n,k]$, the potential amount of interference it causes to an existing link $\mathbf{x}$, by measuring $I^{\mathrm{b}}_{\mathbf{y}}[n,k]$ at the associated BB minislot [13]. Applying the channel reciprocity property of TDD, $G_{\mathbf{yx}}[n,k] = G_{\mathbf{xy}}[n,k]$, yields

$$I^{\mathrm{d}}_{\mathbf{x}}[n,k] = I^{\mathrm{b}}_{\mathbf{y}}[n,k] \cdot \frac{T^{\mathrm{d}}_{\mathbf{y}}[n,k]}{T^{\mathrm{b}}_{\mathbf{x}}[n,k]}. \tag{6}$$

The maximum CCI $I^{\mathrm{d}}_{\mathbf{x}}[n,k]$ that a candidate transmitter $\mathbf{y}$ may cause to an active receiver $\mathbf{x}$ is determined by the interference threshold I_{th}, which is constant and known to the entire network. When $I^{\mathrm{d}}_{\mathbf{x}}[n,k] < I_{\mathrm{th}}$, transmitter $\mathbf{y}$ is located outside the exclusion range of $\mathbf{x}$. Provided that $T^{\mathrm{b}}_{\mathbf{x}}[n,k]$ is known to the candidate transmitter $\mathbf{y}$, (6) enables $\mathbf{y}$ to verify whether $I^{\mathrm{d}}_{\mathbf{x}}[n,k] < I_{\mathrm{th}}$ by invoking the threshold test [13, 14]

$$I^{\mathrm{b}}_{\mathbf{y}}[n,k] \cdot \frac{T^{\mathrm{d}}_{\mathbf{y}}[n,k]}{T^{\mathrm{b}}_{\mathbf{x}}[n,k]} \leq I_{\mathrm{th}}. \tag{7}$$

In case $T^{\mathrm{d}}_{\mathbf{y}}[n,k] = T^{\mathrm{b}}_{\mathbf{x}}[n,k]$, condition (7) reduces to

$$I^{\mathrm{b}}_{\mathbf{y}}[n,k] \leq I_{\mathrm{th}}. \tag{8}$$

By tuning I_{th}, the maximum CCI $I^{\mathrm{d}}_{\mathbf{x}}[n,k]$ in (2) is adjusted, which determines the size of the exclusion range around active receivers.

4.2. *Extension to Multiple Cells.*

In a multicell scenario, signals from multiple links superimpose at the receiver. The total interference at data receiver $\mathbf{x}$ amounts to

$$I^{\mathrm{d}}_{\mathbf{x}}[n,k] = \sum_{\substack{\mathbf{z}\in\mathcal{T} \\ \mathbf{z}\neq\mathbf{x}}} T^{\mathrm{d}}_{\mathbf{z}}[n,k] \cdot G_{\mathbf{zx}}[n,k], \tag{9}$$

where $\mathcal{T}$ is the set of simultaneously active transmitters. Likewise, the received BB at the data transmitter (BB receiver) $\mathbf{y}$ yields

$$I^{\mathrm{b}}_{\mathbf{y}}[n,k] = \sum_{\substack{\mathbf{z}\in\mathcal{R} \\ \mathbf{z}\neq\mathbf{y}}} T^{\mathrm{b}}_{\mathbf{z}}[n,k] \cdot G_{\mathbf{zy}}[n,k], \tag{10}$$

where $\mathcal{R}$ is the set active receivers (BB transmitters).

Unlike the case when two links compete for resources, $I^{\mathrm{b}}_{\mathbf{y}}[n,k]$ is no longer equivalent to $I^{\mathrm{d}}_{\mathbf{x}}[n,k]$ in the threshold test (8). This is because in (9) the interference powers from multiple transmitters $\mathcal{T}$ add up. Consequently, the total CCI at data receiver $\mathbf{x}$ may exceed the tolerable threshold such that $I^{\mathrm{d}}_{\mathbf{x}}[n,k] > I_{\mathrm{th}}$, although the BB power (10) observed by the individual interferers $\mathbf{y} \in \mathcal{T}$ is below the threshold, $I^{\mathrm{b}}_{\mathbf{y}}[n,k] \leq I_{\mathrm{th}}$. On the other hand, in (10) the interfering BB powers from multiple simultaneously active receivers observed at $\mathbf{y} \in \mathcal{T}$ add up. It is, therefore, possible that $I^{\mathrm{b}}_{\mathbf{y}}[n,k] > I_{\mathrm{th}}$, so that link $\mathbf{y}$ is prohibited from accessing chunk (n,k), although its individual CCI contribution, $T^{\mathrm{d}}_{\mathbf{y}}[n,k] \cdot G_{\mathbf{yx}}[n,k]$ would be below I_{th}. Note that the former effect partly compensates the latter. Moreover, in many cases the interference is dominated by one strong interfering source. Therefore, the threshold test (8) provides a good approximation to the level of interference potentially caused to the active receivers.

4.3. Initial Access in Contention. Initial access of unreserved slots in BB-OFDMA is carried out in contention. During contention, two or more transmitters from adjacent cells may access chunk (n, k) simultaneously. As a result, one or several links may encounter a collision on chunk (n, k), where the SINR target is not met. To reduce the occurrence of simultaneously accessed chunks in contention, a p-persistent chunk allocation procedure is applied to BB-OFDMA, where chunk (n, k) in cell q is accessed with probability p. Denoting the outcome of the p-persistent chunk allocation with the binary random variable $\chi_q[n, k] \in \{0, 1\}$, the access probability yields $\mathrm{P}(\chi_q[n, k] = 1) = p$. The impact of p on the system performance is investigated in Section 6.1.

4.4. Decentralised Chunk Reservation with BB Signalling. The BB-OFDMA protocol enables a link $\mathbf{x} = (\nu, q)$ to contend for a chunk once it is ensured that the CCI caused to the coexisting links $\mathbf{y}$ in the neighbouring cells is below a given threshold (8). Prior to accessing chunk (n, k), transmitter $\mathbf{x} = (\nu, q)$ listens to the associated BB minislot. Whether a user ν within cell q may contend for chunk (n, k) in (4) is controlled by

$$\epsilon_{\nu,q}[n,k] = \begin{cases} 0, & I^{\mathrm{b}}_{\nu,q}[n,k] \leq I_{\mathrm{th}} \text{ and } \chi_q[n,k] = 1, \\ \infty, & \text{otherwise.} \end{cases} \tag{11}$$

Chunks, where $a_q[n, k] = \nu$ in (4), are allocated to user ν. Those chunks where the achieved SINR is above a required SINR target, $\gamma_{\nu,q}[n, k] \geq \Gamma$, are reserved by setting the reservation indicator $\beta_q[n, k] = \nu$ in (4), and are subsequently protected from CCI by BB broadcast. The BB broadcast from the intended data receiver is observed as a *surge* in the received BB power [14], which effectively notifies the transmitter that the data of chunk (n, k) has been correctly received. User ν then reserves chunk n in the next frame $k + 1$ by setting $b_q[n, k + 1] = \nu$ in (5). On the other hand, if the transmitter does not detect a BB surge, it is understood that the SINR target was not met due to high CCI. These chunks are released by a reset of the reservation indicator to $\beta_q[n, k] = 0$ and setting $\epsilon_{\nu,q}[n, k] \to \infty$, so that chunk $(n, k + 1)$ may be assigned to other users.

4.5. Balancing System Throughput and Fairness. Cell-edge users are particularly affected by CCI for two reasons. First, the desired signal levels $R^{\mathrm{d}}_{\mathbf{x}}[n, k]$ are, on average, much weaker compared to users in close vicinity to the desired BS due to relatively low channel gains on their intended links $G_{\mathbf{x}}[n, k]$. Second, cell-edge users suffer from high CCI in the downlink, or cause high CCI to the adjacent cells in the uplink.

By tuning the interference threshold I_{th} in (8), the amount of CCI $I^{\mathrm{d}}_{\mathbf{x}}[n, k]$ caused to the receiver of a preestablished and coexisting link $\mathbf{x} = (\nu, q)$ is adjusted. Lowering I_{th} enforces a larger exclusion region around a vulnerable receiver. This enables cell-edge users to meet their SINR target Γ with a greater likelihood. On the other hand, by augmenting I_{th}, the number of simultaneously served links increases, giving rise to an enhanced system throughput. However, the cell-edge users are less likely to maintain their SINR target as interference protection is gradually eliminated. The chunks are released where the SINR target is not met, which means that these chunks are no longer reserved. Since the cell-centre users are less exposed to CCI, the chunks released by the cell-edge users are likely to be reallocated to the cell-centre users. As the allocation of the resources is shifted from the cell-edge users towards the cell-centre users, fairness is compromised. Hence, by adjusting I_{th}, system throughput is traded off for fairness.

A common measure to quantify fairness is Jain's fairness index [22], defined by

$$F = \frac{\left| \sum_{\nu=1}^{U} | \mathcal{B}_{\nu,q} | \right|^2}{U \sum_{\nu=1}^{U} | \mathcal{B}_{\nu,q} |^2}, \tag{12}$$

where U is the number of users in a given cell q. The user throughput $|\mathcal{B}_{\nu,q}|$ accounts for the number of successfully transmitted/received bits by user ν, as defined in (5). A fairness index of $F = 1$ represents a perfectly fair system where all users achieve the same throughput. On the other extreme, a fairness index of $1/U$ represents an unfair system where one user is served while all other users starve. We note that the fairness definition (12) is a relative measure, which ignores the absolute achieved throughput per user. To this end, a good fairness index F may coincide with poor spectrum utilisation. For instance, a system where two users achieve 1 Mbps and 2 Mbps would result in a poorer fairness index than a system where both users achieve only 0.5 Mbps. When analysing fairness, the fairness definition (12) should therefore be considered jointly with user throughput results.

(1) Consequences for the Downlink. In the downlink, MSs at the cell edge are exposed to high CCI from transmitters in adjacent cells (see Figure 2(a)). Note that the CCI observed at a given cell (cell 1 in Figure 2(a)) is independent of the user distribution in adjacent cells (cell 2 in Figure 2(a)), assuming a constant transmit power $T^{\mathrm{d}}_{\mathbf{x}}[n, k]$. This implies that if BS_2 lies within the exclusion region of MS_1, resources reserved by MS_1 cannot be spatially reused by *any* of the links in cell 2. However, if I_{th} is increased such that BS_2 is located outside the exclusion region of MS_1, *all* links in cell 2 qualify for a spatial reuse of the resources reserved by MS_1. However, the SINR target at MS_1 is less likely to be met. Should the SINR target at MS_1 not be met, this would cause the chunk allocated to MS_1 to be released and reallocated to another user served by BS_1- possibly a user that is located closer to the the serving BS_1. Therefore, the cell-edge throughput would suffer.

(2) Consequences for the Uplink. In the uplink, the transmitters (MSs) are distributed uniformly over the coverage area of the BS (see Figure 2(b)). Unlike the downlink, the CCI at the tagged BS depends on which MS transmits in the adjacent cell. To this end, the CCI observed at BS_1 in Figure 2(b) depends on whether MS_2 or MS_3 transmits to BS_2. Suppose that in cell 2 both MS_2 and MS_3 contend with MS_1 in cell 1 for chunks (n, k) and (n', k). In case MS_2 and

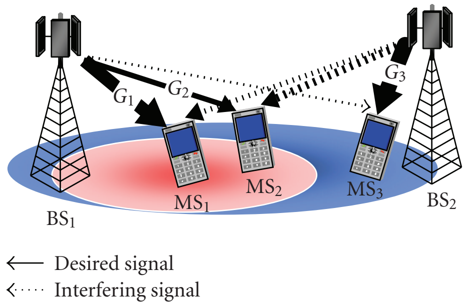


Figure 3: Busy burst with interference tolerance signalling (BB-ITS) in the downlink. The ovals represent the exclusion region formed with BB-ITS.

MS_1 simultaneously access chunk (n,k), while MS_3 and MS_1 simultaneously access chunk (n',k), the SINR at BS_1 tends to be superior on chunk (n',k) due to the lower CCI caused by MS_3. While MS_2 causes excessive CCI to BS_1, MS_1 and MS_3 may share chunk (n',k), although both users might be located near the cell boundary. Thus the uplink benefits from *interference diversity* due to the distributed location of mobile users. As a result, the degradation of performance at the cell edge at high I_{th} in uplink mode is less severe compared to the downlink.

4.6. Interference Tolerance Signalling via Busy Bursts. With fixed power BB signalling, the same level of interference protection is given to all links, disregarding the quality of the intended link. In case two receivers MS_1 and MS_2 with respective channel gains $G_1 > G_2$ are exposed to the same interference, as illustrated in Figure 3, the SINR target Γ is more likely met for MS_1 than for MS_2. By allowing MS_1 and MS_2 to transmit a BB with variable power, the individual amount of interference that can be tolerated by MS_1 and MS_2 is signalled to candidate transmitters in adjacent cells. Exclusion regions of different size are effectively formed around MS_1 and MS_2, as illustrated in Figure 3.

For busy burst with interference tolerance signalling (BB-ITS), the objective is that a given SINR target, $\gamma_{\mathbf{x}}[n,k] \geq \Gamma$, is maintained for an active receiver $\mathbf{x}$. This means that the maximum allowable interference depends on the intended link quality $R^{\mathrm{d}}_{\mathbf{x}}[n,k]$. Let $I^{\mathrm{tol}}_{\mathbf{x}}[n,k]$ denote the interference limit, for which the SINR (2) approaches $\gamma_{\mathbf{x}}[n,k] = \Gamma$. Then the tolerable interference at receiver $\mathbf{x}$ is upper bounded by

$$I^{\mathrm{d}}_{\mathbf{x}}[n,k] \leq I^{\mathrm{tol}}_{\mathbf{x}}[n,k] = \frac{R^{\mathrm{d}}_{\mathbf{x}}[n,k]}{\Gamma} - N. \tag{13}$$

Adjusting the tolerable interference level (13) implies that larger exclusion regions are formed for links with weak desired signal levels $R^{\mathrm{d}}_{\mathbf{x}}[n,k]$ and vice versa.

To signal the variable interference tolerance level $I^{\mathrm{tol}}_{\mathbf{x}}[n,k]$ of a victim receiver $\mathbf{x}$ to candidate transmitters $\mathbf{y}$ in adjacent cells, the BB transmit power $T^{\mathrm{b}}_{\mathbf{x}}[n,k]$ is adjusted, such that the simple threshold test $I^{\mathrm{b}}_{\mathbf{y}}[n,k] \leq I_{\mathrm{th}}$ in (8) is retained. Hence no additional information for channel sensing is required for BB-ITS. The received BB power approaches a fixed threshold, $I^{\mathrm{b}}_{\mathbf{y}}[n,k] = I_{\mathrm{th}}$, if the CCI approaches $I^{\mathrm{d}}_{\mathbf{x}}[n,k] = I^{\mathrm{tol}}_{\mathbf{x}}[n,k]$. Inserting $I^{\mathrm{d}}_{\mathbf{x}}[n,k] = I^{\mathrm{tol}}_{\mathbf{x}}[n,k]$ and $I^{\mathrm{b}}_{\mathbf{y}}[n,k] = I_{\mathrm{th}}$ into (6) yields the variable BB power $T^{\mathrm{b}}_{\mathbf{x}}[n,k] = T^{\mathrm{d}}_{\mathbf{y}}[n,k] \cdot I_{\mathrm{th}}/I^{\mathrm{tol}}_{\mathbf{x}}[n,k]$. Assuming that $T^{\mathrm{d}}_{\mathbf{y}}[n,k]$ is fixed and denoted by T^{d}, the BB transmit power is adjusted as follows [23]:

$$T^{\mathrm{b}}_{\mathbf{y}}[n,k] = \min\left(\frac{I_{\mathrm{th}} \cdot T^{\mathrm{d}}}{R^{\mathrm{d}}_{\mathbf{x}}[n,k]/\Gamma - N}, T^{\mathrm{b}}_{\max}\right), \tag{14}$$

where $T^{\mathrm{b}}_{\max}$ is the maximum BB transmit power. The min operator ensures that $T^{\mathrm{b}}_{\mathbf{x}}[n,k] \leq T^{\mathrm{b}}_{\max}$. Note that when $R^{\mathrm{d}}_{\mathbf{x}}[n,k]/\Gamma < N$, we get $\gamma_{\mathbf{x}}[n,k] < \Gamma$. In this situation, the chunk is released and no BB is transmitted. Therefore, it is ensured that $T^{\mathrm{b}}_{\mathbf{x}}[n,k]$ in (14) always has a positive value. We note that $I^{\mathrm{b}}_{\mathbf{x}}[n,k] = T^{\mathrm{b}}_{\mathbf{y}}[n,k] \cdot G_{\mathbf{xy}}[n,k]$ and $T^{\mathrm{b}}_{\max} = T^{\mathrm{d}}_{\mathbf{y}}[n,k] = T^{\mathrm{d}}_{\mathbf{x}}[n,k]$. It can be checked by plugging (14) into (8) that the threshold test (8) effectively checks if $I^{\mathrm{d}}_{\mathbf{y}}[n,k] \leq I^{\mathrm{tol}}_{\mathbf{y}}[n,k]$, regardless of the threshold used, as long as the BB transmit power is not clipped. In this paper, we choose $I_{\mathrm{th}} = -90\,\mathrm{dBm}$ because the probability of BB transmit power being clipped was found to be lower than 0.05 for the given deployment scenario with $\Gamma = 11.3\,\mathrm{dB}$ used. Furthermore, with this threshold, the received BB at the intended transmitter (the lower bound of which is approximated by $I_{\mathrm{th}} \cdot \Gamma$) remains well above the noise floor $-117.8\,\mathrm{dBm}$, such that it can be reliably detected.

4.7. Link Adaptation with BB Signalling. Receiver feedback based on BB-ITS (see Section 4.6) allows for receiver-driven link adaptation, where the chosen transmission rate is adapted to the instantaneous channel conditions. Let $\mathcal{M} = \{1,\ldots,M\}$ be the set of supported modulation schemes. Associated to each modulation scheme $m \in M$ is an SINR target $\Gamma = \Gamma_m$ that must be achieved to satisfy a given frame error rate (FER).

Provided that the channel response does not change between successive frames, changes in Γ_m may be signalled from receiver to transmitter through (14), since any fluctuation in received BB power $R^{\mathrm{b}}_{\mathbf{x}}[n,k] = G_{\mathbf{x}}[n,k]T^{\mathrm{b}}_{\mathbf{x}}[n,k]$ is due to a change of Γ_m in (14). In summary, BB-ITS serves two important objectives. First, by adjusting the SINR target Γ_m, the receiver implicitly signals to the transmitter through BB-ITS that the transmission format should be changed; second, by varying the BB power $T^{\mathrm{b}}_{\mathbf{x}}[n,k]$ in (14), the size of the exclusion region around the active receiver is adjusted, so that the required SINR target Γ_m is met in successive frames.

Link adaptation with BB-ITS is carried out in two phases: the *contention phase,* where the link is established and the *link adaptation* (LA) phase, where the receiver adjusts its transmission format to the current channel conditions.

Contention Phase. In contention, multiuser chunk allocation is carried out as described in Section 4.3. To contend for an unreserved chunk (n,k), transmitter $\mathbf{x} = (\nu,q)$ initially uses the modulation scheme with the lowest spectral efficiency $m_{\mathbf{x}}[n,k] = 1$. Chunks that satisfy $\gamma_{\mathbf{x}}[n,k] \geq \Gamma_1$ are reserved in the next frame $k+1$ by BB signalling (see Section 4.4), where the transmit power $T^{\mathrm{b}}_{\mathbf{x}}[n,k]$ in (14) is set using $\Gamma = \Gamma_1$. Then the transmission proceeds to the link adaptation phase.

Link Adaptation Phase. The objective of the link adaptation phase is to select the modulation scheme $m_{\mathbf{x}}[n,k] \in \mathcal{M}$ for chunk (n,k), which yields the highest spectral efficiency, for which $\gamma_{\mathbf{x}}[n,k] \geq \Gamma_{m_{\mathbf{x}}[n,k]}$ holds. By utilising BB-ITS link, adaptation is accomplished without any explicit feedback. The receiver executes the following algorithm.

(1) Calculate the achieved SINR $\gamma_{\mathbf{x}}[n,k]$ at chunk (n,k).

(2) Increment the number of bits per symbol based on $\gamma_{\mathbf{x}}[n,k]$

$$m_{\mathbf{x}}[n,k+1] = \begin{cases} m_{\mathbf{x}}[n,k]+1, & \gamma_{\mathbf{x}}[n,k] \geq \Gamma_{m_{\mathbf{x}}[n,k]+1}, \\ & m_{\mathbf{x}}[n,k] < M, \\ m_{\mathbf{x}}[n,k]-1, & \gamma_{\mathbf{x}}[n,k] < \Gamma_{m_{\mathbf{x}}[n,k],} \\ m_{\mathbf{x}}[n,k], & \text{otherwise}. \end{cases} \tag{15}$$

(3) If $m_{\mathbf{x}}[n,k+1] \geq 1$, adjust the BB power (14) using the SINR target $\Gamma = \Gamma_{m_{\mathbf{x}}[n,k+1]}$ and transmit the BB.

(4) If $m_{\mathbf{x}}[n,k+1] < 1$, terminate the link adaptation phase and return to the contention phase.

The transmitter senses the BB minislot associated to chunk (n,k). In order to determine the modulation scheme $m_{\mathbf{x}}[n,k{+}1]$ requested by the receiver, the transmitter executes the following algorithm.

(1) Measure the busy signal power received from the intended data receiver $R_{\mathbf{x}}^{\mathrm{b}}[n,k]$ and compute the difference to the BB power received from intended data receiver in the preceding slot, $\Delta R = R_{\mathbf{x}}^{\mathrm{b}}[n,k] - R_{\mathbf{x}}^{\mathrm{b}}[n,k-1]$.

(2) The modulation format is adjusted based on ΔR as follows:

$$\widehat{m}_{\mathbf{x}}[n,k+1] = \begin{cases} \widehat{m}_{\mathbf{x}}[n,k]+1, & \Delta R \geq I_{\mathrm{th}}\Delta\Gamma_m - \varepsilon, \\ \widehat{m}_{\mathbf{x}}[n,k]-1, & \Delta R < I_{\mathrm{th}}\Delta\Gamma_{m-1} + \varepsilon, \\ \widehat{m}_{\mathbf{x}}[n,k], & \text{otherwise}, \end{cases} \tag{16}$$

where $\Delta\Gamma_m = \Gamma_m - \Gamma_{m+1}$, $m = \widehat{m}_{\mathbf{x}}[n,k]$. The constant $\varepsilon > 0$ introduces a detection margin to enhance the robustness towards estimation errors in $\widehat{R}_{\mathbf{x}}^{\mathrm{b}}[n,k]$ due to channel variations and noise.

(3) If $\widehat{m}_{\mathbf{x}}[n,k+1] \geq 1$, transmit data on chunk $(n,k+1)$ using the new modulation scheme $\widehat{m}_{\mathbf{x}}[n,k+1]$.

(4) If $\widehat{m}_{\mathbf{x}}[n,k+1] < 1$, terminate the link adaptation phase and return to the contention phase.

Estimation errors due to channel variations and noise may cause detection errors, so that $\widehat{m}_{\mathbf{x}}[n,k] \neq m_{\mathbf{x}}[n,k]$. Mismatch between the selected modulation schemes at transmitter and receiver can be mitigated if the transmitter announces $\widehat{m}_{\mathbf{x}}[n,k]$ together with payload data on chunk (n,k).

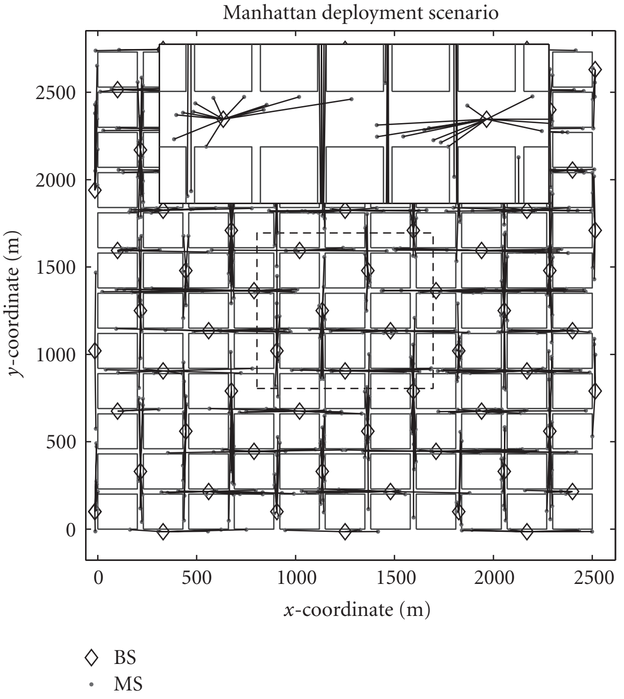


FIGURE 4: Manhattan grid urban microcell deployment.

4.8. Benchmark System. Full-frequency reuse with adaptive score-based chunk allocation (ASCA) is considered as the benchmark system. This means that neither chunk reservation nor interference avoidance mechanisms is in place. In order to maintain a fair comparison, the same fair scheduling algorithm (3) as in BB-OFDMA is applied. With ASCA, the score-based scheduler assigns chunk (n,k) to user ν whose score (3) is minimised

$$a_q[n,k] = \arg\min_{\nu} s_{\nu,q}[n,k]. \tag{17}$$

Chunk allocation for ASCA (17) corresponds to (4) by setting the reservation indicator to zero, $\beta_q[n,k] = 0$, and by allowing a cell to access all chunks, which is achieved by setting $\epsilon_{\nu,q}[n,k] = 0$ for all n,k in (3).

5. Manhattan Grid Deployment

An urban microcell deployment with a rectangular grid of streets (Manhattan grid) as defined in scenario B1 in WINNER [17] is considered, where antennas are mounted below the rooftop. The deployment scenario consists of building blocks of dimensions 200 m × 200 m, interlaced with the streets of width 30 m, forming a regular structure called the Manhattan grid, as shown in Figure 4. The network consists of 11 × 12 building blocks (72 BSs). However, the performance statistics are collected only over the central core of 3 × 3 building blocks (6 BSs), so as to reduce edge effects.

On average $U = 10$ MSs are served by one cell, uniformly distributed in the streets and moving with a constant velocity of 5 km/h. BSs are placed in the middle of the street canyons with an inter-BS distance of 4 building blocks, as depicted in Figure 4. Distance dependent path loss, log-normal shadowing, and frequency selective fading are taken into account, as specified in [24], channel model B1. While the effect of user mobility on the channel response due to the Doppler effect is taken into account, movement of users along the streets is not considered during the duration of one snapshot. Links where transmitter and receiver are located on the same street are modelled as line-of-sight (LoS) channels, with significantly lower path loss attenuation than nonline-of-sight (NLoS) links [24]. WINNER channel models B1-LOS and B1-NLOS [24] are used to model the LoS and NLoS channels, respectively. MSs are connected to the BS with the least path loss. A network synchronised in time and frequency is assumed.

The traffic in the system is modeled as a burst of 100 protocol data units (PDUs) whose interarrival time is exponentially distributed. A PDU of 112 bit is assumed, which is the smallest unit of data that can be transmitted in one chunk. The average offered load per user L_u is adjusted by the interburst duration. It is considered that the arrival times for different users are independent. The maximum number of chunks that a user can be assigned in a given slot is the total number of available chunks in a frame. The simulation parameters are summarised in Table 1.

A 3/4-rate convolutional code and the SINR targets Γ_m for a given modulation scheme m are selected to attain a packet error ratio of 10^{-2} per PDU, given in Table 2. For non-adaptive modulation, we consider a 16-QAM constellation with $m = 4$ and a corresponding SINR target of $\Gamma_4 = 11.3$ dB. For link adaptation, the modulation schemes $m \in \mathcal{M}$ are chosen based on the achieved SINR targets Γ_m.

Table 1: Simulation parameters.

Parameters	Value
Carrier centre frequency	3.95 GHz
System bandwidth B	89.84 MHz
No. of subcarriers (SCs)	1840
Subcarriers spacing Δf	48.8 kHz
OFDM symbols/frame $2n_{os}$	30
OFDM symbol duration T_{sym}	22.48 μs
Frame duration	0.6912 ms
No. of chunks/frame N_C	230
Chunk size $n_{sc} \times n_{os}$	8 (freq.) × 15 (time) = 120
PDU size	112 bits
Access probability p	0.3
No. of sectors/cell	1
No. of users/cell U	10
Tx power/chunk T^d	16.4 dBm
Antenna gain	0 dBi
Noise level/chunk N	−117.8 dBm
No. of snapshots	500
Snapshot duration	75 ms
User load L_u	30 Mbps

Table 2: Look up table for modulation scheme.

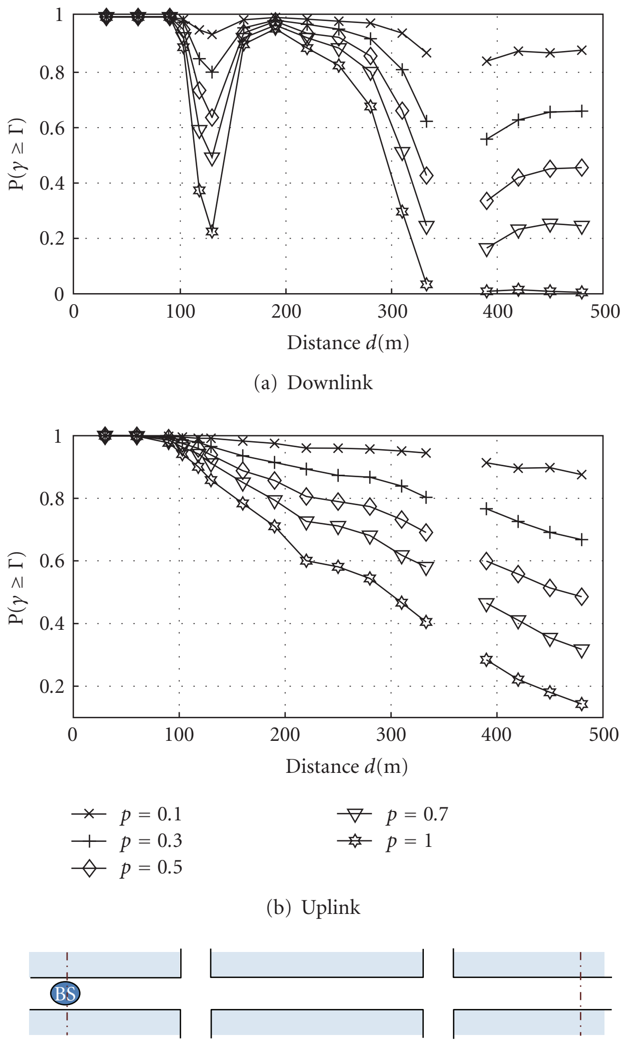
Modulation, No. of link PDUs per slot	Achieved SINR γ (dB)
No transmission $m = 0$	$-\infty < \gamma < 2.2$
BPSK $m = 1$	$2.2 \leq \gamma < 5.2$
QPSK $m = 2$	$5.2 \leq \gamma < 9.1$
Cross 8-QAM $m = 3$	$9.1 \leq \gamma < 11.3$
16-QAM $m = 4$	$11.3 \leq \gamma < 14.4$
Cross 32-QAM $m = 5$	$14.4 \leq \gamma < 16.6$
64-QAM $m = 6$	$16.6 \leq \gamma < 19.5$
Cross 128-QAM $m = 7$	$19.5 \leq \gamma < 22.5$
256-QAM $m = 8$	$22.5 \leq \gamma < \infty$

6. Results and Discussion

The performance of BB-OFDMA and the benchmark system (ASCA) are evaluated in terms of user and system throughput. User throughput is defined as the number of successfully received bits per user per unit time. A transmission is considered successful if the SINR target Γ is met at the receiver. The system throughput is defined as the aggregate throughput of all users per cell.

6.1. Collisions Based on Access Probability. The likelihood of achieving the SINR target during the initial access in contention is depicted in Figure 5 for $m = 4$ with $\Gamma_4 = 11.3$ dB, where m is the number of bits per symbol. The cell-edge region suffers from collisions (SINR target not met) both in the uplink (Figure 5(a)) and the downlink (Figure 5(b)). Decreasing the access probability p substantially reduces the occurrence of collisions, since the probability of simultaneous access of chunks in contention reduces (see Section 4.3). In the downlink, cell-edge users suffer from weaker desired signal power and at the same time experience strong CCI. Furthermore, the users located at the street crossings at $d = 115$ m are exposed to strong LoS interference from BSs in the perpendicular streets. In the uplink, however, these users cause CCI to the neighbouring cells; which may impact either users at the cell-edge or users closer to the intended BS. Consequently, the SINR target is met with less likelihood at street crossings and the cell edge in the downlink mode compared to the uplink mode.

6.2. Setting the Threshold for Fixed Power BB Signalling. The impact of the choice of interference threshold on the mean system throughput is shown in Figure 6 for fixed 16-QAM modulation with $m = 4$. It is seen that for lower values of I_{th}, the amount of allocated resources (Set $\mathcal{A}$) and the achieved throughput (Set $\mathcal{B}$) are approximately equal. This is because at low I_{th}, larger exclusion regions around active receivers are enforced. Thus, CCI is mitigated at the expense of spatial reuse. By increasing I_{th}, the system throughput gradually improves until the maximum is reached. However, increasing I_{th} introduces additional links that cause more CCI to the existing links. As a result, some of the links (mainly cell-edge users) are less likely to meet the SINR target. Although it is desirable to maximise the spectral



(a) Downlink

(b) Uplink

Figure 5: Probability of meeting the SINR target $\Gamma = 11.3$ dB in contention for different access probabilities p, as a function of the BS-MS distance d. At $d = 115$ m, links are exposed to strong LOS interference from cells in perpendicular streets, which causes collisions in the downlink, while at $d = 345$ m, the MSs are connected to BSs in a perpendicular street due to better channel gains.

efficiency, it may be necessary to maintain a fair distribution of resources to all users. Achieving a balance between maximising spectral efficiency and enhancing fairness is addressed in Section 6.3.

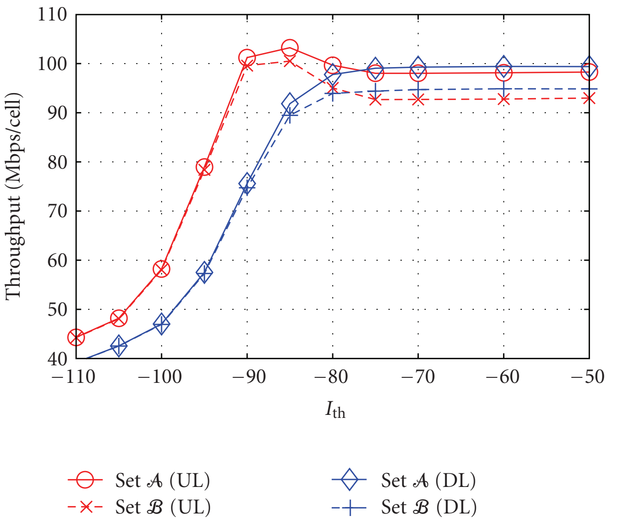
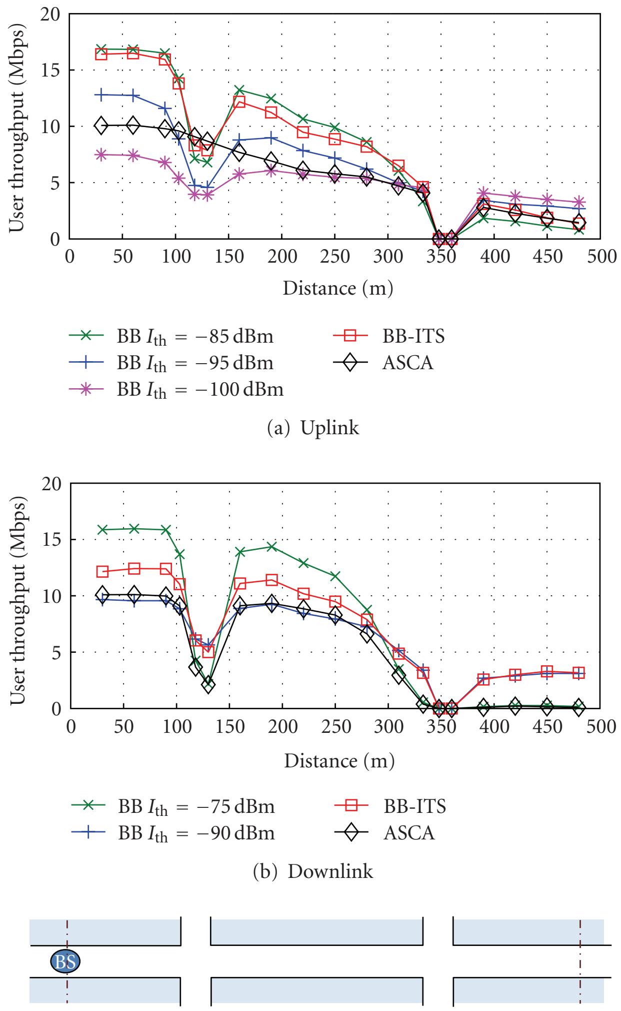


Figure 6: Mean system throughput versus I_{th} for BB-OFDMA with 16-QAM modulation using fixed BB transmit power. The mean system throughput is maximised for $I_{\text{th}} = -85$ dBm in the UL and $I_{\text{th}} = -75$ dBm in the DL.

6.3. Impact of Interference Threshold on Fairness. Figure 7 shows the average user throughput versus distance d from the serving BS. It is observed that the performance of BB-OFDMA is sensitive to the chosen threshold I_{th}. The system throughput is maximised for $I_{\text{th}} = -75$ dBm in the downlink and for −85 dBm in the uplink (see Figure 6). However, these thresholds severely affect cell-edge user throughput. Increasing interference protection by lowering I_{th} enhances user throughput at the cell edge at the expense of system throughput. In the uplink (Figure 7(a)), the cell edge throughput (measured at $d = 420$ m from the desired BS) improves from 1.5 Mbps ($I_{\text{th}} = -85$ dBm) to 3.08 Mbps ($I_{\text{th}} = -95$ dBm), an approximately onefold increase, whereas in the downlink (Figure 7(b)), user throughput increases from 267 kbps ($I_{\text{th}} = -75$ dBm) to 2.9 Mbps ($I_{\text{th}} = -90$ dBm), an approximately tenfold increase. At $d = 115$ m, MSs are exposed to LOS interference from BSs in perpendicular streets in the downlink. Consequently, high CCI compromises throughput for these users. In the uplink, MSs located at street crossings at $d = 115$ m transmit, so that these users are not exposed to LOS interference. Hence the uplink throughput of ASCA is not affected at $d = 115$ m. For BB-OFDMA, however, MSs located at street crossings are exposed to strong BB signals from BSs in perpendicular streets, which reduces the number of chunks such users can compete for, causing a drop of throughput for users located at street crossings.

Fairness is numerically quantified using Jain's fairness index (12). The cdf of the fairness distribution is presented in Figure 8(a) for the uplink and Figure 8(b) for the downlink. Applying the interference threshold that maximises system throughput, $I_{\text{th}} = -75$ dBm in the downlink and −85 dBm in the uplink, results in median fairness index of $F = 0.56$ and 0.66, respectively. Increasing the interference protection by lowering I_{th} improves fairness, as this enables cell-edge users to meet their SINR target. To this end, using $I_{\text{th}} = -95$ dBm in the uplink and −90 dBm in the downlink, approximately 22% of system throughput, is traded off for median fairness indices of $F \approx 0.72$. In the uplink, the median fairness index can be further improved to 0.78 by setting $I_{\text{th}} = -100$ dBm. However, the improved fairness significantly degrades system throughput (see Figure 6).

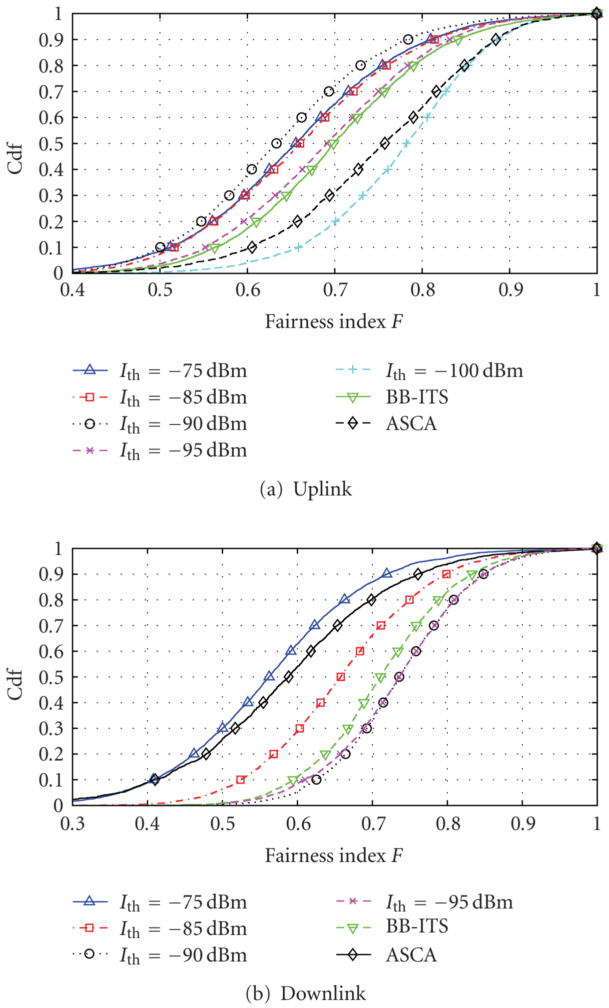
On the other hand, with BB-ITS, median fairness indices of ≈0.7 are achieved. The corresponding average uplink and downlink user throughputs at the cell edge amount to



(a) Uplink

(b) Downlink

FIGURE 7: Mean user throughput versus distance from the serving BS, d, for BB-OFDMA with 16-QAM modulation for different interference thresholds I_{th}. For comparison, results for full-frequency reuse without interference protection termed ASCA are also included. Note that at $d = 115$ m, links are exposed to strong LOS interference (data in downlink, BB in uplink) from cells in perpendicular streets, which compromises throughput, while at $d = 345$ m, the MSs are connected to the BS in a perpendicular street due to better channel gains.



(a) Uplink

(b) Downlink

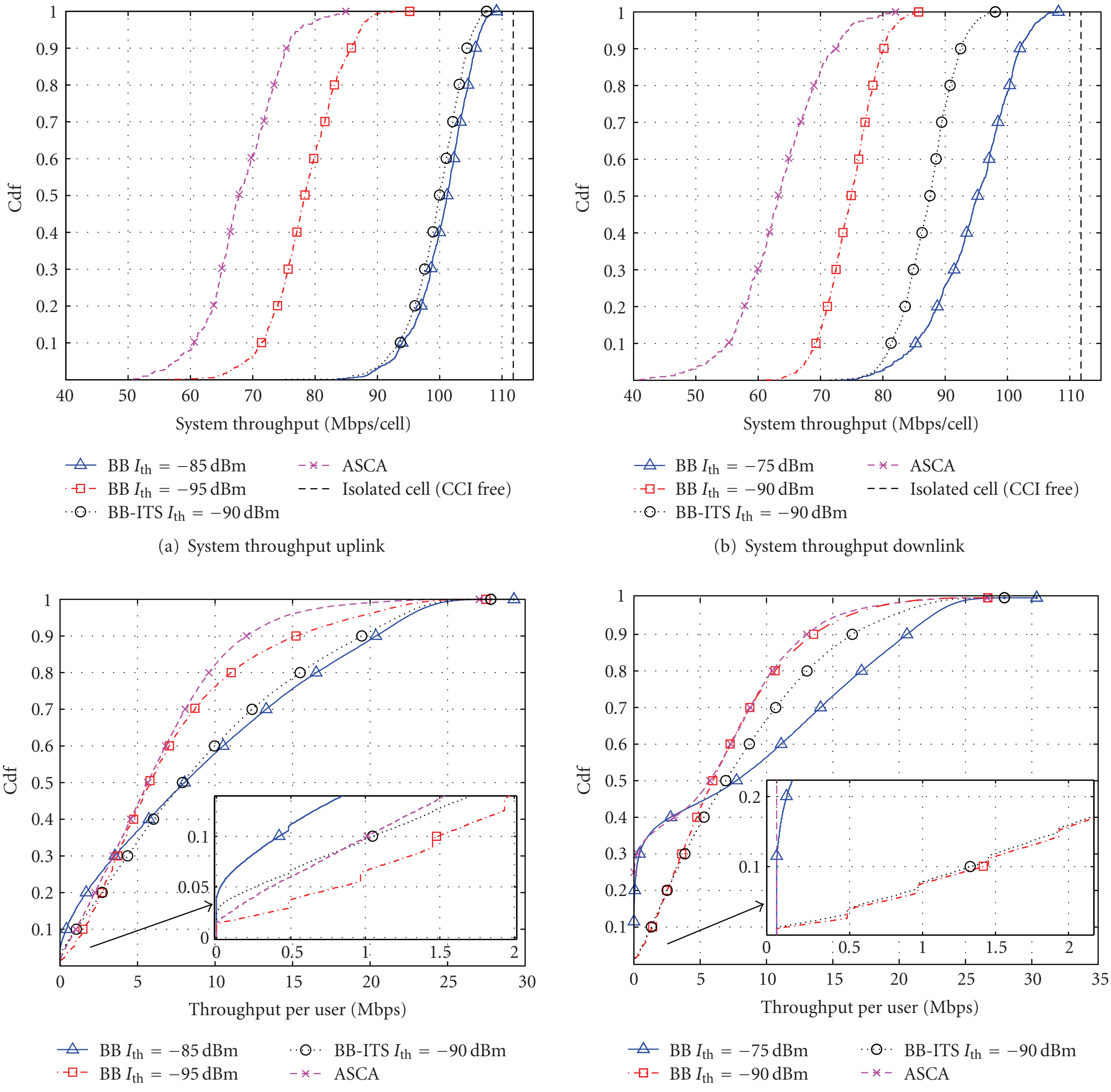
FIGURE 8: Cumulative distributive function (cdf) of Jain's fairness index (12) for BB-OFDMA compared to full-frequency reuse without interference avoidance (ASCA) both with 16-QAM modulation.

2.57 Mbps and 2.99 Mbps, respectively. The corresponding reduction in system throughput compared to the respective optimal thresholds with fixed power BB is only 1% in the uplink and 8% in the downlink. Note that BB-OFDMA with fixed BB power requires a 22% reduction in system throughput for a comparable performance at the cell edge. In light of this, BB-ITS results in a better tradeoff between system throughput and fairness.

For comparison, the median fairness resulting from ASCA is $F = 0.79$ in the uplink and 0.59 in the downlink. The corresponding average user throughputs at the cell edge are 2.278 Mbps and 208 kbps, respectively. This means that ASCA is more fair in the uplink compared to the downlink. The reason is that in the downlink cell-edge users are exposed to high CCI, while in the uplink cell-edge users cause high CCI to adjacent cells. Hence the detrimental effects of interference on the uplink tend to be more equally distributed among all users.

6.4. Comparison between BB-OFDMA and ASCA. Figures 9(a)–9(d) depict the cumulative distribution function (cdf) of the user throughput and the system throughput. The results shown in Figures 9(a)-9(b) demonstrate that BB-enabled interference avoidance exhibits a gain in median system throughput of up to 50% compared to ASCA, both in uplink and downlink. Using a modulation format of m = 4 bits per symbol and a 3/4-rate convolutional code, the upper bound on system throughput achieved in an isolated cell (CCI free system) is 111.8 Mbps. With $I_{th} = -85$ dBm in the uplink and −75 dBm in the downlink, a median system throughput of about 90% and 85% of the upper bound (CCI free system) is achieved.

Figures 9(c)-9(d) show the cdf of the user throughput for BB-OFDMA and ASCA. When fairness is the primary



(a) System throughput uplink

(b) System throughput downlink

(c) User throughput uplink

(d) User throughput downlink

Figure 9: Cumulative distributive function (cdf) of system throughput and user throughput for BB-OFDMA with fixed BB power and BB-ITS. The performance for full-frequency reuse without interference protection termed ASCA is included for comparison. BB-ITS results in a favourable tradeoff between fairness and system throughput both in uplink and downlink.

concern, I_{th} = −95 dBm in the uplink and I_{th} = −90 dBm in the downlink are preferable. Then the 10%-ile of the achieved user throughput amounts to 1.48 Mbps in the uplink (see Figure 9(c)) and 1.42 Mbps in the downlink (see Figure 9(d)). In contrast, ASCA fails to deliver any downlink throughput to more than 20% of the users. In the uplink, the 10%-ile of the user throughput of BB-OFDMA is improved by 40% compared to ASCA. With these uplink and downlink thresholds of I_{th} = −95 dBm and −90 dBm, the median system throughput of BB-OFDMA is still 15% and 18% higher than that achieved with ASCA (see Figures 9(a)-9(b)).

The results of BB-OFDMA with variable BB power, termed BB-ITS, are also included in Figures 9(a)–9(d). With BB-ITS, the lower 10%-ile of user throughput achieved is 1.04 Mbps in uplink and 1.416 Mbps in downlink (see Figures 9(c)-9(d)), at a modest degradation in system throughput (see Figures 9(a)-9(b)) compared to BB-OFDMA with fixed threshold that maximises the respective system throughput. BB-ITS, therefore, not only avoids the need for tuning the interference threshold so as to match a certain interference scenario (e.g., in uplink or downlink), but

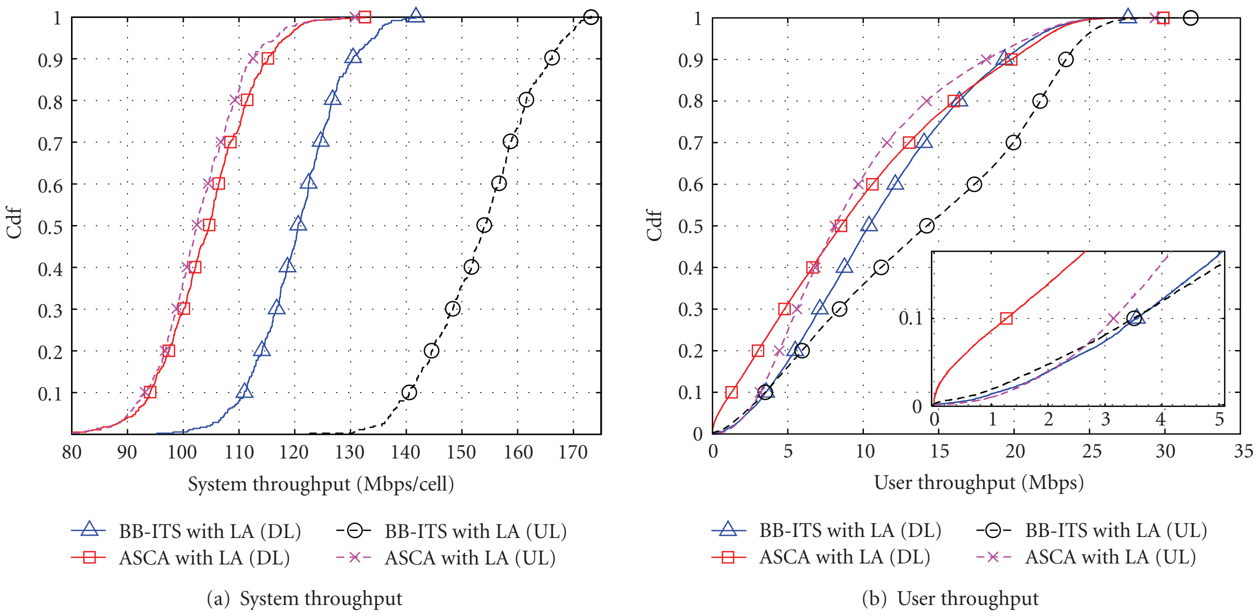


Figure 10: Cdfs of system and user throughputs for BB-ITS and ASCA with LA. In the DL, the users that are located at the cell-edge benefit whereas in the UL the users that are located closer to their desired BS benefit.

also achieves a preferable compromise between maximising system throughput and maintaining fairness.

6.5. Link Adaptation with BB-Signalling. Figures 10(a)-10(b) compare the system and user throughput achieved by performing link adaptation (LA) with BB-ITS and ASCA. Both BB-ITS and ASCA utilise the same link adaptation algorithm presented in Section 4.7; the only difference is that for ASCA interference protection is omitted. The results shown in Figure 10(a) reveal that BB-ITS with link adaptation attains an improvement of 50% (uplink) and 13% (downlink) in median system throughput compared to ASCA with link adaptation. Furthermore, Figure 10(b) shows that the BB-ITS outperforms ASCA by a factor of 2.75 in terms of the lower 10%-ile of the downlink user throughput. On the other hand, the cell-edge user throughput of BB-ITS and ASCA in the UL is comparable, while significant improvements of up to 70% are observed for higher percentiles of the user throughput in Figure 10(b).

By performing link adaptation with BB-ITS, the cell-edge users benefit in the downlink, whereas the users that are closer to their desired BS benefit in the uplink. The reason for this opposite trend for the uplink and the downlink is elaborated in the following. Due to the specific point-to-multipoint structure in the downlink, the CCI observed by the cell-edge users is dominated by the interference originating from the closest BS. When a chunk is assigned to a cell-edge user in the downlink, interference tolerance signalling enforces that this chunk cannot be spatially reused by the closest BS in an adjacent cell. By ensuring that, this dominant interferer does not access this chunk, the achieved SINR is greatly improved, potentially enough to meet the higher SINR target(s), thus allowing for the higher-order modulation schemes. In the uplink, on the other hand, the chunks assigned to the cell-edge users are more likely to be reused in the adjacent cells due to the distributed location of the MSs transmitters (see Section 4.5). Consequently, it is less likely that a more spectrally efficient modulation scheme can be used by the cell-edge users. Furthermore, in the uplink, the distance between the MSs (transmitters) and the victim BSs (receivers) in neighbouring cells is larger for the cell-centre MSs than the cell-edge users. Hence the cell-centre users are more likely to be located outside the exclusion range of BSs receivers (BB transmitters). This results in a larger number of chunks that are available to be spatially reused for the cell-centre users. Lastly, the cell-centre users also benefit from higher SINRs as a result of which throughput is particularly boosted by performing link adaptation.

7. Conclusions

In this paper, the busy signal concept for decentralised and self-organised interference aware medium access has been applied to OFDMA-TDD systems operated in Manhattan grid deployment scenarios. An exclusion zone around victim receivers is established by means of receiver feedback in the form of time-multiplexed busy bursts (BBs), wherein no active transmitter from an adjacent cell may be located. BB enabled interference avoidance exhibits impressive gains in system and user throughputs compared to the benchmark system, with full-frequency reuse without interference avoidance, both in the uplink and the downlink. The impact of the BB specific threshold parameter that controls the interference imposed on coexisting links in neighbouring cells has been studied.

By adjusting this threshold parameter, the system benefits from flexible operation of either achieving high system throughput or enhanced fairness in terms of cell-edge user throughput. A onefold (uplink) and tenfold (downlink) improvement in average cell-edge user throughput is achieved at a reduction in system throughput of about 22% (≈20 Mbps/cell) in both cases. BB-enabled interference avoidance is therefore particularly powerful in enhancing downlink cell-edge user throughput, since in the downlink high interference is coupled with low-desired signal levels, resulting in poor average SINRs at the cell edge. In the uplink, on the other hand, cell-edge users cause high CCI, so that the detrimental effects of uplink interference are distributed more equally among all users, giving rise to *interference diversity*.

By allowing each receiver to signal the amount of interference it can tolerate, by using a variable busy burst power, an even better tradeoff between system throughput and fairness is achieved. Especially in the downlink, a tenfold improvement has been achieved at the cost of only 8% reduction in maximum system throughput. Furthermore, this scheme also alleviates the need to adjust the BB threshold parameter. The latter property is particularly important for self-organising wireless networks, as the optimum choice of the BB threshold is sensitive to changes in the network topology, and may not be known *a priori*.

Finally, link adaptation has been combined with busy burst-enabled interference avoidance, where changes in the transmission format are implicitly signalled to the transmitter by virtue of a variable BB power. BB signalling with link adaptation attained a superior performance than the benchmark system with link adaptation, both in terms of system throughput and user throughput. Due to the particular interference scenario, the cell-edge users achieved larger gains in the downlink whereas the cell-centre users benefitted more in the uplink. Consequently, larger gains in system throughput in the uplink mode were achieved compared to the gains achieved in the downlink mode.

Acknowledgments

Initial parts of this work have been supported by DFG Grant HA 3570/1-2 within the program SPP-1163 (Techniken, Algorithmen und Konzepte für zukünftige COFDM Systeme-TakeOFDM) while some latter parts of this work have been performed within the framework of the IST Project IST-4-027756 WINNER, which is partly funded by the European Union. Harald Haas acknowledges the Scottish Funding Council support of his position within the Edinburgh Research Partnership in Engineering and Mathematics between the University of Edinburgh and Heriot Watt University. This work was presented in part at the IEEE International Symposium of Personal, Indoor and Mobile Radio Communications (PIMRC) 2008, Cannes, France.

References

[1] IEEE802.11a-1999, “Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications; High-Speed Physical Layer in the 5 GHz Band,” IEEE Standard Institution, Piscataway, NJ, USA, 1999.

[2] ETSI EN 300 744 v1.5.1 (2004-06), “Digital Video Broadcasting (DVB); Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television,” European Telecommunications Standards Institute (ETSI), June 2004.

[3] 3rd Generation Partnership Project (3GPP), “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation (Release 8),” Technical Specification Group Radio Access Network, 3GPP TS 36.211 V8.2.0, 3GPP March 2008.

[4] M. Sternad, T. Svensson, T. Ottosson, A. Ahlen, A. Svensson, and A. Brunstrom, “Towards Systems Beyond 3G Based on Adaptive OFDMA Transmission,” *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2432–2455, 2007.

[5] S.-E. Elayoubi, O. Ben Haddada, and B. Fourestie, “Performance evaluation of frequency planning schemes in OFDMA-based networks,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, part 1, pp. 1623–1633, 2008.

[6] M. C. Necker, “Local interference coordination in cellular OFDMA networks,” in *Proceedings of the 66th IEEE Vehicular Technology Conference (VTC ’07)*, pp. 1741–1746, Baltimore, Md, USA, September-October 2007.

[7] S. M. Heikkinen, H. Haas, and G. J. R. Povey, “Investigation of adjacent channel interference in UTRA-TDD system,” in *Proceedings of the IEE Colloquium on UMTS Terminals and Software Radio*, pp. 13/1–13/6, Glasgow, UK, April 1999.

[8] IST-4-027756 WINNER II, “D6.13.14 version 1.1 WINNER II System Concept Description,” January 2008, http://www.ist-winner.org/WINNER2-Deliverables/D6.13.14 v1 .1.pdf.

[9] Z. Bharucha and H. Haas, “Application of the TDD underlay concept to home nodeB scenario,” in *Proceedings of the 67th IEEE Vehicular Technology Conference (VTC ’08)*, pp. 56–60, Singapore, May 2008.

[10] F. A. Tobagi and L. Kleinrock, “Packet switching in radio channels—part II: the hidden terminal problem in carrier sense multiple-access and the busy-tone solution,” *IEEE Transactions on Communications*, vol. 23, no. 12, pp. 1417–1433, 1975.

[11] Z. J. Haas and J. Deng, “Dual busy tone multiple access (DBTMA)—a multiple access control scheme for ad hoc networks,” *IEEE Transactions on Communications*, vol. 50, no. 6, pp. 975–985, 2002.

[12] R. Zhao, B. Walke, and M. Einhaus, “Constructing efficient multi-hop mesh networks,” in *Proceedings of the 30th Anniversary IEEE Conference on Local Computer Networks (LCN ’05)*, pp. 166–173, Sydney, Australia, November 2005.

[13] P. E. Omiyi and H. Haas, “Improving time-slot allocation in 4th generation OFDM/TDMA TDD radio access networks with innovative channel-sensing,” in *Proceedings of the IEEE International Conference on Communications (ICC ’04)*, vol. 6, pp. 3133–3137, Paris, France, June 2004.

[14] P. Omiyi, H. Haas, and G. Auer, “Analysis of TDD cellular interference mitigation using busy-bursts,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 7, pp. 2721–2731, 2007.

[15] H. Haas, V. D. Nguyen, P. Omiyi, N. Nedev, and G. Auer, “Interference aware medium access in cellular OFDMA/TDD networks,” in *Proceedings of the IEEE International Conference on Communications (ICC ’06)*, vol. 4, pp. 1778–1783, Istanbul, Turkey, July 2006.

[16] B. Ghimire, H. Haas, and G. Auer, “Busy burst enabled interference avoidance in winner-TDD,” in *Proceedings of the 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC ’07)*, pp. 1–5, Athens, Greece, September 2007.

[17] IST-4-027756 WINNER II, "D6.13.7 v1.00, WINNER II Test Scenarios and Calibration Cases Issue 2," December 2006, http://www.ist-winner.org/WINNER2-Deliverables/D6.13.7.pdf.

[18] H. Haas and S. McLaughlin, Eds., *Next Generation Mobile Access Technologies: Implementing TDD*, Cambridge University Press, Cambridge, UK, 2008.

[19] T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 2001.

[20] W. Wang, T. Ottosson, M. Sternad, A. Ahlén, and A. Svensson, "Impact of multiuser diversity and channel variability on adaptive OFDM," in *Proceedings of the 58th IEEE Vehicular Technology Conference (VTC '03)*, vol. 1, pp. 547–551, Orlando, Fla, USA, October 2003.

[21] T. Bonald, "A score-based opportunistic scheduler for fading radio channels," in *Proceedings of the European Wireless Conference (EWC '04)*, pp. 283–292, Barcelona, Spain, February 2004.

[22] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Tech. Rep. TR-301, DEC, Maynard, Mass, USA, 1984.

[23] P. Agyapong, H. Haas, A. Tyrrell, and G. Auer, "Interference tolerance signaling using TDD busy tone concept," in *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC '07)*, pp. 2850–2854, Dublin, Ireland, April 2007.

[24] IST-4-027756 WINNER II, "D1.1.2 v1.2 WINNER II Channel Models," November 2007, http://www.ist-winner.org/WINNER2-Deliverables/D1.1.2v1 .1.pdf.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 726495, 14 pages
doi:10.1155/2009/726495

Research Article

A Fair Opportunistic Access Scheme for Multiuser OFDM Wireless Networks

Cédric Gueguen and Sébastien Baey

*Université Pierre et Marie Curie (UPMC) - Paris 6, Laboratoire d'Informatique de Paris 6 (LIP6),
104 avenue du Président Kennedy, 75016 Paris, France*

Correspondence should be addressed to Cédric Gueguen, cedric.gueguen@lip6.fr

Received 2 July 2008; Revised 13 December 2008; Accepted 1 February 2009

Recommended by Zhu Han

We propose a new access scheme for efficient support of multimedia services in OFDM wireless networks, both in the uplink and in the downlink. This scheme further increases the benefits of opportunistic scheduling by extending this cross-layer technique to higher layers. Access to the medium is granted based on a system of weights that dynamically accounts for both the experienced QoS and the transmission conditions. This new approach enables the full support of multimedia services with the adequate traffic and QoS differentiation while maximizing the system capacity and keeping a special attention on fairness. Performance evaluation shows that the proposed access technique outperforms existing wireless access schemes and demonstrates that choosing between high fairness and high system throughput is no more required.

1. Introduction

Providing mobile multimedia transmission services with an adequate QoS is very challenging. In contrast with wired communications, wireless transmissions are subject to many channel impairments such as path loss, shadowing, and multipath fading [1–4]. These phenomena severely affect the transmission capabilities and in turn the QoS experienced by applications, in terms of data integrity but also in terms of the supplementary delays or packet losses which appear when the effective bit rate at the physical layer is low. The past decades have witnessed intense research efforts on wireless digital communications. Among all the studied transmission techniques, IOrthogonal Frequency Division Multiplexing (OFDM) has clearly emerged for future broadband wireless multimedia networks (4G systems) and is already widely implemented in most recent wireless systems like 802.11a/g or 802.16. The basic principle of OFDM for fighting the effects of multipath propagation is to subdivide the available channel bandwidth in subfrequency bands of width inferior to the coherence bandwidth of the channel (inverse of the delay spread). The transmission of a high-speed signal on a broadband frequency selective channel is then substituted with the transmission on multiple subcarriers of slow speed signals which are very resistant to intersymbol interference and subject to flat fading. This subdivision of the overall bandwidth in multiple channels provides frequency diversity which added to time, and multiuser diversity may result in a very spectrally efficient system subject to an adequate scheduling.

MAC protocols currently used in wireless local area networks were originally and primarily designed in the wired local area network context. However, conventional access methods like Round Robin (RR) and Random Access (RA) are not well adapted to the wireless environment and provide poor throughput. Much interest has recently been given to the design of scheduling algorithms that maximize the performance of multiuser OFDM systems. Opportunistic scheduling techniques take advantage of multiuser diversity by preferably allocating the resources to the active mobile(s) with the most favourable channel conditions at a given time. This technique was explored first in single carrier communications [5]. More recently, opportunistic scheduling has been exploited in multicarrier systems [6, 7]. These schemes are derived from the maximum signal-to-noise ratio (MaxSNR), also known as maximum carrier

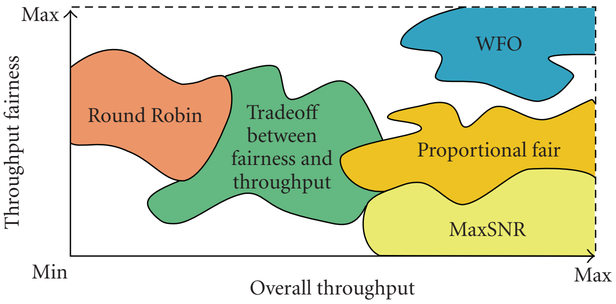


Figure 1: Tradeoff between overall throughput and fairness.

to interference ratio (MaxC/I), technique which allocates the resource at a given time to the active mobile with the greatest SNR. Dynamically adapting the modulation and coding allows then to always make the most efficient use of the radio resource and come closer to the Shannon limit. This maximizes the system capacity of an information theory point of view. However, it assumes that the user with the most favourable transmission conditions has information to transmit at the considered time instant. It does not take into account the variability of the traffic and the queuing aspects.

Pure opportunistic scheduling does not take into account the delay constraints of the flows to convey and suffer of a lack of fairness. References [8, 9] introduce opportunistic schemes coupled with a system of quota. This improves fairness but reduces the efficiency of utilization of the multiuser diversity with prejudice on system throughput. Proportional fair (PF) algorithms have recently been proposed to incorporate a certain level of fairness while keeping the benefits of multiuser diversity [10–14]. The basic principle is to allocate resources to a user, when its channel conditions are the most favourable with respect to its time average. In these schemes, fairness consists in guaranteeing an equal share of the total available bandwidth to each mobile, whatever its position or channel conditions.

However, performance analysis of PF-based protocols has shown that fairness issues persist since these algorithms do not ensure an equal throughput [15, 16]. The main issues are fairness considering mobiles with unequal spatial positioning, different traffic types, or different QoS targets. PF scheduling does not take into account the delay constraints and is not well adapted to multimedia services which introduce heterogeneous users, new traffic patterns with highly variable bit rates and stringent QoS requirements in terms of delay, and packet loss. Recently, [17] proposed the multimedia adaptive OFDM proportional fair (MAOPF) algorithm, an evolution of the classical PF that considers multimedia services. The principle of the MAOPF is to share the total available bandwidth among users proportionally to their bit rate requirement. Although this enables the coexistence of applications with unequal bit rates, heterogeneous QoS requirements are still not well supported. Moreover, the MAOPF allocates all OFDM subcarriers to the same mobile. This does not fully take advantage of the multiuser diversity and has a negative impact on the system capacity.

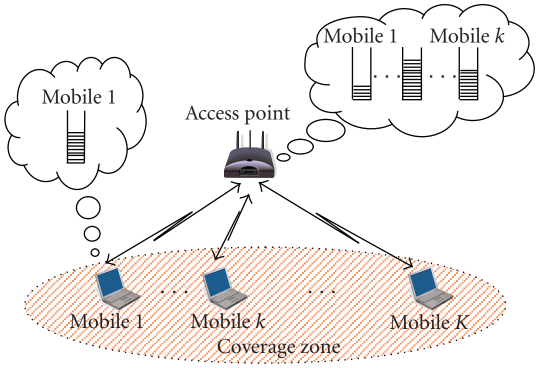


Figure 2: Allocation of radio resources among the set of mobiles situated in the coverage zone of an access point.

This paper proposes a new MAC protocol for efficient support of multimedia services in multiuser OFDM wireless networks. This protocol, which we call the “Weighted Fair Opportunistic (WFO)” protocol, applies cross-layer design concepts taking into account both the OFDM physical layer specificities (transmission conditions) and the higher layer constraints (traffic patterns, QoS constraints). Physical layer information are used in order to take advantage of the time, frequency, and multiuser diversity and maximize the system capacity. Higher layer information are exploited in a weighted system that introduces dynamic priorities between flows for ensuring the same QoS level to all mobiles. This result in an efficient scheme which guarantees the differentiated QoS constraints (data integrity and delay targets) of heterogeneous traffic flows like those generated by multimedia applications. Additionally, this bandwidth management avoids trading capacity for fairness as illustrated in Figure 1.

The paper is organized as follows. Section 2 provides a detailed description of the system under study. Section 3 introduces the QoS management principle embodied in the proposed protocol. Section 4 describes the integrated scheduling algorithm. In Section 5, we present a detailed performance evaluation through a simulation study. Section 6 concludes the paper.

2. System Description

We focus on the proper allocation of radio resources among the set of mobiles situated in the coverage zone of an access point (see Figure 2). We consider a centralized approach. The packets originating from the backhaul network are buffered in the access point which schedules the downlink transmissions. In the uplink, the mobiles signal their traffic backlog to the access point which builds the uplink resource mapping.

We assume that the physical layer is operated using the structure described in Figure 3 which ensures a good compatibility with the OFDM-based transmission mode of the IEEE 802.16-2004 [18, 19]. The total available bandwidth is divided in subfrequency bands or subcarriers. The radio resource is further divided in the time domain in frames.

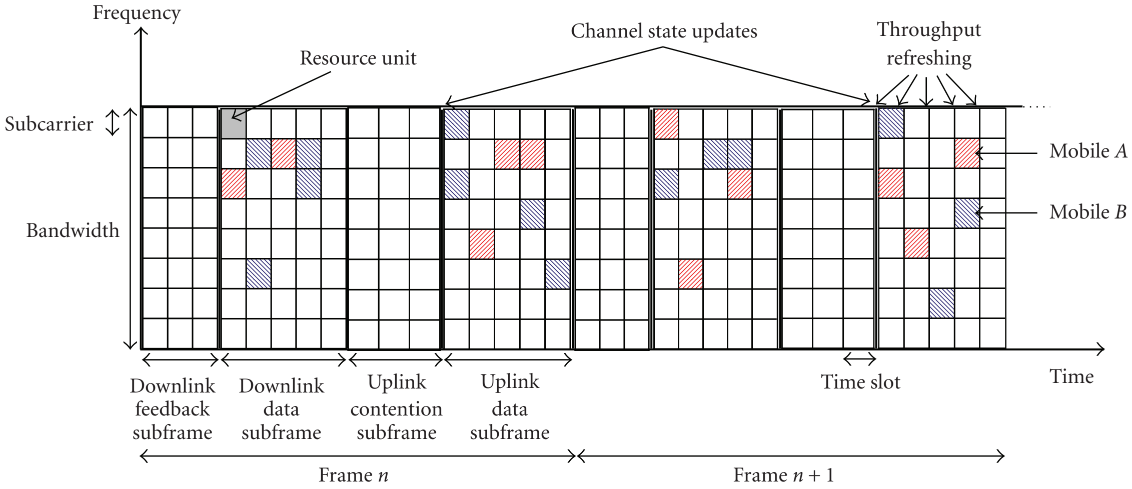


Figure 3: WFO frame structure in TDD mode.

Each frame is itself divided in time slots of constant duration. The time slot duration is an integer multiple of the OFDM symbol duration. The number of subcarriers is chosen so that the width of each subfrequency band is inferior to the coherence bandwidth of the channel. Moreover, the frame duration is fixed to a value much smaller than the coherence time (inverse of the Doppler spread) of the channel. With these assumptions, the transmission on each subcarrier is subject to flat fading with a channel state that can be considered static during each frame.

The elementary resource unit (RU) is defined as any (subcarrier, time slot) pair. Each of these RUs may be allocated to any mobile with a specific modulation order. Transmissions performed on different RUs by different mobiles have independent channel state variations [20]. On each RU, the modulation scheme is QAM with a modulation order adapted to the channel state between the access point and the mobile to which it is allocated. This provides the flexible resource allocation framework required for opportunistic scheduling.

The system is operated using time division duplexing with four subframes: the *downlink feedback subframe*, the *downlink data subframe*, the *uplink contention subframe*, and the *uplink data subframe*. The uplink and downlink data subframes are used for transmission of user data. In the downlink feedback subframe, the access point sends control information towards its mobiles. This control information is used for signalling to each mobile the RU(s) which have been allocated in the next uplink and downlink data subframes, the modulation order selected for each of these RUs and the recommended emission power in the uplink. In the uplink contention subframe, the active mobiles send their current traffic backlog and information elements such as QoS measures and transmit power. The uplink contention subframe is also used by the mobiles for establishing their connections. This frame structure supposes a perfect time and frequency synchronization between the mobiles and the access point as described in [21]. Therefore, each frame starts with a preamble used for synchronization purposes. Additional preambles may also be used in the frame.

3. The WFO Protocol QoS Management Principle

The crucial objective of the WFO protocol is to fully support multimedia transmission services, including the widest range of services: VoIP, videoconference, email, and file transfer. This requires the coexistence of delay sensitive flows as well as non-real-time traffic with looser delay constraints but with tight data integrity targets. In order to deal with the various and heterogeneous QoS requirements of multimedia services, the WFO protocol relies on a generic approach of QoS management.

We define a *service flow* as a traffic stream and its QoS profile, in a given transmission direction. A mobile may have multiple service flows both in the uplink and the downlink. An application may also use several service flows enabling for instance the implementation of Unequal Error Protection schemes in the physical layer. Each service flow possesses its own transmission buffer. In the following, index k is used to designate a given service flow among the set of service flows to be scheduled in a given transmission direction.

The QoS profile is defined as the set of parameters that characterizes the QoS requirements of a service flow mainly in terms of data integrity and delay. In the following, data integrity requirements are specified by a bit error rate (BER) target, which we denote by $BER_{target,k}$ for service flow k. Delay requirements are specified at the packet level. We assume traffic streams are organized at the MAC level in blocks of bits of constant size that we call packets. The packet delay is defined as the time between the arrival of the packet in the transmission buffer and the time of its reception by the mobile or the access point. This delay is roughly equal to the packet waiting time in the service flow transmission buffer neglecting the transmission and propagation delays.

Adequately specifying the delay requirements is challenging. We believe that the meaningful constraint is a limitation of the occurrences of large delay values. By analogy with the concept of outage used in system coverage planning, we define the concept of *delay outage*. A service flow k is in delay outage when its packets experience a delay greater than a given application specific threshold denoted T_k. We define the packet delay outage ratio ($PDOR_k$) experienced by each service flow k as the percentage of packets that do not meet the delay threshold T_k in the total number of packets transmitted. The experienced PDOR value is tracked all along the lifetime of the service flows; at each transmission of a packet of service flow k, the total number of packets whose delay exceeded the delay threshold T_k divided by the total number of packets transmitted since the beginning of the connection is computed. Additionally, we define the packet delay outage ratio target, denoted $PDOR_{target}$, as the maximum ratio of packets that may be delivered after the delay threshold. This characterizes the delay requirements of any service flow in a generic approach. Figure 4 illustrates an example cumulative distribution of the packet delay of service flow k at a given time instant. The objective of the WFO protocol is to regulate the experienced PDOR along the lifetime of the service flow such as its value stays below the PDOR target. This ensures the satisfaction of the delay requirements at a short-time scale.

In the WFO protocol, QoS management is organized in two parts: data integrity management and delay management. Data integrity is guaranteed by the physical layer mainly by adapting the modulation scheme and the transmit power to the mobile specific channel state. This is achieved considering each service flow independently. Delay management is performed considering all service flows jointly and scheduling the packets according to their distance to the PDOR target. Fairness is provided by guaranteeing the same level of satisfaction of delay constraints to all service flows, that is, guaranteeing the same PDOR to all service flows. The joint satisfaction of the delay constraints relies on the dynamics of the traffic streams that are multiplexed. Data integrity and delay management are integrated using the WFO scheduling algorithm.

4. The WFO Scheduling Algorithm

The core of the WFO protocol is its scheduling algorithm. This scheduling is performed during the uplink data transmission phase. The scheduler, located in the access node, grants RUs to each service flow as a function of

(i) its QoS profile (BER target, delay threshold, and PDOR target),

(ii) its currently experienced QoS (BER and PDOR),

(iii) its traffic backlog,

(iv) its channel state.

The QoS profile is signaled in the connection establishment phase. In the uplink, the currently experienced PDOR and the traffic backlog (buffer occupancy) are signaled by the mobile in the contention subframe. The experienced BER is tracked directly by the access node. Reciprocally, in the downlink, the currently experienced PDOR and the traffic backlog are calculated by the access node, and the experienced BER is signaled.

Additionally, knowledge of the channel state is supposed to be available at the receiver [22]. The current channel attenuation on each subcarrier and for each mobile is estimated by the access node based on the SNR of the signal sent by each mobile during the uplink contention subframe. Assuming that the channel state is stable on a scale of 50 milliseconds [23], and using a frame duration of 2 milliseconds, the mobiles will transmit their control information alternatively on each subcarrier so that the access node may refresh the channel state information once every 25 frames.

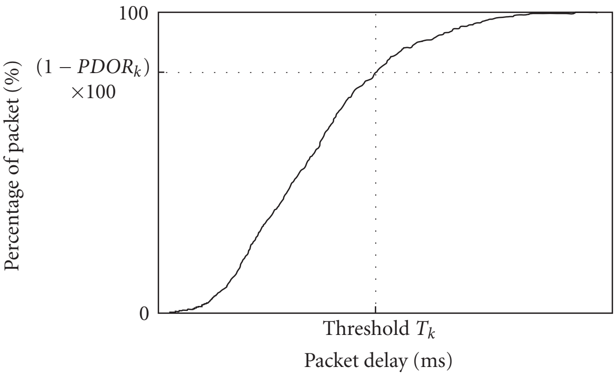


Figure 4: An example of packet delay CDF and experienced PDOR.

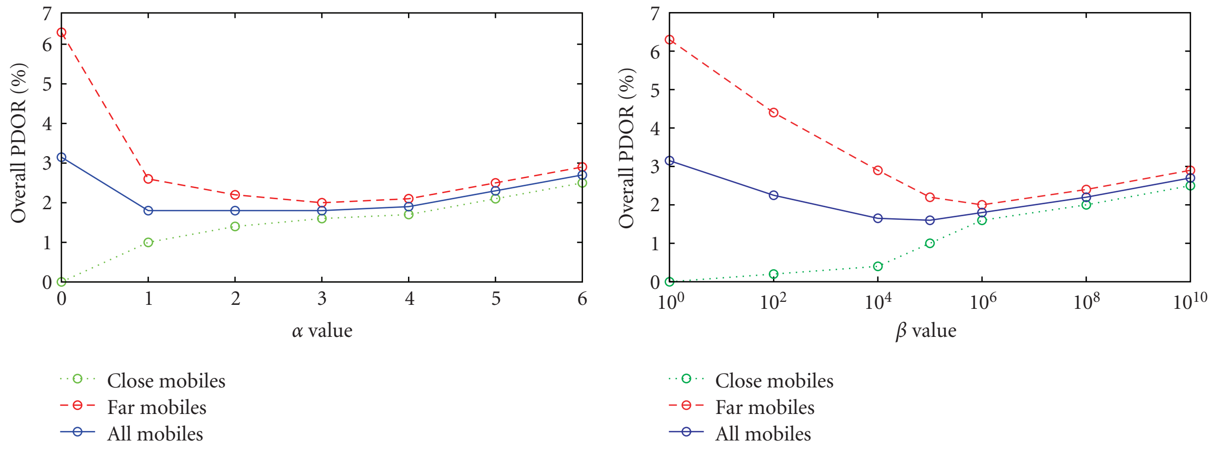
The WFO scheduling algorithm relies on weights that set the dynamic priorities for allocating the resource. These weights are built in order to satisfy two major objectives: system throughput maximization and fairness as explained below.

4.1. System Throughput Maximization. The WFO maximizes the system throughput in a MAC/PHY opportunistic approach. Data integrity requirements of the service flows are enforced considering each service flow independently adapting the modulation scheme and the transmit power to the mobile specific channel state. At each scheduling epoch, the scheduler computes the maximum number of bits $m_{k,n}$ that can be transmitted in a time slot of subcarrier n if assigned to service flow k, for all k and all n. This number of bits is limited by two main factors: the data integrity requirement and the supported modulation orders.

The bit error probability is upper bounded by the symbol error probability [6], and the time slot duration is assumed equal to the duration T_s of an OFDM symbol. The required received power $P_r(q,k)$ for transmitting q bits in an RU while keeping below the data integrity requirement $\mathrm{BER}_{\mathrm{target},k}$ of service flow k is a function of the modulation type, its order, and the single-sided power spectral density of noise N_0. For QAM and a modulation order M on a flat fading channel [1],

$$P_r(q,k) = \frac{2N_0}{3T_s}\left[\operatorname{erfc}^{-1}\left(\frac{\mathrm{BER}_{\mathrm{target},k}}{2}\right)\right]^2 (M-1), \quad (1)$$

where $M = 2^q$, and erfc is the complementary error function. $P_r(q,k)$ may also be determined in practice based on BER history and updated according to information collected on experienced BER.



(a) Exponent parameter calibration

(b) Normalization parameter calibration

Figure 5: α and β calibration.

The transmit power $P_{k,n}$ of service flow k on subcarrier n is upper bounded to a value $P_{\max}$ which complies with the transmit power spectral density regulation:

$$P_{k,n} \leq P_{\max}. \tag{2}$$

Given the channel gain $a_{k,n}$ experienced by service flow k on subcarrier n (including path loss and Rayleigh fading),

$$P_r(q,k) \leq a_{k,n} P_{\max}. \tag{3}$$

Hence, the maximum number of bits $q_{k,n}$ of service flow k which can be transmitted on a time slot of subcarrier n while keeping below its BER target is

$$q_{k,n} \leq \left\lfloor \log_2 \left(1 + \frac{3P_{\max} \times T_s \times a_{k,n}}{2N_0 \left[\operatorname{erfc}^{-1}\left(\mathrm{BER}_{\mathrm{target},k}/2\right)\right]^2} \right) \right\rfloor. \tag{4}$$

We further assume that the supported QAM modulation orders are limited such as q belongs to the set $S = \{0, 2, 4, \ldots, q_{\max}\}$. Hence, the maximum number of bits $m_{k,n}$ that will be transmitted on a time slot of subcarrier n if this RU is allocated to the service flow k is

$$m_{k,n} = \max\{q \in S,\ q \leq q_{k,n}\}. \tag{5}$$

MaxSNR-based schemes allocate the resources to the flows which have the greatest $m_{k,n}$ values. This bandwidth allocation strategy maximizes the bandwidth usage efficiency but suffers of a significant lack of fairness. In order to provide fairness while preserving the system throughput maximization, a new parameter is introduced which modulates this pure opportunistic resource allocation.

4.2. Fairness Support. The second major objective of the WFO is to provide fairness, that is, guaranteeing the same PDOR to all service flows as explained in Section 3. This is achieved by extending the above cross-layer design to higher layers. A new weighted fair (WF) parameter is introduced based on the current estimation of the PDOR of service flow k:

$$\mathrm{WF}_k = f(\mathrm{PDOR}_k), \tag{6}$$

where f is a strictly positive and monotonically increasing function. The WFO scheduling principle is then to allocate a time slot of subcarrier n to the mobile k which has the greatest WFO parameter value $\mathrm{WFO}_{k,n}$ with

$$\mathrm{WFO}_{k,n} = \mathrm{WF}_k \times m_{k,n}. \tag{7}$$

Based on the PDOR, the WF parameters directly account for the level of satisfaction of the delay constraints for an efficient QoS management. The PDOR is more relevant and simpler to use than the service flow throughput, the buffer occupancy, or the waiting time of each packet to schedule which would introduce a great complexity in the scheduling algorithm. The WFO parameters introduce dynamic priorities that delay the flows which currently easily respect their delay threshold to the benefit of others which go through a critical period.

Our studies on the algorithm performance have shown that a polynomial function f suits well

$$f(x) = 1 + \beta x^{\alpha}. \tag{8}$$

The exponent parameter α allows being more sensitive and reactive to PDOR fluctuations which guarantees fairness at a short-time scale. β is a normalization parameter that ensures that WF_k and $m_{k,n}$ are in the same order of magnitude. Given that $\mathrm{PDO}R_k$ has an order of magnitude 10^{-2}, β should be set to $10^{2\alpha}$. With this choice, WF_k is always in the same order of magnitude as $m_{k,n}$ and allows to manage both fairness and system throughput maximization.

By extensive simulations, we analyzed the influence of the value of the pair (α, β) on the performances of the WFO scheduling scheme and adequately tuned $f(x)$. Figures 5(a) and 5(b) illustrate the calibration study. Here, half mobiles are close to the access point and the second half, twice

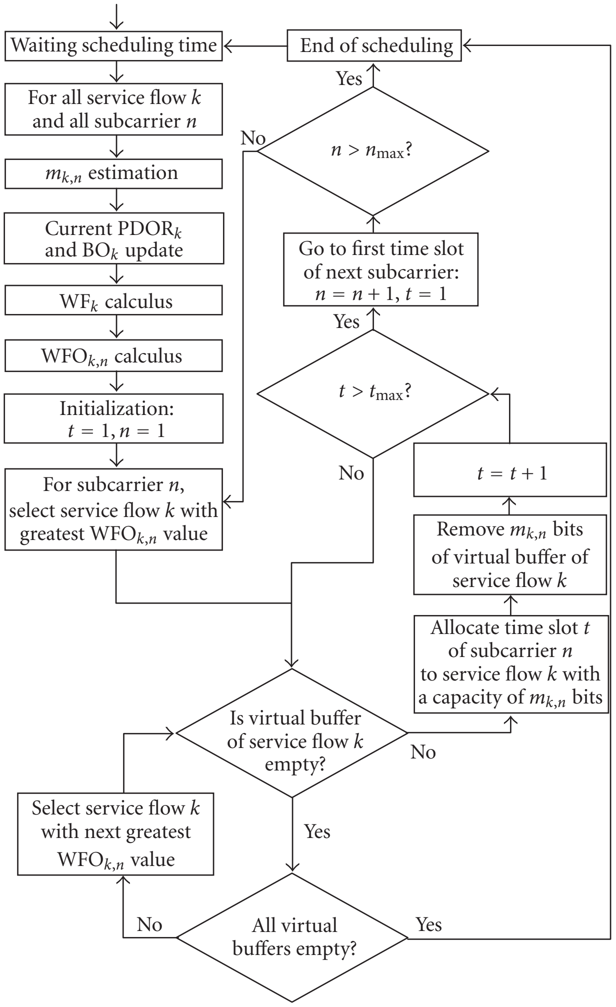

Waiting scheduling time
End of scheduling
Yes
For all service flow k and all subcarrier n
No
$n > n_{\max}$?
$m_{k,n}$ estimation
Current PDOR$_k$ and BO$_k$ update
Go to first time slot of next subcarrier: $n = n + 1$, $t = 1$
WF$_k$ calculus
Yes
WFO$_{k,n}$ calculus
$t > t_{\max}$?
Initialization: $t = 1$, $n = 1$
No
$t = t + 1$
For subcarrier n, select service flow k with greatest WFO$_{k,n}$ value
Remove $m_{k,n}$ bits of virtual buffer of service flow k
Allocate time slot t of subcarrier n to service flow k with a capacity of $m_{k,n}$ bits
Is virtual buffer of service flow k empty?
No
Select service flow k with next greatest WFO$_{k,n}$ value
Yes
No
All virtual buffers empty?
Yes

FIGURE 6: WFO scheduling algorithm flow chart.

other farther. All mobiles run a same application with same delay and BER requirements as described in Section 5.1. Figure 5(a) represents the overall PDOR (computed on all transmitted packets) obtained for different values of α coupled with a β value of $10^{2\alpha}$ as defined above. A cubic exponent suits well offering sufficient reactivity to PDOR fluctuations. Hence, in the following α is assumed to be equal to 3. Figure 5(b) shows the WFO performances obtained for each β value when α is set to 3. It confirms that when β is too small, the weighted parameter has no influence and fairness is lost. On the contrary, if β is too high, $m_{k,n}$ looses weight in the scheduling, and the system throughput maximization decrease. Good values for β range between 10^5 and 10^6. In the following, β is taken equal to 10^6.

Additionally, Figures 5(a) and 5(b) show the potential of the WFO. Indeed, when α or β equals zero, the function f is constant and $m_{k,n}$ only has influence in the scheduling. With this setting, the WFO behaves as the MaxSNR yielding unfair performances. In contrast, the adequate tuning of α and β brings the wanted fairness.

The dynamic priorities introduced by the WFO algorithm evolve as a function of the specific channel conditions and currently experienced QoS of each service flow in a cross-layer higher layers/MAC/PHY approach. This result in a well-balanced resource allocation which keeps a maximum number of service flows active across time but with continuously low traffic backlogs. Preserving this multiuser diversity allows to continuously take a maximal benefit of opportunistic scheduling and thus maximize the bandwidth usage efficiency. Additionally, this also achieves a time uniform fair allocation of the RUs to the service flows ensuring the required short term fairness [24, 25].

4.3. Global WFO Scheduling Algorithm Description. The WFO scheduling algorithm is detailed in Figure 6. The scheduling is run subcarrier by subcarrier and on a time slot basis for improved granularity. In the allocation process of a given time slot, the priority of a service flow with respect to another is determined by the magnitude of its WFO parameter. All service flows are scheduled simultaneously in a single run of the algorithm, whatever their QoS profile is. QoS differentiation is achieved by means of the WFO parameters. Service flows with low delay constraints like best effort traffic are qualified with a quite high delay threshold. As a result, their PDOR is always very small compared to other low latency traffic whose priority increases dramatically as soon their smaller delay threshold is not respected. In the following, we describe the proposed scheduling algorithm step by step.

Step 1. The scheduler refreshes the current PDOR$_k$ and buffer occupancy BO$_k$ values of each service flow k and computes the $m_{k,n}$, WF$_k$, and WFO$_{k,n}$ parameters for each service flow and each subcarrier. Then, n and t are initialized to 1.

Step 2. For subcarrier n, the scheduler selects the service flow k with the greatest WFO$_{k,n}$ value.

Substep 2.1. If the virtual buffer occupancy (we define the virtual buffer occupancy as the current buffer occupancy of service flow k minus the number of bits already allocated to this service flow) of service flow k is positive, the schedulers go to Substep 2.2. Else, if all virtual buffers are null or negative, the scheduler goes to Step 3. Otherwise, the scheduler selects the next service flow k with the greatest WFO$_{k,n}$ value and restarts Substep 2.1.

Substep 2.2. The scheduler allocates time slot t of subcarrier n to service flow k with a capacity $m_{k,n}$ bits, removes $m_{k,n}$ bits of its virtual buffer, and increments the value of t. If t is smaller than the maximum number $t_{\max}$ of time slots by subcarrier, go to Substep 2.1 for allocating the next time slot. Else, go to next substep.

Substep 2.3. Increment the value of n. If n is smaller than the maximum number $n_{\max}$ of subcarriers, go to Step 2 for allocating the time slots of the next subcarrier. Otherwise, go to Step 3.

Step 3. All virtual buffers are empty; or all time slots of all subcarriers are allocated and the scheduling ends.

5. Performance Evaluation

In this section, we compare the proposed weighted fair opportunistic scheduling with the Round Robin (RR), MaxSNR, PF, and MAOPF schemes implemented with subcarrier by subcarrier allocation. Performance evaluation results are obtained using OPNET discrete event simulations.

In the simulations, we assume 128 subcarriers and 5 time slots in a frame. The channel gain model on each subcarrier considers free space path loss and multipath Rayleigh fading [4]. We introduce a reference distance d_{ref} for which the free space attenuation equals a_{ref}. As a result the channel gain is given by

$$a_{k,n} = a_{\text{ref}} \times \left(\frac{d_{\text{ref}}}{d_k}\right)^{3.5} \times \alpha_{k,n}^2, \tag{9}$$

where d_k is the distance to the access point of the mobile owning the service flow k, and $\alpha_{k,n}^2$ represents the flat fading experienced by this service flow k if transmitted on subcarrier n. In the following, $\alpha_{k,n}$ is Rayleigh distributed with an expectancy equal to unity.

The maximum transmit power satisfies

$$10\log_{10}\left(\frac{P_{\max} T_s}{N_0} \times a_{\text{ref}}\right) = 31\ \text{dB}. \tag{10}$$

The BER target is taken equal to 10^{-3}. With this setting, the value of $m_{k,n}$ for the mobiles situated at the reference distance is 6 bits when $\alpha_{k,n}^2$ equals unity.

We assume all mobiles run the same videoconference application. This demanding type of application generates a high volume of data with high sporadicity and requires tight delay constraints which substantially complicates the task of the scheduler. Each mobile has only one service flow with a traffic composed of an MPEG-4 video stream [26] and an AMR voice stream [27].

The problem we are studying is quite different with the sum-rate maximization with water filling for instance. The purpose of the scheduler proposed in this paper is to maximize the traffic load that can be admitted in the wireless access network while fulfilling delay constraints. This is achieved by both taking into account the radio conditions but also the variations in the incoming traffic. In this context, we cannot for instance assume that each mobile has some traffic to send at each scheduling epoch. Traffic overload is not realistic in a wireless access network because it corresponds to situations where the excess traffic experiences an unbounded delay. This is why, in all our simulations, the traffic load (offered traffic) does not exceed the system capacity. In these conditions, the offered traffic is strictly equal to the traffic carried over the wireless interface and all mobiles get served sooner or later. The bit rate sent by each mobile is equal to its incoming traffic. Fairness in terms of bit rate sent by each mobile is rigorously achieved. The purpose of the scheduler is to dynamically assign the resource units to the mobiles at the best time in order to meet the traffic delay constraints. This is why we adopted the PDOR as a measure of the fairness in terms of QoS level obtained by each mobile.

Table 1: First scenario setup.

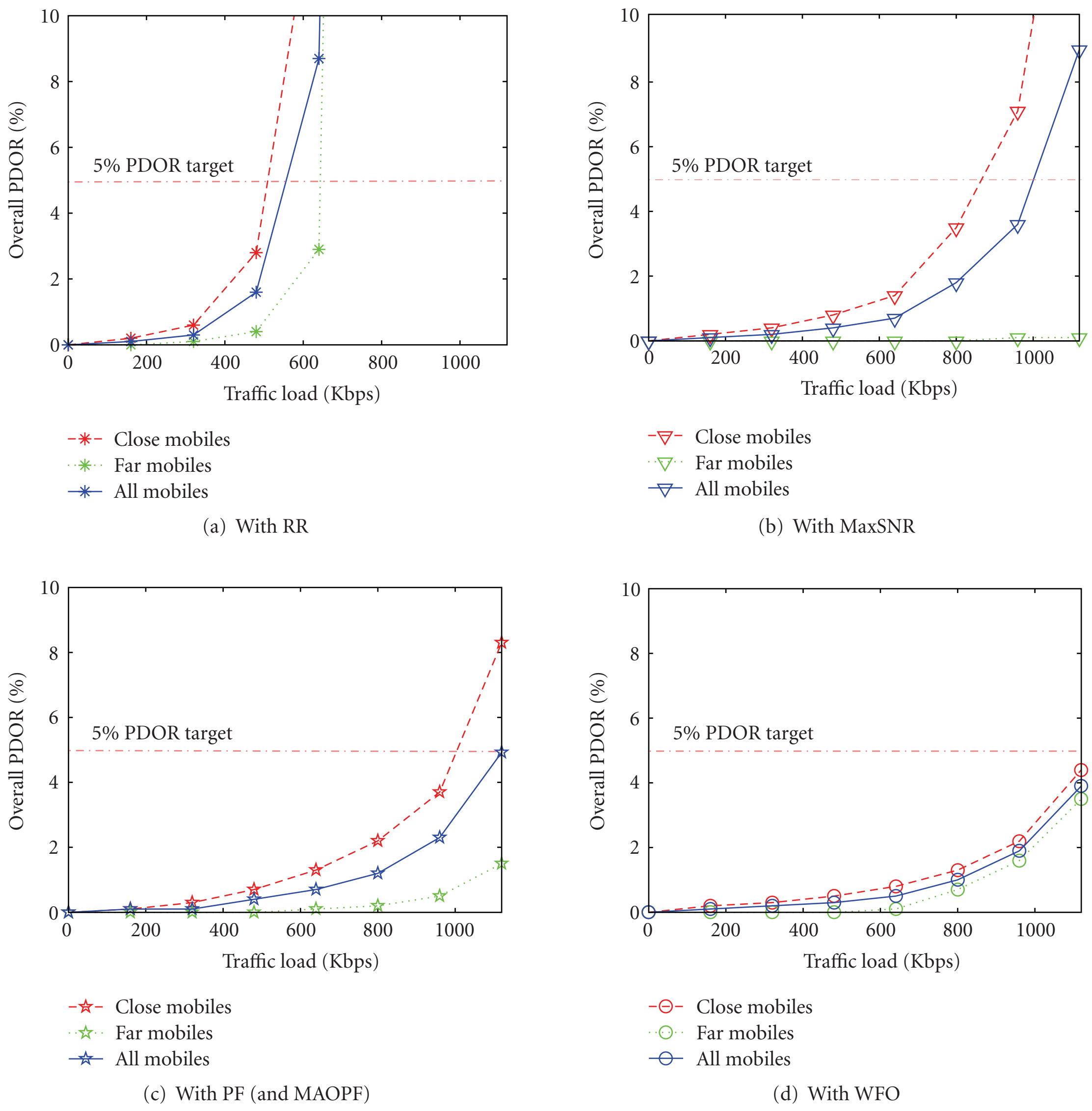
Group	Distance d_k	Delay threshold T_k	Data rate
1	2 d_{ref}	80 ms	80 Kbps
2	3 d_{ref}	80 ms	80 Kbps

Four simulation scenarii were used in the performance evaluation. In the first scenario, we analyzed the behavior of the schedulers when mobiles occupy different geographical positions. The second scenario examines the performance of the schedulers when mobiles have heterogeneous bit rate requirements. QoS differentiation is evaluated in the third scenario. The fourth simulation scenario considers mobiles with both heterogeneous geographical positions, bit rate, and QoS requirements.

5.1. First Scenario: Influence of the Distance on the Schedulers Performances. In wireless networks, it is well known that the closest mobiles to the access point generally obtain better QoS than mobiles more distant thanks to their higher spectral efficiency. In order to study the influence of the distance on the scheduling performances, a first half of mobiles are situated close to the access point and a second half 1.5 farther. The other parameters are identical for all the mobiles as described in Table 1. The total number of mobiles sets the traffic load.

First we focus on the fairness provided by each scheduler. Figures 7(a), 7(b), 7(c), and 7(d) display the overall PDOR for different traffic loads considering the influence of the distance on the scheduling. The classical RR fails to ensure the same PDOR to all mobiles. Actually, the RR fairly allocates the RUs to the mobiles without taking in consideration that far mobiles have a much lower spectral efficiency than closer ones. Moreover, the RR does not take benefit of multiuser diversity which results in a bad utilization of the bandwidth and in turn, poor system throughput. Consequently, an acceptable PDOR target of 5% is exceeded even with relatively low traffic loads. Based on opportunistic scheduling, the three other schemes globally show better QoS performances supporting a higher traffic load. However, MaxSNR, PF, and MAOPF still show severe fairness deficiencies (in this context where all mobiles have an equal source bit rate, the MAOPF and PF perform the same scheduling). Close mobiles easily respect their delay requirement while far mobiles experience much higher delays and go past the 5% PDOR target when the traffic load increases. In contrast, the WFO provides the same QoS level to all mobiles whatever their respective position. The WFO is the only one to guarantee a totally fair allocation. This allows to reach higher traffic loads with an acceptable PDOR for all mobiles. Additionally, looking at the overall PDOR for all mobiles at different traffic loads shows that, besides fairness, the WFO provides a better overall QoS level as well.

Observing the mean buffer occupancy in Figure 8(a), the WFO clearly limits the buffer occupancy to a same and reasonable value whatever the position of the mobile. This allows to stay under the PDOR target for any traffic load. With its system of weights, the WFO dynamically adjusts the



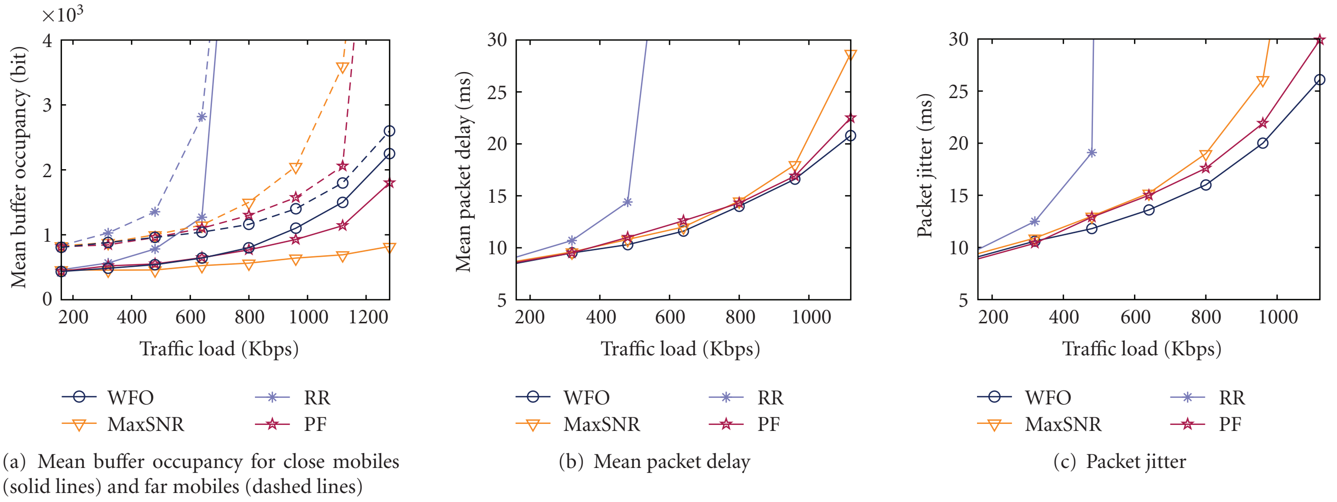
(a) With RR

(b) With MaxSNR

(c) With PF (and MAOPF)

(d) With WFO

Figure 7: Measured QoS with respect to distance.

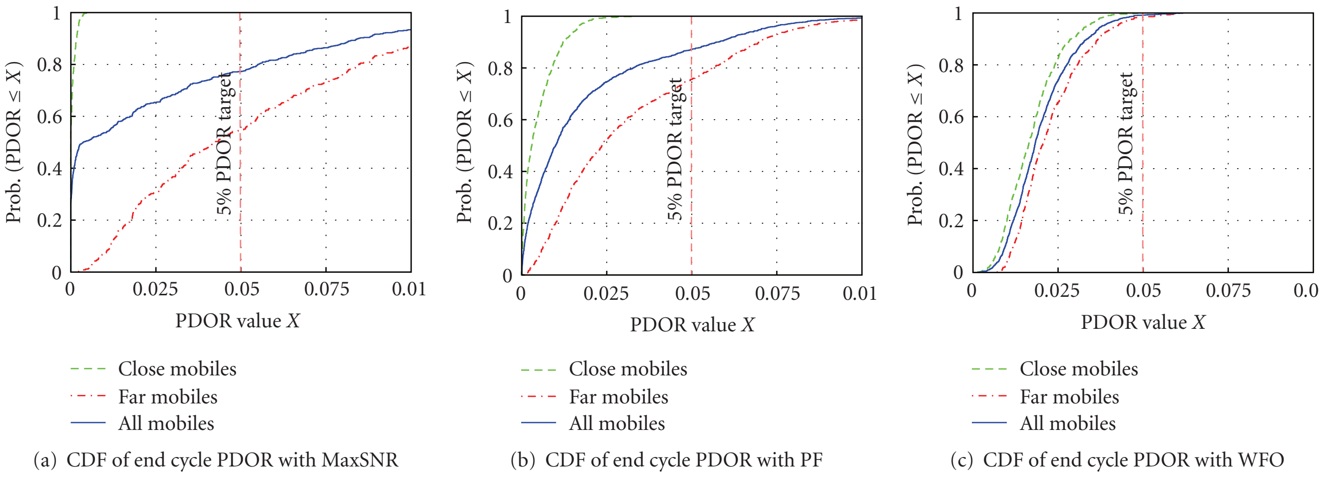


(a) Mean buffer occupancy for close mobiles (solid lines) and far mobiles (dashed lines)

(b) Mean packet delay

(c) Packet jitter

Figure 8: Buffer occupancy, delay and jitter.



(a) CDF of end cycle PDOR with MaxSNR

(b) CDF of end cycle PDOR with PF

(c) CDF of end cycle PDOR with WFO

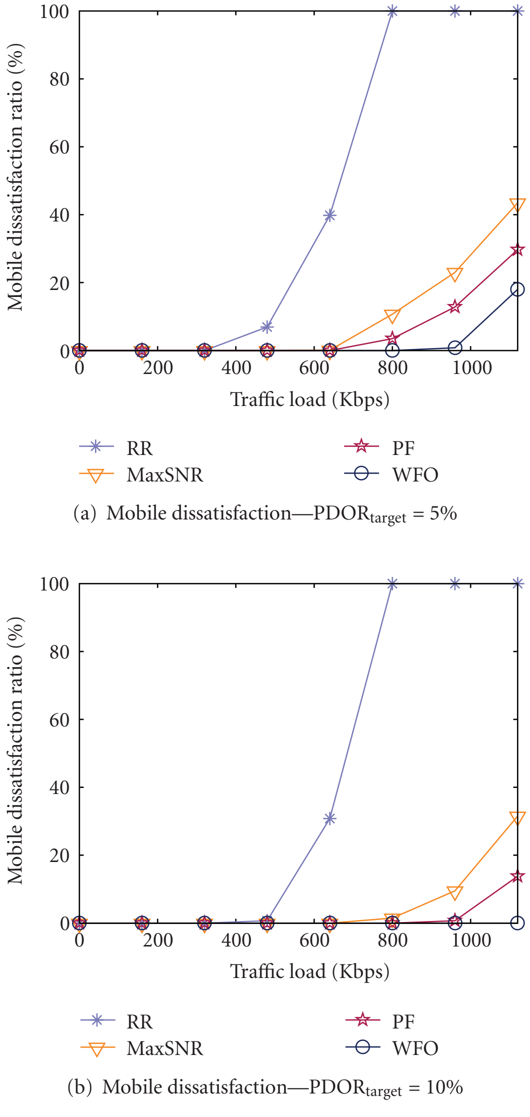
Figure 9: Perceived QoS with different allocation schemes.

relative priority of the flows according to their experienced delay. With this approach, sparingly delaying the closer mobiles, the WFO builds on the breathing space offered by the easy respect of the delay constraints of the closer mobiles (with better spectral efficiency) for helping the farther ones. The WFO interesting performance results are corroborated in Figures 8(b) and 8(c), where the overall values of the mean packet delay and jitter obtained using the WFO are smaller.

We then had a look at the QoS satisfaction level that each mobile perceives across the lifetime of a connection. We divided the connection of each mobile in cycles of five minutes and measured the PDOR at the end of each cycle. Figure 9 shows the CDF of end cycle PDOR values for a traffic load of 960 Kbps, using, respectively, the MaxSNR, the PF, and the WFO schemes (RR performances are not presented here since they are not able to support this high traffic load). We also estimated the mobile dissatisfaction ratio. We checked if at the end of each cycle the delay constraint is met or not. We then computed the mobile dissatisfaction ratio defined as the number of times that the mobiles are not satisfied (experienced PDOR≥ $\text{PDOR}_{\text{target}}$) divided by the total number of cycles (cf. Figure 10).

Highly unfair, MaxSNR fully satisfies the required QoS of close mobiles at the expense of the satisfaction of far mobiles. Indeed, only 54.5 percents of these latter experience a final PDOR inferior to a PDOR target of 5% (cf. Figure 9(a)). Unnecessary priorities are given to close mobiles which easily respect their QoS constraints while more attention should be given to the farther. These inadequate priority management dramatically increases the global mobile dissatisfaction which reaches 23% as shown in Figures 9(a) and 10(a).

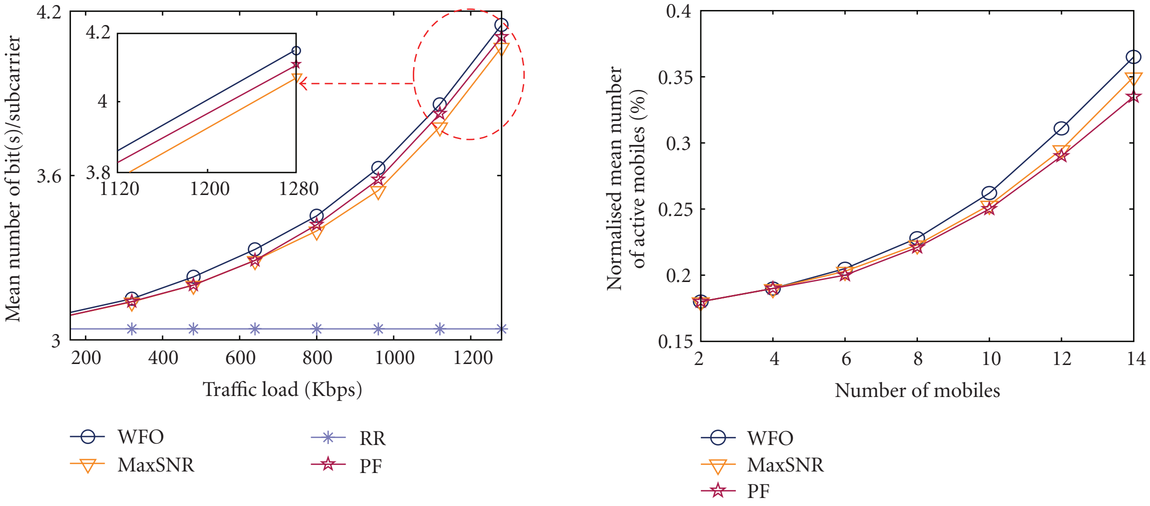
PF brings more fairness and allocates more priority to far mobiles. Compared to MaxSNR, PF offers a QoS support improvement with only 12.8% of dissatisfied mobiles (cf. Figures 9(b) and 10(a)). Fairness is still not total since the farther mobiles have a lower spectral efficiency than the closer ones due to path loss. All mobiles do not all benefit of an equal average throughput despite they all obtain an equal share of bandwidth. This induces heterogeneous delays



(a) Mobile dissatisfaction—$\text{PDOR}_{\text{target}}$ = 5%

(b) Mobile dissatisfaction—$\text{PDOR}_{\text{target}}$ = 10%

Figure 10: Analysis of the respect of QoS constraints for different targeted QoS.



(a) Spectral efficiency

(b) Multiuser diversity

Figure 11: Bandwidth usage efficiency.

and unequal QoS. This fairness improvement compared to MaxSNR indicates however that some flows can be slightly delayed to the benefit of others without significantly affecting their QoS.

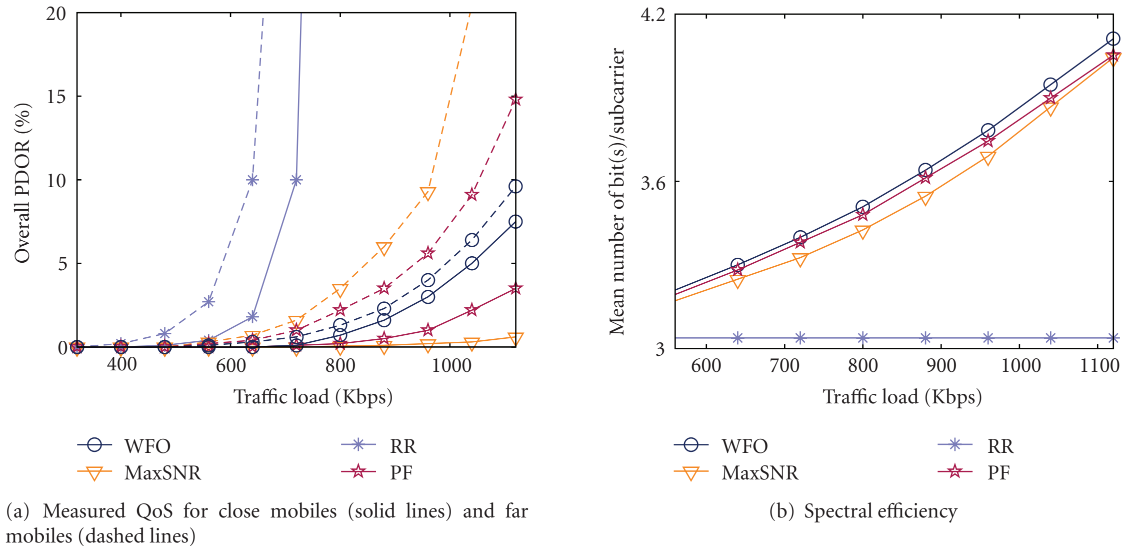
The WFO was built on this idea. The easy satisfaction of close mobiles (with better spectral efficiency) offers a degree of freedom which ideally should be exploited in order to help the farther ones. WFO allocates to each mobile the accurate share of bandwidth required for the satisfaction of its QoS constraints, whatever its position is. With WFO, only 0.8 percents of the mobiles are dissatisfied (cf. Figures 9(c) and 10(a)). Additionally, compared to Figures 9(a), 9(b), and 9(c) exhibits superimposed curves which proves the WFO high fairness, included at short term.

Figure 10 shows that the WFO brings the largest level of satisfaction. Indeed, for a tight PDOR target of 5% (see Figure 10(a)), the dissatisfaction ratio with a high traffic load of 1120 Kbps is equal to 18% with the WFO versus 29.7% with the best of the other scheduling schemes. If we set the PDOR target to 10%, the dissatisfaction ratio with a high traffic load of 1120 Kbps is 0% with the WFO versus 13.8% with the best of the other scheduling schemes (PF).

We finally studied the system capacity offered by the four scheduling algorithms. Figure 11(a) shows the average number of bits carried on a used subcarrier by each tested scheduler under various traffic loads. As expected, the nonopportunistic Round Robin scheduling provides a constant spectral efficiency, that is, an equal bit rate per subcarrier whatever the traffic load since it does not take advantage of the multiuser diversity. The three other tested schedulers show better results. In contrast with RR, with the opportunistic schedulers (MaxSNR, PF, WFO), we observe an interesting inflection of the spectral efficiency curve when the traffic load increases. The join analysis of Figures 11(a) and 11(b) shows that the spectral efficiency of opportunistic scheduling is an increasing function of the number of active mobiles, thanks to the exploitation of this supplementary multiuser diversity. Consequently, MaxSNR, PF, and WFO increase their spectral efficiency with the traffic load, and the system capacity is highly extended compared to networks which use classical scheduling algorithms. With these three schedulers, all mobiles are served even at the highest traffic load of 1280 Kbps.

The performance of the four schedulers can be further qualified by computing the theoretical maximal system throughput. Considering the Rayleigh distribution, it can be noticed that $\alpha_{k,n}^2$ is greater or equal to 8 with a probability of only 0.002. In these ideal situations, close mobiles can transmit/receive 6 bits per RU while far mobiles may transmit/receive 4 bits per RU. If the scheduler always allocated the RUs to the mobiles in these ideal situations, an overall efficiency of 5 bits per RU would be obtained which yields a theoretical maximal system throughput of 1600 Kbps. Comparing this value to the highest traffic load in Figure 11(a) (1280 Kbps) further demonstrates the good efficiency obtained with the opportunistic schedulers that nearly always serve the mobiles when their channel conditions are very good. This result also shows that the WFO scheduling has slightly better performances than the two other opportunistic schedulers. Keeping more mobiles active (cf. Figure 11(b)) but with a relatively lower traffic backlog (cf. Figure 8(a)), the WFO scheme preserves multiuser diversity and takes more advantage of it obtaining a slightly higher bit rate per subcarrier (cf. Figure 11(a)).

In the results described above, the traffic load was varied by increasing or decreasing the number of mobiles in the system, which modified the multiuser diversity. This exhibited the opportunistic behavior of the schedulers and especially their ability to take advantage of the multiuser diversity brought with the increase of the number of mobiles. We also studied below the ability of each scheduler to take profit of the multiuser diversity brought by a given number of users. In Figure 12, we provide complementary results obtained in a context where the traffic load variation is done



(a) Measured QoS for close mobiles (solid lines) and far mobiles (dashed lines)

(b) Spectral efficiency

Figure 12: Performances of schedulers with fixed multiuser diversity.

Table 2: Second scenario setup.

Group	Number of mobiles	Distance	Delay threshold	Data rate
1	9	1.6 d_{ref}	80 ms	80 Kbps
2	3	1.6 d_{ref}	80 ms	240 Kbps

Table 3: Third scenario setup.

Group	Number of mobiles	Distance	Delay threshold	Data rate
1	7	2.7 d_{ref}	80 ms	80 Kbps
2	7	2.7 d_{ref}	250 ms	80 Kbps

through just increasing the mobile bit rate requirement and keeping a constant number of users (10 mobiles). The results in Figure 12(a) show that, like above, the WFO outperforms the other scheduling schemes. With its weighted algorithms, the WFO dynamically adjusts the mobiles priority and ensures a completely fair allocation. WFO is the only one which allows to reach higher traffic loads with an acceptable PDOR for all mobiles. Additionally, even if the traffic load increases without variation in the number of mobiles, the WFO keeps more mobiles active across the time than the other schemes and takes better advantage of the multiuser diversity. The analysis of Figure 12(b). confirms that WFO maximizes the average bit rate per subcarrier.

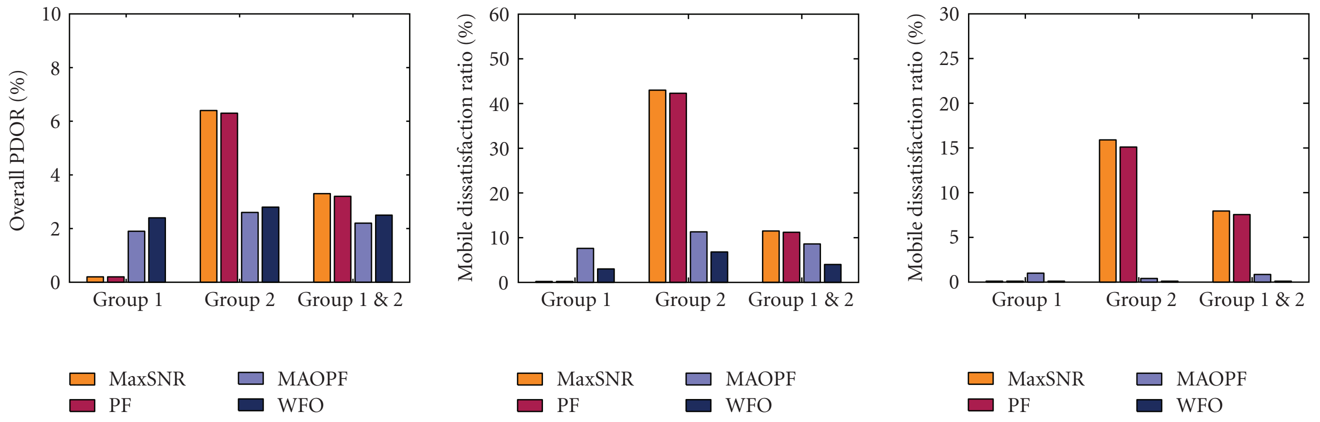
5.2. Second Scenario: Performance with Heterogeneous Bit Rate Sources. In this simulation scenario, mobiles are divided in two groups that differ only by their data rate as described in Table 2.

The four opportunistic scheduling strategies provide the same bandwidth usage ratio of 82% (RR performances are not reported here and in the following because its poor performances do not support the tested configurations). However, delay management considerably differs. Figure 13(a) shows the overall ratio of packets delivered after the threshold time, respectively, in Group 1, Group 2, and globally. The results show that the MaxSNR and the PF easily respect the delay constraints of low bit-rate mobiles but fail for the second group of mobiles. In contrast, the MAOPF and the WFO schemes provide fairness with an equal and moderate ratio of packets in delay outage whatever the source bit rate. The overall PDOR obtained with the MAOPF and the WFO is smaller than with the two other schemes. Here, the two multimedia oriented schedulers provide fair QoS management and better QoS support. Regarding the perceived QoS, Figures 13(b) and 13(c) show that the WFO outperforms the other schedulers including the MAOPF which do not directly manage the PDOR fluctuations.

5.3. Third Scenario: Performance with Heterogeneous Delay Constraints. We then studied the influence of heterogeneous delay requirements on the scheduling performances. In this simulation scenario, mobiles are divided in two groups that differ only by their delay requirements (cf. Table 3).

In this context where all mobiles have an equal source bit rate, the MAOPF and PF perform the same scheduling. Figure 14 clearly shows that the WFO outperforms the three other schemes ensuring fair QoS support and provides the largest QoS satisfaction level. This is processed with the WFO weighted system which dynamically controls the delay in a generic manner by monitoring the distance to the delay threshold thanks to a continuous and efficient regulation of the PDOR. This provides full QoS differentiation.

As explained above, the sum of incoming traffics of the mobiles is inferior to the system throughput. In this context, the traffic of each mobile is served sooner or later, and the bit rate sent by each mobile is equal to its incoming traffic. Fairness is absolute in terms of bit rate sent by each mobile. High-delay-sensitive mobiles are not served more often than

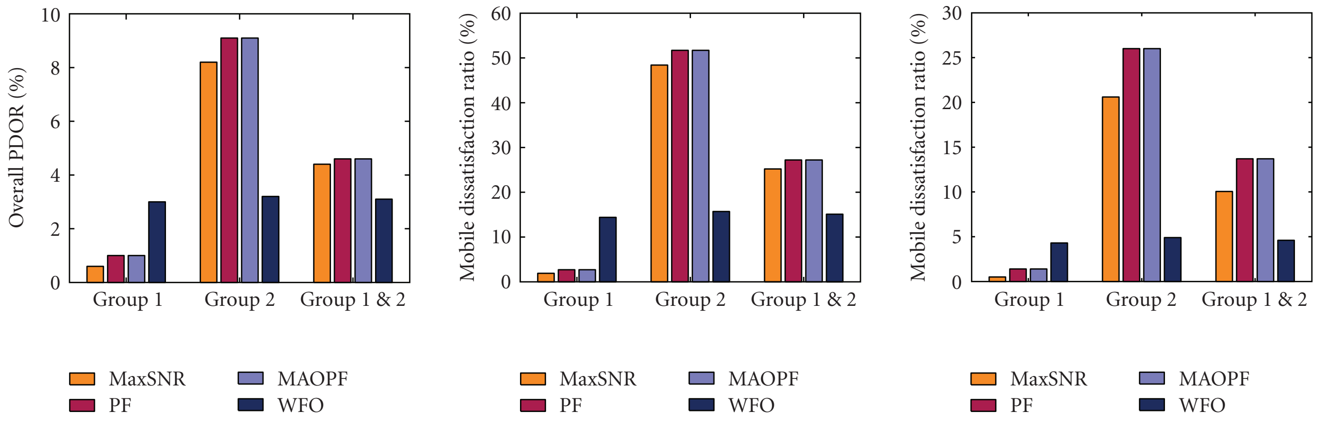


(a) Overall PDOR

(b) Mobile dissatisfaction—$\text{PDOR}_{\text{target}} = 5\%$

(c) Mobile dissatisfaction—$\text{PDOR}_{\text{target}} = 10\%$

Figure 13: Measured QoS with heterogeneous sources in terms of bit rate.



(a) Overall PDOR

(b) Mobile dissatisfaction—$\text{PDOR}_{\text{target}} = 5\%$

(c) Mobile dissatisfaction—$\text{PDOR}_{\text{target}} = 10\%$

Figure 14: Measured QoS with heterogeneous sources in terms of delay requirement.

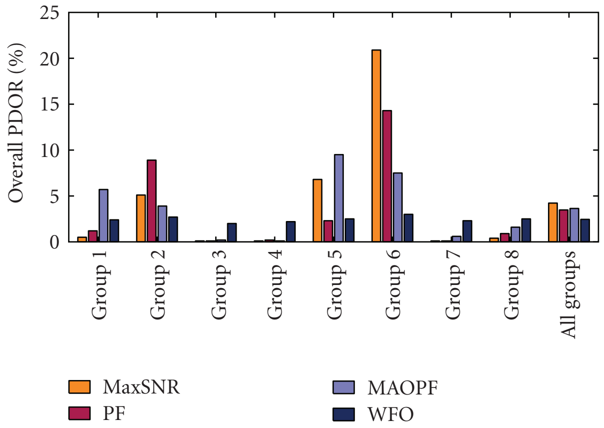


Figure 15: Overall PDOR.

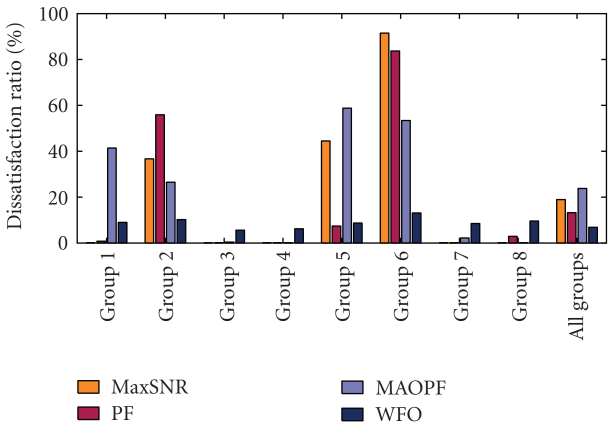


Figure 16: Mobile dissatisfaction when $\text{PDOR}_{\text{target}} = 5\%$.

other mobiles but earlier. It is only the time instant at which each high-delay-sensitive mobile and background mobile is served that differs. The purpose of the tested schedulers is to set dynamic priorities between the different types of traffics.

5.4. Fourth Scenario: Global Scheduling Performances Analysis. So far, we have analyzed the behavior of the schedulers in simple contexts considering one criterion at a time for better understanding its influence on the performances. In order to

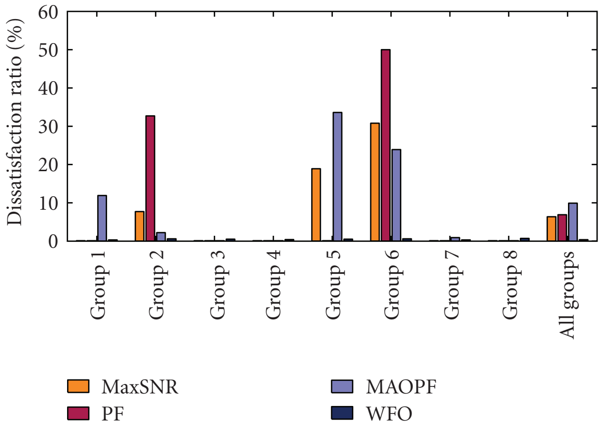


Figure 17: Mobile dissatisfaction when $\text{PDOR}_{\text{target}} = 10\%$.

Table 4: Fourth scenario setup.

Group	Number of mobiles	Distance	Delay threshold	Data rate
1	2	2 d_{ref}	80 ms	80 Kbps
2	1	2 d_{ref}	80 ms	160 Kbps
3	2	2 d_{ref}	250 ms	80 Kbps
4	1	2 d_{ref}	250 ms	160 Kbps
5	2	2.4 d_{ref}	80 ms	80 Kbps
6	1	2.4 d_{ref}	80 ms	160 Kbps
7	2	2.4 d_{ref}	250 ms	80 Kbps
8	1	2.4 d_{ref}	250 ms	160 Kbps

corroborate the good results of the WFO, we study in this section the performance of the tested protocols in a more general context. Eight groups of mobiles are considered here as described in Table 4.

Figures 15, 16, and 17, respectively, show the overall packet loss ratio and the dissatisfaction ratio with a PDOR target set to 5% and 10% for each group of mobiles and on the right, for all groups. MaxSNR provides a very poor QoS in groups 2, 5, and 6, that is, when delay requirements are stringent and the path loss or the source bit rate is high. This result confirms that MaxSNR severely lacks fairness in realistic scenarii. Mobile position has less consequences on fairness with PF. However, PF still shows deficiencies for mobiles with high data rate and tight delay threshold (groups 2 and 6). In comparison with PF, MAOPF brings more fairness between mobiles with heterogeneous data rate. Groups 2 and 6 experience less difficulties but at the expense of the satisfaction of groups 1 and 5. Globally, MaxSNR, PF, and MAOPF provide comparable performance results, each of them penalizing selectively some of the groups of mobiles. In contrast, WFO performs an efficient multiplexing and jointly manages all the mobiles so that they are all satisfied in a same proportion whatever their respective QoS constraints, positions, or data rate specificities. WFO allows to respect the delay thresholds in equity for all mobiles and satisfy the largest number.

6. Conclusion

In this paper, we propose a new MAC protocol for wireless multimedia networks, called “weighted fair opportunistic (WFO)” protocol. This access scheme operates on top of an OFDM-based physical layer and shows a good compatibility with the existing 802.16 standard. Full support of evolved multimedia services and QoS differentiation is enabled with the introduction of generic QoS attributes. Based on a system of weights, the WFO scheduling introduces dynamic priorities between the mobiles according to their transmission conditions and the delay they currently experience in a higher layers/MAC/PHY cross-layer approach. With its well-balanced resource allocation, the WFO scheme keeps a maximum number of service flows active across time but with relatively low traffic backlogs. Preserving the multiuser diversity, it takes a maximal benefit of the opportunistic scheduling technique for maximizing the system capacity. Simulation results show that the WFO outperforms other wireless OFDM-based scheduling schemes providing efficient QoS management. Fairness is ensured whatever the mobile position, the bit rate, or the delay constraints and without never sacrificing system capacity.

Acknowledgments

This work has been partially supported by the IST European Project WIP under Contract no. 27402. The authors also thanks the reviewers for their constructive comments.

References

[1] J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, NY, USA, 3rd edition, 1995.

[2] R. Steele and L. Hanzo, *Mobile Radio Communications*, IEEE Computer Society Press, Los Alamitos, Calif, USA, 2000.

[3] A. Goldsmith, *Wireless Communications*, Cambridge University Press, Cambridge, UK, 2005.

[4] J. D. Parsons, *The Mobile Radio Propagation Channel*, John Wiley & Sons, New York, NY, USA, 1992.

[5] R. Knopp and P. A. Humblet, “Information capacity and power control in single-cell multiuser communications,” in *Proceedings of IEEE International Conference on Communications (ICC ’95)*, vol. 1, pp. 331–335, Seattle, Wash, USA, June 1995.

[6] C. Y. Wong, R. S. Cheng, K. Ben Letaief, and R. D. Murch, “Multiuser OFDM with adaptive subcarrier, bit, and power allocation,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

[7] X. Wang and W. Xiang, “An OFDM-TDMA/SA MAC protocol with QoS constraints for broadband wireless LANs,” *Wireless Networks*, vol. 12, no. 2, pp. 159–170, 2006.

[8] Y. Fukui, N. Yamagaki, H. Tode, and K. Murakami, “Packet transfer scheduling scheme with throughput compensated considering wireless conditions,” in *Proceedings of IEEE International Conference on Computer Communications and Networks (ICCCN ’03)*, pp. 11–16, Dallas, Tex, USA, October 2003.

[9] H. Zhu and K. R. Liu, “Throughput maximization using adaptive modulation in wireless networks with fairness constraint,” in *Proceedings of IEEE International Wireless Communications*

and Networking Conference (WCNC '03), vol. 1, pp. 243–246, New Orleans, La, USA, March 2003.

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

[11] H. Kim, K. Kim, Y. Han, and J. Lee, "An efficient scheduling algorithm for QoS in wireless packet data transmission," in *Proceedings of the 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '02)*, vol. 5, pp. 2244–2248, Lisbon, Portugal, September 2002.

[12] W. Anchun, X. Liang, Z. Shidong, X. Xibin, and Y. Yan, "Dynamic resource management in the fourth generation wireless systems," in *Proceedings of the International Conference on Communication Technology (ICCT '03)*, vol. 2, pp. 1095–1098, Beijing, China, April 2003.

[13] P. Svedman, S. K. Wilson, and B. Ottersten, "A QoS-aware proportional fair scheduler for opportunistic OFDM," in *Proceedings of the 60th IEEE Vehicular Technology Conference (VTC '04)*, vol. 1, pp. 558–562, Los Angeles, Calif, USA, September 2004.

[14] H. Kim, K. Kim, Y. Han, and S. Yun, "A proportional fair scheduling for multicarrier transmission systems," in *Proceedings of the 60th IEEE Vehicular Technology Conference (VTC '04)*, vol. 1, pp. 409–413, Los Angeles, Calif, USA, September 2004.

[15] J.-G. Choi and S. Bahk, "Cell-throughput analysis of the proportional fair scheduler in the single-cell environment," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 2, pp. 766–778, 2007.

[16] J. M. Holtzman, "Asymptotic analysis of proportional fair algorithm," in *Proceedings of the 12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '01)*, vol. 2, pp. 33–37, San Diego, Calif, USA, September-October 2001.

[17] J. Z. Haiying and R. H. M. Hafez, "Novel scheduling algorithms for multimedia service in OFDM broadband wireless systems," in *Proceedings of IEEE International Conference on Communications (ICC '06)*, vol. 2, pp. 772–777, Istanbul, Turkey, July 2006.

[18] IEEEStd 802.16-2004, "IEEE standard for local and metropolitan area networks, part 16: air interface for fixed broadband wireless access systems," October 2004.

[19] C. Hoymann, "Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16," *Computer Networks*, vol. 49, no. 3, pp. 341–363, 2005.

[20] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–154, 2001.

[21] J.-J. van de Beek, P. O. Börjesson, M.-L. Boucheret, et al., "A time and frequency synchronization scheme for multiuser OFDM," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 11, pp. 1900–1914, 1999.

[22] Y. Li, N. Seshadri, and S. Ariyavisitakul, "Channel estimation for OFDM systems with transmitter diversity in mobile wireless channels," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 3, pp. 461–471, 1999.

[23] T. E. Truman and R. W. Brodersen, "A measurement-based characterization of the time variation of an indoor wireless channel," in *Proceedings of the 6th IEEE International Conference on Universal Personal Communications Record (ICUPC '97)*, vol. 1, pp. 25–32, San Diego, Calif, USA, October 1997.

[24] X. Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 2053–2064, 2001.

[25] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling: generalizations to include multiple constraints, multiple interfaces, and short term fairness," *Wireless Networks*, vol. 11, no. 5, pp. 557–569, 2005.

[26] S. Baey, "Modeling MPEG4 video traffic based on a customization of the DBMAP," in *Proceedings of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS '04)*, pp. 705–714, San Jose, Calif, USA, July 2004.

[27] P. Brady, "A model for generating on-off speech patterns in two-way conversation," *The Bell System Technical Journal*, vol. 48, no. 9, pp. 2445–2472, 1969.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 104548, 12 pages
doi:10.1155/2009/104548

Research Article

Decentralized Utility Maximization in Heterogeneous Multicell Scenarios with Interference Limited and Orthogonal Air Interfaces

Ingmar Blau,[1] Gerhard Wunder,[1] Ingo Karla,[2] and Rolf Sigle[2]

[1] *Fraunhofer German-Sino Lab for Mobile Communications (MCI), Fraunhofer-Institute for Telecommunications, Heinrich-Hertz-Institut, Einsteinufer 37, 10587 Berlin, Germany*
[2] *Bell Labs, Alcatel-Lucent Deutschland AG, 70435 Stuttgart, Germany*

Correspondence should be addressed to Ingmar Blau, blau@hhi.fhg.de

Received 6 August 2008; Revised 18 November 2008; Accepted 6 January 2009

Recommended by Mohamed Hossam Ahmed

Overlapping coverage of multiple radio access technologies provides new multiple degrees of freedom for tuning the fairness-throughput tradeoff in heterogeneous communication systems through proper resource allocation. This paper treats the problem of resource allocation in terms of optimum air interface and cell selection in cellular multi-air interface scenarios. We find a close to optimum allocation for a given set of voice users with minimum QoS requirements and a set of best-effort users which guarantees service for the voice users and maximizes the sum utility of the best-effort users. Our model applies to arbitrary heterogeneous scenarios where the air interfaces belong to the class of interference limited systems like UMTS or to a class with orthogonal resource assignment such as TDMA-based GSM or WLAN. We present a convex formulation of the problem and by using structural properties thereof deduce two algorithms for static and dynamic scenarios, respectively. Both procedures rely on simple information exchange protocols and can be operated in a completely decentralized way. The performance of the dynamic algorithm is then evaluated for a heterogeneous UMTS/GSM scenario showing high-performance gains in comparison to standard load-balancing solutions.

1. Introduction

In today's wireless scenarios, new radio access technologies (RATs) are emerging at frequent intervals. Although operators quickly introduce new wireless systems to the market they still have a strong interest in exploiting their legacy systems. Consequently, scenarios where an operator is in charge of multiple air interfaces with overlapping coverage are a common business case. Dense urban environments in Europe, where users are often in the coverage of a cellular TDMA-based GSM and CDMA-based UMTS systems, serve as a good example. In this case, if services are offered independently of the radio access technology and terminals support multiple wireless standards, the operator has the freedom to assign users to a cell and air interface of its choice.

Over the last years there has been growing interest in academics and industry in which way these degrees of freedom should be used and how users should be assigned in heterogeneous wireless scenarios to exploit resources more efficiently, incorporate fairness, and increase reliability. Established concepts include load-balancing, service-based, and cost-based strategies. Load-balancing strategies assign users such that overload situations are avoided in one RAT as long as there are resources left in a collocated radio system [1]. More advanced approaches are service-based strategies which select an RAT also in dependence of the requested service type [2]. These strategies exploit the fact that one wireless technology might be better suited to support a certain service-class than another one due to different granularities of distributable resources, different coding, and modulation schemes. However, both approaches neglect the fact that also the position and corresponding channel gain of a user influence the efficiency of an RAT supporting a service request. Reasons include different carrier frequencies and corresponding channel models of RATs, base station positioning, different interference situations and sensitivity to it. A concept that considers all earlier mentioned factors, like the system load, service class, interference situation, characteristics of the RAT, and users' positions, is the cost-based approach, introduced and analyzed in [3, 4]. There, it was observed that all characteristics can be bundled together in one cost parameter per user and RAT which suffice to calculate a close to optimum assignment that maximizes the total number of supportable voice users under static conditions. Alternative approaches can be found in [5] and references therein.

In this paper, we analyze in which way users of different service classes should be assigned in a heterogeneous scenario, thereby extending ideas from [3, 4]. Users request either a fixed minimum data rate, for example, as needed for voice services, or unconstrained best-effort (BE) data services. We formulate the user assignment as a utility maximization problem which is constrained by the resources (such as power or bandwidth) of the individual base stations (BSs) as well as users' minimum data rate requirements. The utilities represent quality of service (QoS) indicators of the BE users and, by choosing appropriate utility functions, give operators the freedom to tune the operation point of the heterogeneous system. It is important to note that although our model holds for general concave utility functions we will adopt the concept of α-proportional fairness introduced in [6] which allows to variably shift the operation point between maximum sum throughput, proportional fairness up to max-min fairness by a single, parameterizable utility function. Related work on utility maximization in nonheterogeneous interference limited systems was carried out in [7–9], where the generally nonconvex utility maximization problem was turned into a convex representation (or supermodular game) using specific techniques. The major difference to the approach taken in this paper is that we consider a heterogeneous scenario where the user-wise utilities are a function of the individual link rates; this practical assumption significantly complicates the analysis and neither of the approaches in [7–9] can be applied. Based on the convex formulation and by using structural properties, we present a decentralized algorithm that solves the optimization problem for static scenarios and derive simple assignment rules using the dual representation of the utility problem. The insights gained from the static setup are then adapted to dynamic scenarios and we design a distributed protocol which requires minimal information exchange between users and BSs and still achieves considerable performance gains. Most importantly, both algorithms allow operators to arbitrarily tune the fairness-throughput tradeoff online without any system changes. Although we cannot guarantee the convergence of the simplified algorithm in the dynamic scenario we observe a close to the global optimum operation in case a sufficient number of users requests service and the variation of the channel gains due to mobility is low. This is verified by the derivation of an upper bound and comparison to simulation results. Still, also for low service request rates and stronger channel variations due to mobility and fading considerable gains in terms of throughput and sum utility are obtained in comparison to a load-balancing strategy.

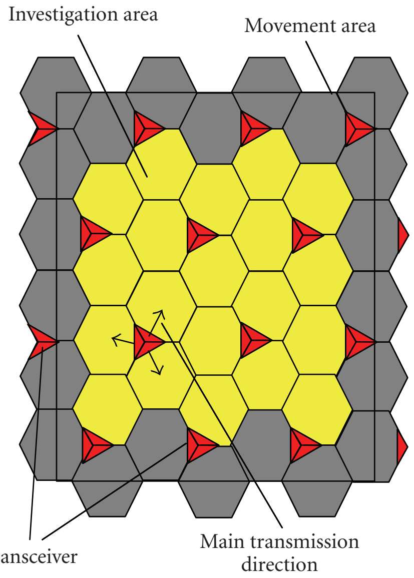


Figure 1: Playground with 40 GSM and 40 UMTS directional transceivers (collocated).

The paper is organized as follows: after the introduction of the system model and the utility concept in Section 2, we will formulate the optimization problem in Section 3. Algorithms that solve the problem in a decentralized way for static and dynamic scenarios are presented in Section 4. There, also the upper performance bound for the dynamic scenario is derived. In Section 5, we eventually evaluate the performance of the dynamic algorithm by comparing it to a load-balancing approach. We conclude the paper in Section 6.

Notations. In this work bold symbols denote vectors or matrices, calligraphic letters sets, and $|\cdot|$ the cardinality of a set. The transpose of a vector is $(\cdot)^T$, x_m is the mth element of $\mathbf{x}$, and $\mathbb{E}(\cdot)$ is the expectation. The summation over sets is defined as $\mathcal{X} = \sum_n \mathcal{X}_n = \{\mathbf{x} : \mathbf{x} = \sum_n \mathbf{x}_n,\ \mathbf{x}_n \in \mathcal{X}_n\}$.

2. System Model

We consider a wireless scenario in the down-link direction where multiple RATs with partly overlapping coverage are arranged in an area called playground. The set of RATs $\mathcal{A} = \mathcal{A}_{\text{orth}} \cup \mathcal{A}_{\text{inf}}$ thereby consists of two subsets: in RATs with orthogonal resources $a \in \mathcal{A}_{\text{orth}}$ time or frequency slots or subcarriers are assigned explicitly and users connected to one BS do not interfere with each other. In interference limited RATs $a \in \mathcal{A}_{\text{inf}}$ all users share the same bandwidth and the power constitutes the distributable resource. Each RAT $a \in \mathcal{A}$ consists of a set of base stations $m \in \mathcal{M}_a$ and one operator is assumed to control the set of all base stations $\mathcal{M} = \bigcup_{a\in\mathcal{A}} \mathcal{M}_a$. An exemplary scenario with one cellular UMTS system belonging to the interference limited class and one cellular GSM/EDGE air interface of the orthogonal class is depicted in Figure 1.

Since commercial wireless systems usually operate on individual frequency bands, we assume that signals of different RATs are orthogonal to each other and no intersystem interference takes place. Users can be affected by intra- and intercell interference within one radio technology, however.

The set of users $\mathcal{I}$ can be divided into two subsets and users are equally distributed on the playground; users $i \in \mathcal{I}_v$ request a voice service with guaranteed data rate and have priority to BE users $i \in \mathcal{I}_b$ who do not have any QoS guarantees. Furthermore, it is assumed that the user equipment is able to cope with all RATs and the service requests are independent of the technology giving the operator the freedom to choose a cell and a RAT for each user that is best suited from its perspective.

Next we will describe the two classes of RATs that are covered in our scenario in more detail.

2.1. Orthogonal RATs. For the class of orthogonal systems we assume a fixed transmission power per BS and that the bandwidth, in terms of time or frequency slots, respectively, is the resource continuously distributable between users. Since commercial TDMA systems like GSM/EDGE usually have low frequency reuse factors we will assume constant intercell interference for this class of systems. The signal to interference and noise ratio (SINR) of user i and a BS m of this class

$$\beta_{i,m} = \frac{g_{i,m}\overline{P}_m}{\eta_m + I_m} \quad \forall m \in \mathcal{M}_a,\ a \in \mathcal{A}_{\text{orth,}} \tag{1}$$

thus depends on the channel gain $g_{i,m}$, the BS power $\overline{P}_m$, the constant intercell interference I_m, the thermal noise η_m, and is independent of the assigned resource. The amount of bandwidth assigned to user i by BS m is denoted by $t_{i,m}$. It is limited by the total, distributable bandwidth per BS $\overline{T}_m$ and the constraint

$$\sum_{i\in\mathcal{I}} t_{i,m} = t_m \leq \overline{T}_m \quad \forall m \in \mathcal{M}_a,\ a \in \mathcal{A}_{\text{orth}}. \tag{2}$$

Due to the orthogonality of the users' signals and since the bandwidth is the distributable resource the relation between a user's data rate $R_{i,m}$ and the assigned resource is linear for this class of RATs:

$$R_{i,m} = \overline{R}_{i,m}t_{i,m.} \tag{3}$$

Here, $\overline{R}_{i,m} := f(\beta_{i,m})$ denotes the link rate per time or frequency slot between user i and base station m where $f(\beta)$ is a positive, nondecreasing SINR-rate mapping curve corresponding to the coding and transmission technology of the RAT $a \in \mathcal{A}_{\text{orth}}$. By substituting (3) into (2) the achievable rate region of each individual BS $m \in \mathcal{M}_a$ results in an I-dimensional simplex, limited by the positive orthant and a hyperplane:

$$\mathcal{R}_m = \left\{\mathbf{R}_m : \sum_{i\in\mathcal{I}} \frac{R_{i,m}}{\overline{R}_{i,m}} \leq \overline{T}_m,\ R_{i,m} \geq 0\ \forall i \in \mathcal{I}\right\}, \tag{4}$$

where $\mathbf{R}_m$ is the i-dimensional vector with entries $R_{i,m}$. Since the rate assignment in one cell does not influence the feasible rate region of neighboring cells the feasible rate region of the whole RAT results in the convex polytope

$$\mathcal{R}_a = \sum_{m\in\mathcal{M}_a} \mathcal{R}_m, \quad a \in \mathcal{A}_{\text{orth.}} \tag{5}$$

2.2. Interference Limited RATs. We assume that all users share the same bandwidth and that resources are distributed in terms of assigned power for BSs in interference limited air interfaces like UMTS $m \in \mathcal{M}_b$, $b \in \mathcal{A}_{\text{inf}}$. The power of each BS is limited by a sum constraint

$$\sum_{i\in\mathcal{I}} p_{i,m} = P_m \leq \overline{P}_m \quad \forall m \in \mathcal{M}_b,\ b \in \mathcal{A}_{\text{inf,}} \tag{6}$$

where $p_{i,m}$ is the power that BS m assigns to user $i \in \mathcal{I}$. Users are sensitive to intracell and intercell interference in interference limited systems and the SINR between BS $m \in \mathcal{M}_b$, $b \in \mathcal{A}_{\text{inf}}$ and user $i \in \mathcal{I}$ is given by

$$\begin{gathered}\beta_{i,m} = \frac{g_{i,m}p_{i,m}}{\rho g_{i,m}\sum_{j\neq i} p_{j,m} + \sum_{n\neq m} g_{i,n}P_n + \eta_{\text{inf}}} \\ m, n \in \mathcal{M}_b, \quad b \in \mathcal{A}_{\text{inf}}, \quad i, j \in \mathcal{I},\end{gathered} \tag{7}$$

with ρ the orthogonality factor which accounts for a reduced intercell interference. In this class of systems all links of one BS share a limited power budget and are impaired by the power assigned to other users in the air interface. A well-known model for the link rate of these systems is given in [10]:

$$\begin{aligned}R_{i,m} &= C_b \log\left(1 + D_b\beta_{i,m}\right) \\ &= C_b \log\left(1 + D_b\frac{g_{i,m}p_{i,m}}{\rho g_{i,m}(\overline{P}_m - p_{i,m}) + \sum_{n\neq m} g_{i,n}\overline{P}_n + \eta_{\text{inf}}}\right).\end{aligned} \tag{8}$$

There, the positive constants C_b, D_b parameterize the system characteristics such as bandwidth, modulation, and bit-error rates. In (8), a user's data rate is in general neither convex nor concave in p (index omitted). Therefore, also the feasible rate region is not convex, which in turn will be a requirement to obtain a convex representation of the utility maximization problem in Section 3. However, assuming that all BS transmit with fixed transmission power and that the SINR of all links is not too high we can approximate the data rate by

$$\begin{aligned}R_{i,m} &= C_b \log\left(1 + D_b\frac{p_{i,m}}{I_{i,m} - \rho p_{i,m}}\right) \\ &\approx \frac{\Delta_b}{I_{i,m}} p_{i,m} \\ &=: \overline{R}_{i,m}p_{i,m},\end{aligned} \tag{9}$$

with

$$I_{i,m} = \frac{\rho g_{i,m}\overline{P}_m + \sum_{n\neq m\in\mathcal{M}_b} g_{i,n}\overline{P}_n + \eta_{\text{inf}}}{g_{i,m}}. \tag{10}$$

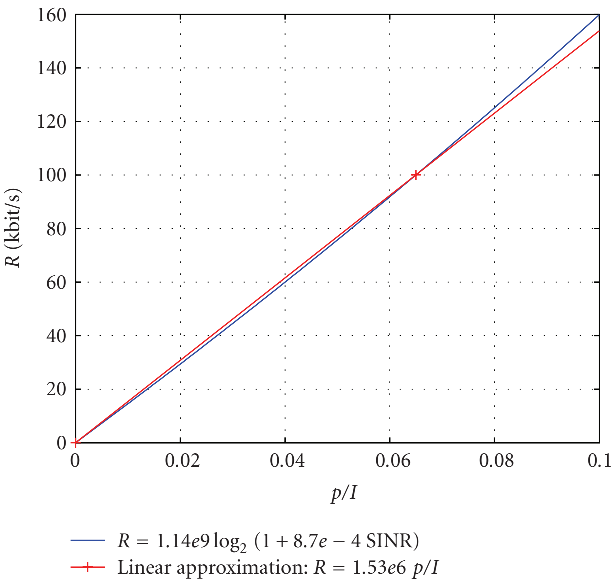


Figure 2: UMTS resource-rate mapping: quality of linear approximation (9).

The approximation in (9) represents the first order Taylor expansion for $p = 0$ if one chooses $\Delta_b = C_b D_b$. Clearly, this approximation holds only for low data rates and since we are interested in a good approximation for typical rates of the UMTS system, it turns out to be practical to use a higher slope $\Delta_b > C_b D_b$. Indeed we plotted the rates in (9) over p/I for UMTS in Figure 2 and chose Δ_b so that it intersects the real rate curve at the origin and 100 kbit/s which covers the range of rates that are typically assigned to users in UMTS in our scenario quite well. Obviously, this is only a model, but works fine for the problem at hand. We refer also to the discussion in Section 5.

By solving the approximation in (9) for p and substitution into (6) the achievable rate region of BS $m \in \mathcal{M}_b$ can be represented by

$$\mathcal{R}_m = \left\{ \mathbf{R}_m : \sum_{i\in\mathcal{I}} \frac{R_{i,m}}{\overline{R}_{i,m}} \le \overline{P}_m,\ R_{i,m} \ge 0\ \forall i \in \mathcal{I} \right\}. \tag{11}$$

Since all BS are assumed to transmit with $P_m = \overline{P}_m$, the intercell interference is independent of the resource assignment and the achievable rate region of the whole RAT results in

$$\mathcal{R}_b = \sum_{m\in\mathcal{M}_b} \mathcal{R}_m, \quad b \in \mathcal{A}_{\text{inf}}, \tag{12}$$

which is a convex polytope as for the orthogonal RAT.

Our approach stands in clear contrast to [8] where a convex feasible rate region for interference limited RATs was obtained with the posynomial transform and assuming $R \approx C\log(D\beta)$. The posinomial approach has the advantage that also the BS sum transmission power P_m can be optimized. However, the corresponding rate approximation is only valid for high SINR and does not hold in our scenario. The linear structure of our approximation will further lead to simple assignment rules in Section 3.

2.3. Utility Concept and α-Proportional Fairness. Instead of maximizing a fixed metric like the system throughput, we will formulate the optimization problem in terms of utility functions, which relate assigned resources, system parameters as the SINR or the data rate to benefits such as revenues, fairness or user satisfaction. More precisely, we focus our investigations on utility functions which are concave, strictly increasing and dependent on the user's data rate in the following form:

$$U = \sum_{i\in\mathcal{I}_b} \psi_i \left(\sum_{m\in\mathcal{M}} R_{i,m} \right). \tag{13}$$

Without loss of generality ψ_i in (13) is given by

$$\psi_i^{\alpha}(R_i) = \begin{cases} w_i \log(R_i), & \text{if } \alpha = 1, \\ \dfrac{w_i}{1-\alpha} R_i^{1-\alpha}, & \text{otherwise}. \end{cases} \tag{14}$$

Utilities defined by (13) and (14) correspond to the well-established weighted α-proportional fairness [6], and are from special interest for operators since they ensure flexible tuning of the system fairness in a wide range. A rate allocation $\mathbf{R}^*$ is said to be α-proportional fair, if for any feasible allocation $\mathbf{R}$

$$\sum_{i\in\mathcal{I}_b} \frac{R_i - R_i^*}{R_i^{*^{\alpha}}} \le 0 \tag{15}$$

holds [6]. The parameter α in (14) hereby tunes the fairness-throughput tradeoff; for $\alpha = 0$ the system throughput will be maximized, which might result in assignments where only very few users are served and which is quite unfair. A selection $\alpha = 1$ leads to proportional fairness which is equivalent to assigning equal shares of resources to all users in our scenario. For $\alpha \to \infty$ the assignment converges to the max-min fairness, where all users will be assigned equal data rates and the overall system throughput will be low [6].

Note that the definition of the utility in terms of the sum of a user's link rates in (13) is more relevant for practical application than, for example, the sum utilities of individual links $U = \sum_i \sum_m \psi(R_{i,m})$ used in [7, 9]. It turns out that it is exactly this so-called nonseparable utility formulation that leads to the desired characteristic that most users will establish only a single link, as will be shown in Section 3. By contrast, the separable utility in [7, 9] will favor multilink operation and therefore the results cannot be applied to our model. This follows from the concavity of ψ and the Jensen's inequality; assume a user is assigned a certain sum rate R_i that can be split between two links $R_{i,m}$ and $R_{i,n}$, $R_i = R_{i,m} + R_{i,n}$. Then, it is beneficial in terms of the separable sum utility to activate both links because $\psi(R_{i,m}) + \psi(R_{i,n}) \ge \psi(R_i)$.

3. Problem Formulation

Having the system model and the utility concept introduced, we now present the formal problem formulation. We want to find the user assignment in a heterogeneous multicell scenario that maximizes the sum utility of all BE users under the constraint that all voice users are assigned at least a minimum data rate $R_{\text{min},i}$. Based on the earlier presented assumptions, the problem can be formulated as

$$\begin{aligned}
&\max_{\mathbf{R}} \sum_{i\in\mathcal{I}_b} \psi_i\left(\sum_{m\in\mathcal{M}} R_{i,m}\right),\\
&\text{subject to} \sum_{i\in\mathcal{I}} \frac{R_{i,m}}{\overline{R}_{i,m}} \leq \Gamma_m \quad \forall m \in \mathcal{M},\\
&\sum_{m\in\mathcal{M}} R_{i,m} \geq R_{\text{min},i} \quad \forall i \in \mathcal{I}_v,\\
&R_{i,m} \geq 0 \quad \forall i, m \in \mathcal{I}, \mathcal{M},
\end{aligned} \tag{P1}$$

with Γ_m denoting available resources, $\Gamma_m = \overline{P}_m\ \forall m \in \mathcal{M}_b$, $b \in \mathcal{A}_{\text{inf}}$ or $\Gamma_m = \overline{T}_m\ \forall m \in \mathcal{M}_a$, $a \in \mathcal{A}_{\text{orth}}$, respectively. Problem (P1) consists of a concave objective over linear constraints and is therefore convex. Consequently, a variety of ready-to-use algorithms exists to solve it [11]. However, neither give these algorithms insights into the problem structure nor do they give a hint to a decentralized solution. We therefore develop a different approach based on duality [11, 12]; instead of solving (P1) directly we transform it into an alternative problem which is known to have the same solution as (P1) but can be solved in a decentralized way. To obtain an expression for the dual transform the Lagrangian function of (P1) is needed, which has the following form:

$$\begin{aligned}
\mathcal{L}(\mathbf{R}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = {} & \sum_{i\in\mathcal{I}_b} \psi_i\left(\sum_{m\in\mathcal{M}} R_{i,m}\right)\\
& - \sum_{m\in\mathcal{M}} \lambda_m \left(\sum_{i\in\mathcal{I}_b} \frac{R_{i,m}}{\overline{R}_{i,m}} - \Gamma_m\right)\\
& + \sum_{i\in\mathcal{I}_v} \mu_i \left(\sum_{m\in\mathcal{M}} R_{i,m} - R_{\text{min},i}\right)\\
& + \sum_{i\in\mathcal{I}} \sum_{m\in\mathcal{M}} \sigma_{i,m} R_{i,m}.
\end{aligned} \tag{16}$$

Here $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ are nonnegative dual parameters. Next, we introduce the dual function of (P1) which is defined as [11]

$$g(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\sigma}) = \max_{\mathbf{R}} \mathcal{L}(\mathbf{R}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\sigma}). \tag{17}$$

Due to nonnegativity of the dual parameters one observes that (17) is always larger than or equal to the solution of (P1). Therefore, minimizing the unconstrained dual function over the dual parameters

$$\min_{\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\sigma}\geq 0} g(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\sigma}) = \min_{\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\sigma}\geq 0} \underbrace{\max_{\mathbf{R}} \mathcal{L}(\mathbf{R}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\sigma})}_{\text{inner problem}} \tag{18}$$

yields an upper bound on the original optimization problem (P1) and is called the dual problem of (P1). Furthermore, by convexity of (P1) and since Slater's conditions [11] hold, the bound is tight and (18) and (P1) have the same solution. Our motivation to use the dual formulation is the possibility to decouple the optimization problem into an inner maximization problem over the primal variables $\mathbf{R}$ and an outer minimization over the dual parameters which will be called outer loop further on. Additionally, the dual problem allows to exploit structural properties which will greatly simplify the algorithm design. The inner problem can be solved by each base station individually as we will see shortly. In addition, there exists a very limited number of degrees of freedom for the selection of meaningful dual parameters in the outer loop. To be more precise, only $\boldsymbol{\lambda}$ has to be optimized iteratively in the outer minimization. A rate allocation $\mathbf{R}(\boldsymbol{\lambda})$ that maximizes the inner problem can be calculated directly for a given $\boldsymbol{\lambda}$ independently of $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$. Before we go into the details the KKT conditions are given, which are necessary and sufficient for the optimum solution of (P1) (or equivalently (18))[11] and will be exploited later:

$$\frac{\partial \mathcal{L}(\mathbf{R}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*, \boldsymbol{\sigma}^*)}{\partial R_{i,m}} = 0 \quad \forall m, i \in \mathcal{M}, \mathcal{I}, \tag{19}$$

$$\lambda_m^* \left(\sum_{i\in\mathcal{I}} \frac{R_{i,m}^*}{\overline{R}_{i,m}} - \Gamma_m\right) = 0 \quad \forall m \in \mathcal{M}, \tag{20}$$

$$\mu_i^* \left(R_{\text{min},i} - \sum_{m\in\mathcal{M}} R_{i,m}^*\right) = 0 \quad \forall i \in \mathcal{I}_v, \tag{21}$$

$$\sigma_{i,m}^* R_{i,m}^* = 0 \quad \forall i, m \in \mathcal{I}, \mathcal{M}. \tag{22}$$

Here $(\cdot)^*$ denotes the variables at the optimum.

3.1. Inner Problem. Rearranging terms in (16) results in the following:

$$\begin{aligned}
\mathcal{L}(\mathbf{R}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\sigma}) = {} & \sum_{i\in\mathcal{I}_b} \psi_i\left(\sum_{m\in\mathcal{M}} R_{i,m}\right)\\
& + \sum_{i\in\mathcal{I}_v} \sum_{m\in\mathcal{M}} R_{i,m} \left(\sigma_{i,m} - \frac{\lambda_m}{\overline{R}_{i,m}} + \mu_i\right)\\
& + \sum_{i\in\mathcal{I}_b} \sum_{m\in\mathcal{M}} R_{i,m} \left(\sigma_{i,m} - \frac{\lambda_m}{\overline{R}_{i,m}}\right)\\
& + \sum_{m\in\mathcal{M}} \lambda_m \Gamma_m - \sum_{i\in\mathcal{I}_v} \mu_i R_{\text{min},i}.
\end{aligned} \tag{23}$$

From (23), one observes that (17) is only finite if and only if

$$\sigma_{i,m} - \frac{\lambda_m}{\overline{R}_{i,m}} + \mu_i = 0 \quad \forall m, i \in \mathcal{M}, \mathcal{I}_v, \tag{24}$$

$$\frac{\lambda_m}{\overline{R}_{i,m}} > \sigma_{i,m} \quad \forall m, i \in \mathcal{M}, \mathcal{I}_b, \tag{25}$$

and hence it follows that (24) and (25) are necessary conditions to obtain a meaningful solution in (18). Furthermore, the first KKT condition (19) has to hold for any rate assignment that solves (17) which after substituting (24) into (23) simplifies to

$$\frac{\partial \mathcal{L}}{\partial R_{i,m}} = \psi_i'\left(\sum_{m\in\mathcal{M}} R_{i,m}\right) + \sigma_{i,m} - \frac{\lambda_m}{\overline{R}_{i,m}} = 0 \quad \forall m, i \in \mathcal{M}, \mathcal{I}_b. \tag{26}$$

Here, $\psi_i'(x) = \partial\psi_i'(x)/\partial x$ and (26) are necessary and sufficient conditions for the maximum of the Lagrangian function which is independent of the voice users. Although the optimization of the dual parameters is formally performed in the outer problem, one observes already here that only certain $\boldsymbol{\sigma}$ can lead to the optimum solution of (P1). More precisely, for a given $\boldsymbol{\lambda}$ only one element $\sigma_{i,m}$ can be chosen freely for each user i so that (26) is not violated. All other elements $\sigma_{i,n}, n \neq m$ result directly from $\sigma_{i,m}$ by (26). This is shown in the following example: assume one element $\sigma_{i,m}$ and $\boldsymbol{\lambda}$ are given for user i from the outer loop. Then, for the rate assignment that maximizes the inner problem $u_i := \psi_i'(\sum_{m\in\mathcal{M}} R_{i,m}) = (\lambda_m/\overline{R}_i, m) - \sigma_{i,m}$ has to hold (from (26)). Since (26) is a necessary condition also for all $n \neq m$ it follows that $\sigma_{i,n} = u_i(\sigma_{i,m}) + (\lambda_m/\overline{R}_i, n)$, $n \neq m$ which is therefore uniquely determined by $\sigma_{i,m}$. This observation reduces the degrees of freedom to select meaningful $\boldsymbol{\sigma}$ to one scalar element per user in the outer loop. From (26), it further follows that $\sigma_{i,m} = 0$ can only hold for $m \in \mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda})$, with

$$\mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda}) = \left\{ m_i' \in \mathcal{M} : m_i' = \arg\min_m \frac{\lambda_m}{\overline{R}_{i,m}} \right\}. \tag{27}$$

This is a direct consequence of the nonnegativity of the dual parameters and u_i based on (26). Having $\sigma_{i,m} = 0$, however, is a necessary condition for $R_{i,m}^* > 0$ since for any optimum rate assignment of (P1) the last KKT condition (22) has to be fulfilled. Therefore, regardless of the outer optimization we can already state here that $\sigma_{i,n} > 0 \, \forall n \notin \mathcal{M}_{\text{opt},i}$, $i \in \mathcal{I}_b$ and only rate assignments

$$R_{i,m} \begin{cases} \geq 0 & \forall m \in \mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda}), \\ = 0 & \text{else} \end{cases} \tag{28}$$

have to be considered as solution for (P1). Furthermore, setting $\sigma_{i,m} = 0$ $m \in \mathcal{M}_{\text{opt},i}$ if possible is required to allow for assignments with $R_{i,m} > 0$. Only if the maximum slope of the utility function $\psi'(0)$ is smaller than $\min_m(\lambda_m/\overline{R}_{i,m})$ this will result in $\sigma_{i,m} > 0 \, \forall m \in \mathcal{M}_{\text{opt},i}$ then so that (26) is not violated. In this case user i will not be assigned any resources. The KKT conditions lead to similar optimality conditions for the voice users; from (24) as well as the argumentation above it follows that

$$\mu_i = \min_m \frac{\lambda_m}{\overline{R}_{i,m}} \quad \forall i \in \mathcal{I}_v, \tag{29}$$

and that (28) is also a necessary condition for the voice users. It is noted here that for a given $\boldsymbol{\lambda}$ the solution of (17) is uniquely determined (see proof of Theorem 1 in Section 4). However, the corresponding rate assignment might not be unique. Multiple optimum rate assignments can exist in the rare case when $\exists\{m, n \in \mathcal{M}, m \neq n : \lambda_m/\overline{R}_{i,m} = \lambda_n/\overline{R}_{i,n}\}$ and therefore $|\mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda})| > 1$. For all other users it follows by (26) and the discussions on $\boldsymbol{\sigma}$ that the rate assignment

$$R_{i,m}(\boldsymbol{\lambda}) = \begin{cases} \psi_i'^{-1}\left(\dfrac{\lambda_m}{\overline{R}_{i,m_i}}\right) & \text{if } \psi_{i,m}'(0) > \dfrac{\lambda_m}{\overline{R}_{i,m}}, \\ & m \in \mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda}),\ \forall i \in \mathcal{I}_b, \\ R_{\min,i} & \text{if } m \in \mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda}),\ \forall i \in \mathcal{I}_v, \\ 0 & \text{else} \end{cases} \tag{30}$$

maximizes the inner problem and solves (17). In this case, the rate assignment is unique and only depends on $\boldsymbol{\lambda}$. In (30), ψ'^{-1} is the inverse of the derivative of the utility function with $\psi'(\psi'^{-1}(x)) = x$.

Equation (30) gives some valuable insights to the optimum cell/RAT selection of users and the corresponding resource assignment. First, it can be shown that almost all users are assigned to exactly one BS since $|\mathcal{M}_{\text{opt},i}| = 1$ in general. Second, this BS can be determined independently by each user if $\boldsymbol{\lambda}$ is known and under the assumption that each user i can measure $\overline{R}_{i.m} \, \forall m \in \mathcal{M}$. Both characteristics rely on the linear connection between the data rate and the assigned resources and on the user based utilities and greatly simplify the distributed solution of (P1). In contrast, one would obtain that $R_{i,m}^* > 0 \, \forall i, m \in \mathcal{I}, \mathcal{M}_b, b \in \mathcal{A}_{\inf}$ under the high SINR assumption in [7, 9], which implies that all users have active connections to all BSs in the interference limited air interface. Third, the maximum slope of the utility function $\psi_i(0)$ defines a threshold which can be tuned to switch off BE users with low $\overline{R}_{i,m}$, as will be described in Section 5.

3.2. Outer Problem. Since for $\boldsymbol{\mu}$ (24) has to hold, $\boldsymbol{\lambda}$ and formally $\boldsymbol{\sigma}$ are the only dual parameters that have to be considered in the outer optimization. In order to minimize the dual (17), clearly all entries of $\boldsymbol{\sigma}$ have to be as small as possible and chosen in a way that (26) holds. Therefore, $\sigma_{i,m_i'} = 0 \, \forall \{i, m_i' : i \in \mathcal{I}_b, m_i' \in \mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda}), \lambda_{m_i'}/\overline{R}_{i,m_i'} \leq \psi(0)\}$. A subgradient approach can be applied to minimize the dual over $\boldsymbol{\lambda}$ [12]. Assume for a given $\hat{\boldsymbol{\lambda}}$

$$\hat{\mathbf{R}} = \arg\max_{\mathbf{R}} \mathcal{L}(\mathbf{R}, \hat{\boldsymbol{\lambda}}) \tag{31}$$

is the solution of inner problem, obtained by (30). Then, the following holds for the dual function [12]

$$g(\boldsymbol{\lambda}) \geq \mathcal{L}(\hat{\mathbf{R}}, \boldsymbol{\lambda}) = \mathcal{L}(\hat{\mathbf{R}}, \hat{\boldsymbol{\lambda}}) + \sum_{m\in\mathcal{M}} (\lambda_m - \hat{\lambda}_m) \left(\Gamma_m - \sum_{i\in\mathcal{I}} \frac{\hat{R}_{i,m}}{\overline{R}_{i,m}} \right), \tag{32}$$

where the last equation is obtained by adding and subtracting the terms $\sum_{m\in\mathcal{M}} \hat{\lambda}_m(\Gamma_m - \sum_{i\in\mathcal{I}}(\hat{R}_{i,m}/\overline{R}_{i,m}))$ to $\mathcal{L}(\hat{\mathbf{R}}, \hat{\boldsymbol{\lambda}})$ and the assumption that $\sigma_{i,m} R_{i,m} = 0 \, \forall i, m \in \mathcal{I}, \mathcal{M}$. Further, it can be shown from (32) that the vector $\boldsymbol{\nu}$, with $\nu_m = (\Gamma_m - \sum_{i \in \mathcal{I}}(\hat{R}_{i,m}/\overline{R}_{i,m}))$ is a subgradient.

A descriptive explanation of the subgradient approach is as follows: for a given $\hat{\boldsymbol{\lambda}} \neq \boldsymbol{\lambda}^*$ the rate assignment $\hat{\mathbf{R}}$ might either violate the feasible rate region constraint or will not exploit all available resources. Both cannot be optimal since the first case is not feasible and in the latter case the assignment of more resources to any BE user would increase the sum utility. Then, the subgradient gives the direction how $\boldsymbol{\lambda}$ should be updated so that the resource constraints are less violated or more resources are assigned. At the global optimum of (P1), all entries of the subgradient will be zero and all resource constraints are met with equality. The subgradient will be used in the decentralized algorithm, which will be presented in Section 4.

4. Algorithm

We will now present two decentralized algorithms for (P1) in a static and dynamic scenario, respectively. In the static setup, all user requests and channel gains are assumed to be fixed, while in the dynamic one the requests and user mobility are subject to stochastic processes. The static algorithm hereby serves as motivation for the dynamic one which is adapted for practical applications with the advantage of requiring almost no signaling information.

4.1. Static Scenario. Based on the optimality conditions of the inner problem and the subgradient of the outer loop in Section 3, we are able to formulate the static Algorithm 1, where l denotes the index of the iteration, $\delta(l)$ is the step size, and ϵ a constant for the stopping criteria. The algorithm consists of an iterative procedure where in each cycle at first all BSs broadcast the BS weights λ_m to all users. Then, each user i evaluates $\boldsymbol{\lambda}_m/\overline{R}_{i,m}$ for all BSs and sends an assignment request (and the corresponding $\overline{R}_{i,m}$ or $R_{\min,i}$) to a BS $m_i' \in \mathcal{M}_{\text{opt},i}$. Next, each BS m individually calculates the rate assignment for all users that sent an assignment request to it. The rate assignment hereby depends on λ_m and might lie either inside, on, or outside the feasible rate region of BS m and thereby either under exploit, meet with equality or violate the resource constraint. Correspondingly, BS m will update λ_m using the subgradient and the cycle starts again by broadcasting the updated BS weight. Although Algorithm 1 might not converge to the optimum rate assignment in case $\exists\{m, n \in \mathcal{M}, m \neq n : \lambda_m^*/\overline{R}_{i,m} = \lambda_n^*/\overline{R}_{i,n}\}$ and therefore results in $|\mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda}^*)| > 1$, we can formulate the following theorem.

Theorem 1. *Assume that for the series* $\lim_{l\to\infty}\delta(l) = 0$, $\limsup_{l\to\infty}\sum_l \delta(l) = \infty$ *holds and that a feasible allocation for the voice users exists, then* Algorithm 1 *converges to the optimum dual weights* $\boldsymbol{\lambda}^*$. *In case* $|\mathcal{M}_{opt,i}(\boldsymbol{\lambda}^*)| = 1\ \forall i \in \mathcal{I}$ *the corresponding rate assignment of* Algorithm 1 *is also optimal. In case* $\exists i \in \mathcal{I} : |\mathcal{M}_{opt,i}(\boldsymbol{\lambda}^*)| > 1$ *an optimum rate assignment that solves* (P1) *can be obtained by solving the set of linear equations:*

$$\begin{aligned}
\sum_{m\in\mathcal{M}_{opt,i}} R_{i,m}^* &= \psi'^{-1}\left(\min\left\{\min_m \frac{\lambda_m^*}{\overline{R}_{i,m}}, \psi'_i(0)\right\}\right), \quad \forall i \in \mathcal{I}_b,\\
\sum_{m\in\mathcal{M}_{opt,i}} R_{i,m}^* &= R_{\min,i}, \quad \forall i \in \mathcal{I}_v,\\
\sum_{i\in\mathcal{I}} R_{i,m}^* &= \Gamma_m, \quad \forall m \in \mathcal{M}.
\end{aligned} \tag{33}$$

Proof. In Section 3.1, it was shown that steps (3) and (4) of Algorithm 1 maximize the inner problem of (18) in case $|\mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda})| = 1\ \forall i \in \mathcal{I}$. Step (5) corresponds to an update of $\boldsymbol{\lambda}$ in direction of the negative subgradient which was derived in Section 3.2. Since (P1) is a convex optimization problem and Slater's condition holds, it is proven in [12] that the dual problem (18) has the same solution as (P1). Further, it is shown in [12] that dual subgradient algorithms like Algorithm 1 converge to the global optimum for the given step-width constraints. The proof can be extended to the case where $\exists i \in \mathcal{I} : |\mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda})| > 1$ by observing the fact that the maximum of the inner problem is independent of the BS $m_i \in \mathcal{M}_{\text{opt},i}$ which is selected by user i in step (3) (however, it clearly matters for complying with the feasible rate region constraints); from (26) it follows that

$$R_i = \sum_m R_{i,m} = \psi'^{-1}\underbrace{\left(\frac{\lambda_m}{\overline{R}_{i,m}} - \sigma_{i,m}\right)}_{\zeta_i} \quad \forall m, i \in \mathcal{M}, \mathcal{I}_b \tag{34}$$

is necessary and sufficient for the maximization of the inner problem and that by (21) $\sum_{m\in\mathcal{M}} R_{i,m} = R_{\min,i}\ \forall i \in \mathcal{I}_v$ holds. Substituting this into the Lagrangian (23) together with (24) results in a dual function

$$\begin{aligned}
g(\boldsymbol{\lambda}) = &\sum_{i\in\mathcal{I}_b} \psi\left(\psi'^{-1}(\zeta_i)\right) - \sum_{i\in\mathcal{I}_b} \zeta_i \psi'^{-1}(\zeta_i)\\
&+ \sum_{m\in\mathcal{M}} \lambda_m \Gamma_m - \sum_{i\in\mathcal{I}_v} \mu_i R_{\min,i},
\end{aligned} \tag{35}$$

which is independent of the actual BS selection of the users. Therefore, Algorithm 1 will converge to the optimum $\boldsymbol{\lambda}^*$ and to the maximum utility also if $\exists i \in \mathcal{I} : |\mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda})| > 1$. The optimum rate assignment of users that are in multilink operation results then from $\boldsymbol{\lambda}^*$ by solving the set of KKT conditions which reduce to (33) since $\lambda_m^* > 0\ \forall m \in \mathcal{M}$, $\mu_i > 0\ \forall i \in \mathcal{I}_v$ for any nontrivial solution. □

4.2. Dynamic Scenario. In a dynamic scenario where users and service requests follow stochastic mobility and traffic models, respectively, applying Algorithm 1 might be a good choice from a theoretic perspective. Practically, however, the procedure is too expensive, since, having the optimum user assignment at any point in time, it would have to be executed any time a user's channel gain or interference

```
**(1)** Each BS initializes $\lambda_m, \nu_m = 1 \; \forall m \in \mathcal{M}, \; l = 0$.
**while** $!((\boldsymbol{\nu}(\boldsymbol{\nu})^T > \epsilon) || (l < l_{\max}))$ **do**
    **(2)** Each BS broadcasts $\lambda_m$ to all users.
    **(3)** Each user $i \in \mathcal{I}$ evaluates $\mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda})$ with (27) and announces an assignment request to $m_i'(\boldsymbol{\lambda}) \in \mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda})$. If $|\mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda})| > 1$ it picks one BS of the set randomly.
    **(4)** Based on the assignment requests each BS calculates the rate assignment that maximizes its sum utility and that fulfills the voice user's rate constraints corresponding to (30).
    **(5)** Each BS evaluates its sub-gradient component $\nu_m = (\Gamma_m - \sum_{i \in \mathcal{I}} (R_{i,m}/\overline{R}_{i,m}))$ and updates its dual weight $\lambda_m(l+1) = \lambda_m(l) - \delta(l)\nu_m; l = l+1$.
**end while**
**(6)** Assign users to $m_i'(\boldsymbol{\lambda}^*)$ with $R_{i,m}$ corresponding to **(3)**, **(4)**.
```

Algorithm 1: Decentralized utility maximization.

situation changes (and therefore $\overline{R}$) and in case a service request arrives or leaves the system. Each execution thereby might trigger reassignments of a whole set of users and a considerable amount of signaling information would have to be exchanged between users and BSs in each iteration. (It is noted here that higher utilities might be obtainable in the dynamic scenario by exploitation of mobility information or, e.g., under the fluid assumptions [13].) We therefore suggest the following adaptation of Algorithm 1 to a dynamic procedure which can be split into two almost independently operating parts, the cell/RAT selection of users and the resource assignment inside each BS.

A user's heterogeneous cell/RAT selection procedure is described in Algorithm 2(a). It is similar to the one in the static setup; the BSs broadcast $\boldsymbol{\lambda}$ and each user selects a BS $m \in \mathcal{M}_{\text{opt},i}$. However, unlike in Algorithm 1 where all users directly update their cell/RAT selection if $\boldsymbol{\lambda}$ is updated the selection is only triggered once at the beginning of a service request or if the user would be dropped from the air interface where it is currently assigned to. For the selection, only local information ($\overline{R}_{i,m}$ can be measured or estimated for all BSs by a user) and the BS weights $\boldsymbol{\lambda}$ are needed similar to the static procedure. After a user selected a cell/RAT or in case that the request, the channel or the interference situation changed, an update of the resource assignment will be triggered in the corresponding base station. Thereby, the triggers are independent for each BS and no information from neighboring cells is needed for the resource assignment. Also, contrary to the static Algorithm 1, the resource update will not trigger the cell/RAT selection of users and users stay assigned to their current BS in general. Only in case a user cannot be supported by a BS anymore and no intrasystem hand-over is possible the user will execute Algorithm 2(a) again leading to a possible intersystem hand-over. The resource assignment in a cell will be updated following the iterative procedure in Algorithm 2(b). Algorithm 2(b) maximizes the sum utility of the BS over all BE users that are assigned to it and assures that all voice users comply with their minimum rate requirement. Thereby, the rates will be assigned in a way that all available resources are exploited and that the resource constraint of the BS is met with equality before $\boldsymbol{\lambda}$ is broadcasted again. This stands in clear contrast to the static algorithm where $\boldsymbol{\lambda}$ is updated based on the subgradient.

Since in Algorithm 2 each user only actively selects a RAT/cell once at its call setup and it does not trigger reassignments of other users in general almost no signaling information has to be exchanged between users and BSs. The simplicity of Algorithm 2 however, comes at the cost of its optimality. The influence of new users on λ, mobility, and the restriction that users stay in the actual air interface if possible lead to situations where a user j might find itself assigned to a BS $m \neq \mathcal{M}_{\text{opt},j}(\boldsymbol{\lambda})$. Wrong assignments will lead to deviations of $\boldsymbol{\lambda}$ and it cannot be guaranteed that the procedure approaches to $\boldsymbol{\lambda}^*$, which would be the optimum weights for the current request and channel situation in the static scenario. Since Algorithm 1 is difficult to implement in our simulation tool, we will derive a simple upper bound. The bound allows us to evaluate the maximum degradation of an assignment obtained with the dynamic procedure from the optimum solution of (P1). Since the bound overestimates (P1), it is also an upper bound for Algorithm 1 and could be used to evaluate the quality of the static Algorithm 1, which might be nonoptimal in case $|\mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda}^*)| > 1$.

4.3. Utility Bound. Assume that the dynamic algorithm approaches $\boldsymbol{\lambda}^+$ and a rate assignment $\mathbf{R}^\epsilon$ at a certain point in time. Then, there exists a corresponding dual function $g(\boldsymbol{\lambda}^+)$ which is an upper bound on (P1):

$$\begin{aligned} g(\boldsymbol{\lambda}^+) &= \max_{\mathbf{R}} \mathcal{L}(\mathbf{R}, \boldsymbol{\lambda}^+) = \mathcal{L}(\mathbf{R}^+, \boldsymbol{\lambda}^+) \geq \mathcal{L}(\mathbf{R}^*, \boldsymbol{\lambda}^*) \\ &\geq \mathcal{L}(\mathbf{R}^\epsilon, \boldsymbol{\lambda}^+) = \sum_{i \in \mathcal{I}_b} \psi_i \left(\sum_{m \in \mathcal{M}} R_{i,m}^\epsilon \right). \end{aligned} \tag{36}$$

Therefore, the deviation to the global optimum of a rate assignment $\mathbf{R}^\epsilon$ can be bounded by the difference of $\mathcal{L}(\mathbf{R}^+, \boldsymbol{\lambda}^+)$ and $\mathcal{L}(\mathbf{R}^\epsilon, \boldsymbol{\lambda}^+)$

$$\begin{aligned} \Delta \mathcal{L} = &\sum_{i \in \mathcal{I}_b \cap \mathcal{I}_\epsilon} \psi_i(R_{i,m}^+) - \psi_i(R_{i,m}^\epsilon) \\ &- \sum_{m \in \mathcal{M}} \lambda_m^+ \left[\sum_{i \in \mathcal{I}_\epsilon} \left(\frac{R_{i,m}^+ - R_{i,m}^\epsilon}{\overline{R}_{i,m}} \right) \right], \end{aligned} \tag{37}$$

```
(a) Cell/RAT Selection of user $i$.
(1) User $i$ measures the channels and evaluates $\overline{R}_{i,m}$ for all BS/RATs in its vicinity
(2) Based on the broadcasted $\boldsymbol{\lambda}$ user $i$ evaluates $\mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda})$ with (27) and sends an assignment request to $m \in \mathcal{M}_{\text{opt},i}$.
(b) Resource Assignment of BS $m$.
(1) Initialize $\nu_m, l = 1$ if not initialized: $\lambda_m = 1$
while $|\nu_m| > \epsilon$ do
    (2) For all users $i$ that are assigned to BS $m$ set $\mathcal{M}_{\text{opt},i} = m$ and calculate $R_{i,m}$ with (30)
    (3) BS $m$ evaluates its sub-gradient $\nu_m = (\Gamma_m - \sum_{i \in \mathcal{I}} (R_{i,m}/\overline{R}_{i,m}))$ and updates its dual weight
    $\lambda_m(l+1) = \lambda_m(l) - \delta(l)\nu_m; l = l+1$
end while
(3) Assign users $R_{i,m}$ corresponding to (2) and broadcast updated $\lambda_m$
```

Algorithm 2

with $\mathcal{I}_\epsilon = \{i \in \mathcal{I}, m_i' \notin \mathcal{M}_{\text{opt},i}(\boldsymbol{\lambda}^+)\}$. Only the rates $\mathbf{R}^+$ are needed for the evaluation of the bound which can be easily calculated by (30).

5. Simulation Results

In this section, the performance of Algorithm 2 will be evaluated by comparing it to a load-balancing algorithm. We therefore employ Alcatel-Lucent's C++ based MRRM-Simulator which is an event driven simulation environment for heterogeneous wireless scenarios. It supports cellular UMTS/HSDPA, GSM/EDGE air interfaces, a WiMAX hotspot, and different service classes such as VoIP, streaming, circuit-switched voice and best-effort data services. For the simulations we consider a 2-RAT scenario consisting of a cellular GSM/EDGE and UMTS air interface with 42 BSs each. The BSs of both RATs are arranged as indicated in Figure 1; on each site there are 3 BSs with directional antennas of both RATs collocated with the distance between sites being 2400 m. All RAT specific parameters are listed in Table 1. Equally distributed inside the rectangular movement area (see Figure 1), there are users that are moving corresponding to the pedestrian mobility model in [14] with 3 km/h and randomly requesting services based on a Poisson process with exponentially distributed service duration with a mean of 120 seconds. For voice services a constant data rate of 12.2 kbit/s is required while no minimum requirements for best-effort services exist.

The load-balancing strategy and Algorithm 2 differ only by the cell/RAT selection procedure which are triggered at a call setup or at an intersystem hand-over request. All other mechanisms like intrasystem hand-overs and the triggers themselves correspond to the standards and stay untouched. Both algorithms perform the resource assignment inside a BS corresponding to Algorithm 2(b) so that the sum utility of each BS is maximized. In case of load balancing a new user that requests service or an intersystem hand-over performs the cell/RAT selection as follows: at first it short-lists one BS of each air interface where the one with the strongest pilot signal that could accept the call in the users vicinity is selected. Usually, these are the closest UMTS and GSM BSs to the user. Then, the user sends the request to the BS with the lower load value. Hereby, the load values are obtained by signaling and are defined as $l_{v,m}$, $l_{b,m}$ in case of a voice or best-effort requests, respectively:

$$
l_{v,m} = \begin{cases} \sum\limits_{i \in \mathcal{I}_v} \dfrac{t_{i,m}}{\overline{T}_m} & \forall m \in \mathcal{M}_a,\ a \in \mathcal{A}_{\text{orth}}, \\[2ex] \sum\limits_{i \in \mathcal{I}_v} \dfrac{p_{i,m}}{\overline{P}_m} & \forall m \in \mathcal{M}_b,\ b \in \mathcal{A}_{\text{inf}}, \end{cases} \qquad (38)
$$

$$
l_{b,m} = \mathbb{E}_{i \in \mathcal{I}_b} \left(\frac{1}{R_{i,m}} \right) \qquad \forall m \in \mathcal{M}
$$

Table 1: Simulation parameters.

$P_{\max,UMTS} = 20\,\text{W}$
$P_{\max,GSM} = 15\,\text{W}$
Time slots GSM $\overline{T}_m = 21$
Antenna pattern: Sector 90° [14]
Path-loss GSM [dB] , r distance in m: $L = 132.8 + 38\,\text{lg}(r - 3)$ [15]
Path-loss UMTS [dB] : $L = 128.1 + 37.6\,\text{lg}(r - 3)$ [14]
Rate-SINR mapping UMTS: $C_b = 1.4e9$ $D_b = 1e - 3$
Thermal noise GSM, UMTS: -100 dBm
Intercell interference GSM: -105 dBm
Orthogonality factor UMTS: $\rho = 0.4$

For the UMTS air interface the used normalized resource-rate mapping curve and the linear approximation corresponding to (9) are shown in Figure 2. The slope of the linear approximation is chosen so that it intersects the real rate mapping curve at the origin and at 100 kbit/s, which corresponds to $\Delta_b = 1.53e6$ bit/s. For the GSM air interface, the envelope of the coding and modulation corresponding to [15] serves as SINR-rate mapping with the additional requirement from the standard that voice users are not able to share a time slot with other users. As utility curve, a shifted version of the α-proportional fair curve with $\alpha = 1/2$ is used, which is a more throughput oriented metric:

$$\psi(R_i) = \sqrt{\frac{R_i}{\text{bit/s}} + 1000} - \sqrt{1000}. \tag{39}$$

The shifting operation leads to a finite slope of the curve at the origin which is essential to enable switching off users. Otherwise, a user in a deep fade might be assigned almost all resources, if $\lim_{x\to 0}\psi'(x) = \infty$.

In the simulation scenario, there are in average 10 voice service call setup requests per second inside the movement area which corresponds to approximately 36 active voice users and a voice traffic load of 440 kbit/s per cell area in average. Additionally, a varying number of BE users request service. For the simulation statistics, only the investigated cells (see Figure 1) are considered. In Figure 3, the throughput of the BE users based on the real SINR-rate mapping and the approximation is shown over the average number of active BE users. As can be observed, Algorithm 2 achieves up to 30% more throughput compared to load-balancing. The real and approximated rates match pretty well in the region for low user request rates, but also at high load the deviation is small compared to the gain. The sum utility per cell area and the upper bound are shown in Figure 4. The utility gain of Algorithm 2 compared to load-balancing is also almost as large as of the throughput because of the low curvature of ψ. The distance to the bound is of special interest; at high call arrival rates the distance is almost zero, indicating that Algorithm 2 performs close to optimum and no significant gains could be achieved by using Algorithm 1 instead. At lower rates this is different. Here, the dynamic procedure pays the price for its simplicity in terms of performance loss. The main reason for the loss results from the fluctuation of $\boldsymbol{\lambda}$. At low request rates a user's call setup or service termination has a great impact on the resource allocation of the other users in the cell and therefore leads to strong variations of $\boldsymbol{\lambda}$ over time. The fluctuation of $\boldsymbol{\lambda}$ directly influences the set of optimum BSs $\mathcal{M}_{\text{opt}}$ of users and therefore often leads to the case that users find themselves assigned to a currently nonoptimal BS. In this case, the dynamic algorithm looses performance since the cell selection is only allowed once per user in general. Higher utility values could be obtained here by allowing users to perform intersystem hand-overs so that each user would be assigned to $\mathcal{M}_{\text{opt}}$ again. This characteristic is also reflected in the looseness of the bound. Unlike to low request rates, if the average number of users in a cell is high the influence of a single-user arrival or departure from a cell on $\boldsymbol{\lambda}$ is diminishing and a user's optimum BS hardly changes during the service time. In this case the performance is almost optimal and the bound gets very tight. The tightness also indicates that the influence of the users pedestrian mobility and therefore the variation of $\overline{R}$ (and on $\mathcal{M}_{\text{opt}}$) is negligible in this scenario.

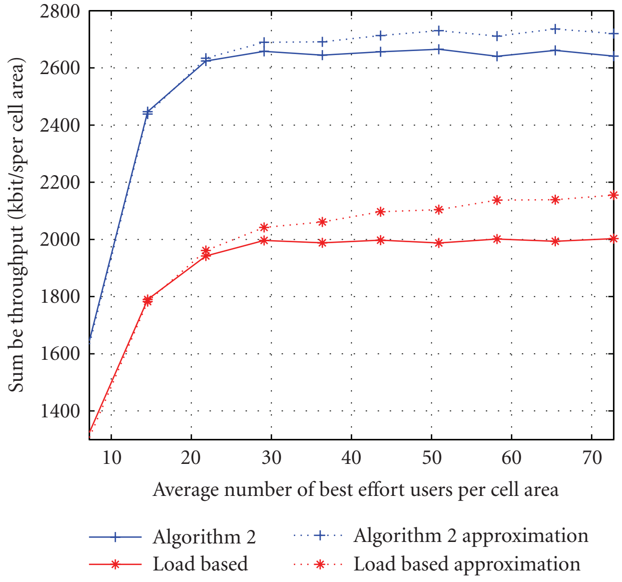


Figure 3: BE throughput with and without linear approximation (9) without slow fading.

For the heterogeneous UMTS GSM/EDGE system the following interpretation of the optimum assignment strategy can be given. One observes that $\overline{R}$ is a monotonically increasing function of a user's SINR for both air interfaces. Therefore, for a given $\boldsymbol{\lambda}$ the optimum cell/RAT selection $m_{\text{opt},i} = \arg\min_m \lambda_m / \overline{R}_{i,m}(\beta_{i,m})$ reduces to an SINR threshold. This threshold depends on the air interface and the service type through $\overline{R}(\beta)$ and on λ which can be interpreted as the load situation of the BS. The threshold characteristic can be observed in Figure 5, where the BE user assignment in terms of the selected RAT is shown by color shades; Algorithm 2 assigns users to UMTS that are in the red area close to the BSs and users in the blue area to GSM. The border of both areas is characterized by the threshold SINR of each RAT which has a lobe pattern because of the directional antenna characteristics. The pattern looks very regular in Figure 5 due to equal average loads in each cell of an air interface (and therefore equal λ for BSs of one RAT) and collocated sites of UMTS and GSM BSs. However, Algorithm 2 will also flexibly adapt itself to the optimum configuration in case of arbitrary, not necessary collocated, BS positioning and varying load situations without any change in configuration of the algorithm. The optimum area pattern will then of course look different. Contrary to the BE users Algorithm 2 will assign almost all voice users to UMTS in the presented scenario. This is due to the fact that time-slot sharing is not possible in GSM for voice users. Therefore, the maximum slot rate of a voice user is much lower than in UMTS. Thus, a much lower λ of the GSM BS compared to the λ of the UMTS BS would be required to make GSM attractive for an assignment. This instance might suggest that also the major part of the gain of Algorithm 2 is based on the low effectivity of voice in GSM, which is not avoided in load balancing. Simulations however show that also for pure BE traffic gains of more than 20% are obtained.

So far slow fading has not been active in the simulations to demonstrate that the utility bound can be tight and to visualize the assignment policy of Algorithm 2 qualitatively. In Figure 6, the sum utility and the bound is shown for the scenario above however this time with slow fading

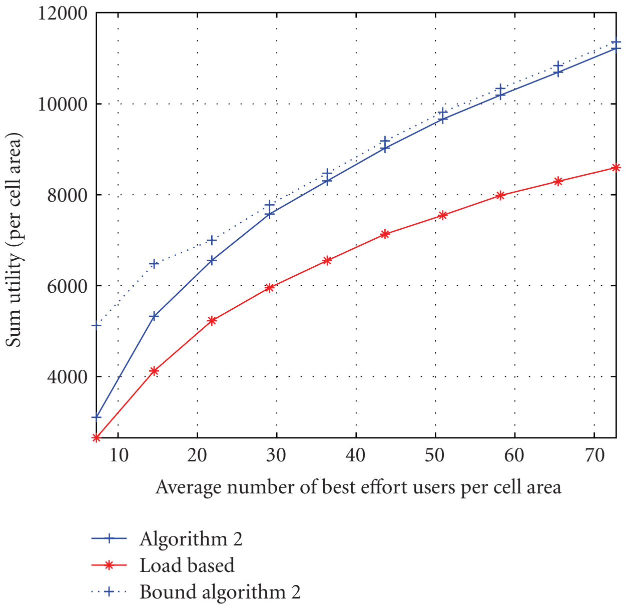


Figure 4: Sum utility $U = \sum_{i \in \mathcal{I}_b} \psi_i(R_i)$ and upper bound $U + \Delta\mathcal{L}$ without slow fading.

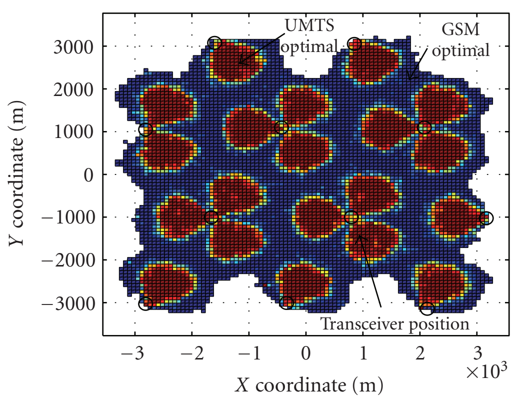


Figure 5: RAT assignment of BE users without slow fading: 1 → 100% assigned to UMTS 0 → 100% assigned to GSM.

corresponding to [14] in both air interfaces with a variance of 6 dB. Considering load balancing, the slow fading does hardly influence the performance. For Algorithm 2 however the users' mobility in connection with the slow fading has a nonnegligible impact. Now, even small changes in position can result in large channel gain and therefore $\overline{R}$ differences which lead to more wrongly assigned users and looseness of the bound. Nevertheless, still a gain of approximately 20% is achieved. Similarly the performance of Algorithm 2 decreases and the bound gets less tight without slow fading in case the velocity is increased. For completeness, it is noted here that in case users do not change their position the tightness of the bound under slow fading is similar to Figure 4.

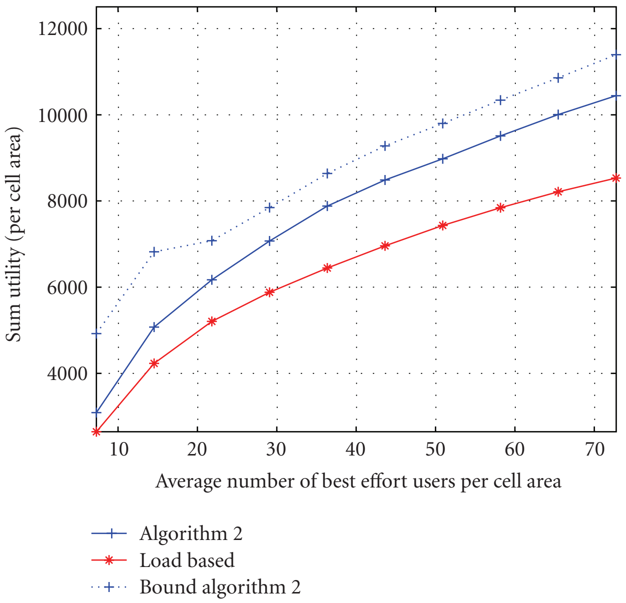


Figure 6: Sum utility and upper bound with slow fading 6 dB.

The observations made in Section 3 and in the simulations open up the way for even more simplified algorithms that might be interesting for practical applications. For given scenarios fixed base station weights $\boldsymbol{\lambda}$ or service dependent SINR, channel or even distance thresholds could be applied for the cell/RAT selection or as triggers for intersystem handovers. Additionally, in case users are subject to strong channel variations, for example, by mobility or fading during a service request updating the cell/RAT selection and therefore executing Algorithm 2(a) at more frequent intervals is an option to improve the performance and get close to the optimum again.

6. Conclusions

In this paper, we developed an optimization framework for wireless heterogeneous multicell scenarios. Having derived the feasible rate regions for air interfaces with orthogonal resource assignment and a convex approximation for interference limited radio access technologies we introduced a convex utility maximization problem formulation for heterogeneous scenarios. We gained general insights on the problem solution and derived simple assignment rules that lead to the global optimum by exploiting the dual problem formulation. These observations were then used to develop decentralized algorithms for static scenarios and then simplified for dynamic settings. Although the simplifications came at the cost of the optimality still high gains in comparison to a simple load-balancing algorithm were obtained and close to optimum performance could be shown by simulations based on a duality bound.

Acknowledgment

The authors are supported in part by the *Bundesministerium für Bildung und Forschung (BMBF)* under Grant FK 01 BU 566.

References

[1] J. Pérez-Romero, O. Sallent, and R. Agustí, "On the optimum traffic allocation in heterogeneous CDMA/TDMA networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 9, pp. 3170–3174, 2007.

[2] A. Furuskär and J. Zander, "Multiservice allocation for multiaccess wireless systems," *IEEE Transactions on Wireless Communications*, vol. 4, no. 1, pp. 174–183, 2005.

[3] I. Blau and G. Wunder, "User allocation in multi-system, multi-service scenarios: upper and lower performance bound of polynomial time assignment algorithms," in *Proceedings of the 41st Annual Conference on Information Sciences and Systems (CISS '07)*, pp. 41–46, Baltimore, Md, USA, March 2007.

[4] I. Blau, G. Wunder, I. Karla, and R. Siegle, "Cost based heterogeneous access management in multi-service, multi-system scenarios," in *Proceedings of the 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, pp. 1–5, Athens, Greece, September 2007.

[5] E. Stevens-Navarro, Y. Lin, and V. W. S. Wong, "An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1243–1254, 2008.

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

[7] S. Stańczak, M. Wiczanowski, and H. Boche, "Distributed utility-based power control: objectives and algorithms," *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 5058–5068, 2007.

[8] M. Chiang, "Balancing transport and physical layers in wireless multihop networks: jointly optimal congestion control and power control," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 104–116, 2005.

[9] J. Huang, R. A. Berry, and M. L. Honig, "Distributed interference compensation for wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 5, pp. 1074–1084, 2006.

[10] A. Goldsmith, *Wireless Communications*, Cambridge University Press, New York, NY, USA, 2005.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.

[12] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, Mass, USA, 2nd edition, 1995.

[13] S. Borst, A. Proutière, and N. Hegde, "Capacity of wireless data networks with intra- And inter-cell mobility," in *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM '06)*, pp. 1–2, Barcelona, Spain, April 2006.

[14] ETSI, "Selection procedures for the choice of radio transmission," Tech. Rep. 101 112 V3.1.0, UMTS, November 2001.

[15] ETSI, "Radio network planning aspects," Tech. Rep. 101 362 V8.3.0, GSM, 1999.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 564692, 15 pages
doi:10.1155/2009/564692

Research Article

Joint Throughput Maximization and Fair Uplink Transmission Scheduling in CDMA Systems

Symeon Papavassiliou[1, 2] and Chengzhou Li[3]

[1] *Network Management and Optimal Design Laboratory (NETMODE), Institute of Communications and Computer Systems (ICCS), 9 Iroon Polytechniou Street, Zografou 157 73, Athens, Greece*
[2] *School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), 9 Iroon Polytechniou Street, Zografou 157 73, Athens, Greece*
[3] *LSI Corporation, 1110 American Parkway NE, Allentown, PA 18109, USA*

Correspondence should be addressed to Symeon Papavassiliou, papavass@mail.ntua.gr

Received 9 July 2008; Revised 10 December 2008; Accepted 20 February 2009

Recommended by Alagan Anpalagan

We study the fundamental problem of optimal transmission scheduling in a code-division multiple-access wireless system in order to maximize the uplink system throughput, while satisfying the users quality-of-service (QoS) requirements and maintaining fairness among them. The corresponding problem is expressed as a weighted throughput maximization problem, under certain power and QoS constraints, where the weights are the control parameters reflecting the fairness constraints. With the introduction of the power index capacity, it is shown that this optimization problem can be converted into a binary knapsack problem, where all the corresponding constraints are replaced by the power index capacities at some certain system power index. A two-step approach is followed to obtain the optimal solution. First, a simple method is proposed to find the optimal set of users to receive service for a given fixed target system load, and then the optimal solution is obtained as a global search within a certain range. Furthermore, a stochastic approximation method is presented to effectively identify the required control parameters. The performance evaluation reveals the advantages of our proposed policy over other existing ones and confirms that it achieves very high throughput while maintains fairness among the users, under different channel conditions and requirements.

1. Introduction

The continuous growth in traffic volume and the emergence of new services have begun to change the structure and requirements of wireless networks. Future mobile communication systems will be characterized by high throughput, integration of services, and flexibility [1–5]. With the demand for high data rate and support of multiple quality of service (QoS), the transmission scheduling plays a key role in the efficient resource allocation process in wireless systems. The transmission scheduling determines the time instances that a mobile user may receive service, as well as the resources that should be allocated to support the requested service, in order to make the resource distribution fair and efficient.

The fundamental problem of scheduling the users transmission and allocating the available resources in a realistic uplink code-division multiple-access (CDMA) wireless system that supports multirate multimedia services, with efficiency and fairness, is investigated and analyzed in this paper. A transmission scheduling method which achieves the maximum system throughput under the constraints of satisfying certain users QoS requirements and maintaining throughput fairness among them is provided and evaluated.

1.1. Related Work and Motivation. A class of scheduling schemes, namely, the opportunistic scheduling schemes, has been proven to be an effective approach to improve the system throughput by utilizing the multiuser diversity effect [6, 7] in wireless communications. Specifically, for a system with many users that have independent varying channels, with high probability there is a user with channel much stronger than its average SNR requirement. Therefore, the system throughput may be maximized by choosing the user with "relatively best" channel for transmission at a given slot. However, some fairness constraints must be imposed on the scheduling policies to ensure the fair resource allocation.

It has been shown in [8] that scheduling users one-by-one can result in higher system throughput for high data rate traffic in the CDMA downlink. However, this work does not exploit the time-varying channel conditions. In [7, 9], a high-speed data rate scheme (HDR) is introduced, where the base station schedules the downlink transmission of a single user at a given time slot with the data rates and slot lengths varying according to the specific channel condition. In [10–12], a transmission scheduling scheme for multiple users, which considers both the channel condition and queueing delay/length, is proposed and shown to be throughput optimal if it is feasible. However, the fairness issue is not explicitly addressed in that work. In [13–15], a framework for opportunistic scheduling that maximizes the system performance by exploiting the time-varying channel conditions of wireless networks is presented. Three categories of scheduling problems—the temporal fairness, utilitarian fairness, and minimum-performance guarantee scheduling—are studied and optimal solutions are given.

Although the downlink transmission assignment is important for several applications, the efficient uplink transmission scheduling plays an important role as well, especially with the prevailing of multimedia communications and applications. It has been argued that the downlink scheduling method is not suitable to be applied to the uplink transmission scheduling, where simultaneous transmissions may result in higher throughput [16, 17]. The uplink transmission scheduling problem is more complicated and requires further consideration of additional elements to make the corresponding scheduling policies feasible [18]. The achievable throughput in such a case depends not only on the service access time, but also on the transmission powers and the corresponding users interference. In addition, multiple users can be scheduled simultaneously to transmit in the same time slot, which is a major difference from the wireline and TDMA-like scheduling schemes, making the respective scheduling processes either inapplicable or inefficient in the CDMA environment. The simple temporal fairness scheduling, where the main resource to be shared is the time, fails to provide rational fairness in this case. As a result, the throughput optimal and fair uplink transmission scheduling problem needs to jointly consider multiple factors such as access time, transmission power, channel conditions, and number of users to be scheduled at the same time. Heuristic approaches to address the problem of short-term fairness and demonstrate the tradeoff between fairness and throughput under some special cases have been introduced in [19–21].

Furthermore, how to maximize the throughput of uplink CDMA system was first analyzed in [16]. The sole purpose of uplink throughput maximization can be achieved by choosing the "best" K users in terms of their received power, when they transmit at their maximum power. However, such throughput maximization does not consider fairness, that is, the equal opportunity for all users to receiving service despite their channel conditions. Therefore, among the objectives of our approach in this paper is to identify the actual "best" users that should transmit in order to maximize the throughput, when the fairness constraints are introduced and respected.

In [22], several scenarios of scheduling uplink CDMA transmission with voice and data services are analyzed. With the number of voice users and their corresponding transmission rates fixed, that work attempted to maximize the throughput of data service. It was shown that when the synchronization overhead is reasonable, a smaller number of simultaneous transmission users achieve higher system throughput and at the same time lower the average transmission power. However, in this case the achievable throughput is affected by the "weakest link." Therefore, this approach can be regarded only as a static analysis that considers the relationship between the performance and the number of users chosen for transmission. The problem of identifying the actual set of users to transmit based on their channel conditions, which may reduce the impact of the "weakest link", has not yet been investigated and is one of the main objectives of our paper.

In addition, the problem of uplink CDMA scheduling is further complicated by the fact that the conventional concept of capacity used in the wireline networks, for example, total bandwidth of the physical media, is not directly applicable in the CDMA systems. In this case, the actual system capacity is not fixed and known in advance, since it is a function of several parameters such as the number of users, the channel conditions, and the transmission powers.

Therefore, in summary the main contributions of this paper are as follows. (1) Jointly consider the factors of channel capacity, number of users and their interference, transmit power, and fairness requirements. (2) Formulate an optimization problem that stresses the fairness requirement under time-varying wireless environment and proves the existence of an optimal solution based on all constraints. (3) Exploit the power index concept and power index capacity, as a novel and effective way, to treat the fairness issue in the transmission scheduling policy under the considered uncertain and dynamic environment. (4) Devise a scheduling policy that achieves throughput fairness among the users and optimal system throughput under certain constraints.

1.2. Paper Outline. The rest of the paper is organized as follows. In Section 2, the system model that is used throughout our analysis is described, and the problem of the uplink scheduling in CDMA systems is rigorously formulated as a multiconstraint optimization problem. It is demonstrated that this problem can be expressed as a weighted throughput maximization problem, under certain power and QoS constraints, where the weights are the control parameters that reflect fairness constraints. Based on the concept of power index capacity, this optimization problem is converted into a simpler linear knapsack problem in Section 3.1, where all the corresponding constraints are replaced by the users power index capacities at some certain system power index. The optimal solution of the latter problem is identified in Sections 3.2 and 3.3, while in Section 3.4, a stochastic approximation method is presented in order to effectively identify the required control parameters. Section 4 contains the performance evaluation of the proposed method, along with some numerical results and discussion, and finally Section 5 concludes the paper.

2. System Model and Problem Formulation

In this paper, we consider a single cell DS-CDMA system with $B(k)$ backlogged users at time slot k. The users channel conditions are assumed to change according to some stationary stochastic process, while the uplink transmission rate is assumed to be adjustable with the variable spreading gain technique [23]. Each user i is associated with some preassigned weight ϕ_i according to its QoS requirement. In the following for simplicity in the presentation, we omit the notation of the specific slot k from the notations and definitions we introduce. Let us denote by r_i the transmission rate of user i in the slot under consideration. We assume that the chip rate W for all mobiles is fixed, and hence the spreading gain G_i of user i is defined as $G_i = W/r_i$. Let us also denote by γ_i the required signal-to-interference and noise ratio (SINR) level of user i, by h_i the corresponding channel gain, and by p_i the user i transmission power at a given slot, which, however, is limited by the maximum power value $p_i^{\max}$. Therefore, the received SINR γ_i' for a user i is given by

$$\frac{h_i p_i G_i}{\alpha \sum_{j=1, j \neq i}^{B(k)} h_j p_j + W\eta_0} = \gamma_i', \quad i = 1, 2, \ldots, B(k), \tag{1}$$

where η_0 is the one-sided power spectral density of additive white Gaussian noise (AWGN), and α determines the proportion of the interference from other users received power. Without loss of generality in the following, we assume $\alpha = 1$. Obviously, to meet the SINR requirement, the received SINR γ_i' has to be larger than the corresponding threshold γ_i, that is, $\gamma_i' \geq \gamma_i$. In the following, we assume perfect power control in the system under consideration, while users are scheduled to transmit at the beginning of every fixed-length slot. The objective of the optimal scheduling policy Q^* is to find the optimal number of allowable users and their transmission rates, which achieves the maximum system throughput while maintaining the fairness property.

2.1. Problem Formulation. Let $R(k) = \sum_{i=1}^{B(k)} r_i(k)$ denote the total throughput in slot k. Our objective function is to maximize the expectation of $R(k)$ by selecting the optimal transmit power vector $(p_1, p_2, \ldots, p_{B(k)})$ and transmit rate vector $(r_1, r_2, \ldots, r_{B(k)})$, that is,

$$\max E\left(\sum_{i=1}^{B(k)} r_i\right) \tag{2}$$

subject to specific SINR, maximum transmit power, and fairness constraints as follows:

$$\begin{gathered} \frac{h_i p_i G_i}{\sum_{j=1, j \neq i}^{B(k)} h_j p_j + W\eta_0} \geq \gamma_i, \quad \text{for } i = 1, 2, \ldots, B(k), \\ p_i \leq p_i^{\max}, \quad \text{for } i = 1, 2, \ldots, B(k), \\ \frac{\overline{r}_i}{\phi_i} = \frac{\overline{r}_j}{\phi_j} \quad \text{for } 1 \leq i,\ j \leq B(k), \end{gathered} \tag{3}$$

where $\overline{r}_i = E(r_i)$ denotes the mean throughput of user i in the corresponding backlogged period. It has been shown in [15, 24] that the above-constrained optimization problem can be considered as equivalent to the following problem (4), where Z is the minimal value among all $\overline{r}_i/\phi_i$, that is, $Z = \min_i\{\overline{r}_i/\phi_i\}$. In (4), we transform the objective function (2) into finding the optimal transmit powers and rates that maximize the minimal normalized average rate Z. Therefore,

$$\begin{gathered} \max Z, \\ \text{s.t.} \quad Z \leq \frac{\overline{r}_i}{\phi_i}, \quad 1 \leq i \leq B(k), \\ \frac{h_i p_i W/r_i}{\sum_{j=1, j \neq i}^{B(k)} h_j p_j + W\eta_0} \geq \gamma_i \quad i = 1, 2, \ldots, B(k), \\ p_i \leq p_i^{\max}, \quad 1 \leq i \leq B(k). \end{gathered} \tag{4}$$

Apparently, the solution of the above problem will finally make $Z = \overline{r}_i/\phi_i$ for $1 \leq i \leq B(k)$ since one can always reduce its throughput for the benefit of other users in order to maximize Z. With the constraint $Z = \overline{r}_i/\phi_i$, the objective function then is generalized to

$$\max \sum_{i=1}^{B(k)} w_i \overline{r}_i, \tag{5}$$

where w_i is an arbitrary positive number. Here the crucial observation [24] is that the optimal scheduling policy will be the one that maximizes the sum of weighted throughputs and equalizes the normalized throughput. The maximization of mean-weighted rate in (5) is obtained by the maximization of the weighted rate in every slot, that is, $\max\sum_{i=1}^{B(k)} w_i r_i$ for every slot k. In conclusion, to obtain the optimal uplink throughput while keeping fairness, we must solve the following problem:

$$\max \sum_{i=1}^{B(k)} w_i r_i, \tag{6}$$

$$\text{s.t.} \quad \frac{h_i p_i W/r_i}{\sum_{j=1, j \neq i}^{B(k)} h_j p_j + W\eta_0} \geq \gamma_i, \quad i = 1, 2, \ldots, B(k), \tag{7}$$

$$p_i \leq p_i^{\max}, \quad 1 \leq i \leq B(k). \tag{8}$$

The fairness constraint, that is, $\overline{r}_i/\phi_i = \overline{r}_j/\phi_j$, is represented by the choice of w_i. By adjusting the value of w_i, the user will get more or less opportunities to transmit data, and hence the corresponding normalized throughput is balanced. As we discuss later in this paper, the value of w_i can be approximated by a stochastic approximation algorithm, which has already found its application in [14, 15] under similar situations. Note that since we assume perfect power control in the CDMA system under consideration, only the equality case of (7) is considered here.

The following Proposition 1 states that the optimal solution is achieved when a user either transmits at full power or does not transmit at all.

Proposition 1. *The optimal solution that maximizes the weighted throughput of problem* (6) *is such that*

$$p_i(k) \in \{0, p_i^{\max}\}, \quad \text{for } i = 1, 2, \dots, B(k). \tag{9}$$

Proof. In order to minimize the multiple access interference, users transmit with the minimum required power to meet the required threshold γ_i. Therefore, we consider the equality case of constraint (7). To maintain exactly the threshold γ_i for user i, the achievable transmit rate is represented as

$$r_i(k) = \frac{h_i p_i W}{\gamma_i\left(\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0\right)}. \tag{10}$$

The objective function then becomes

$$Z = \sum_{i=1}^{B(k)} w_i r_i = \sum_{i=1}^{B(k)} \frac{w_i h_i W}{\gamma_i} \frac{p_i}{\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0}. \tag{11}$$

Differentiating twice with respect to the transmit power of a user m, we obtain

$$\frac{\partial^2 Z}{\partial p_m^2} = 2\sum_{i=1, i\neq m}^{B(k)} \frac{w_i h_i W}{\gamma_i} \frac{p_i h_m^2}{\left(\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0\right)^3}. \tag{12}$$

Since w_i is positive number, obviously (12) is nonnegative, while the objective function is a convex function of p_m. Hence, the optimal solution of this problem is that the transmit power obtains the value of its boundary, that is, either 0 or $p_i^{\max}$. □

In Section 3, the corresponding optimization problem is transformed to an equivalent problem of a simpler form, which facilitates the identification of the optimal solution. However, in the following we first introduce the concept of power index capacity which is used to represent the corresponding constraints, under the problem transformation.

2.2. Power Index Capacity. It has been shown in [25] that by solving the constraints (7) and (8), the following inequality must be satisfied if there exists a feasible power assignment $\mathbf{p} = [p_1, p_2, \dots, p_{B(k)}]$ that meets the QoS requirements:

$$\begin{aligned}\sum_{i=1}^{B(k)} g_i &\leq 1 - \frac{\eta_0 W}{\min_{1\leq i\leq B(k)}\{p_i^{\max} h_i (G_i/\gamma_i + 1)\}} \\ &= 1 - \frac{\eta_0 W}{\min_{1\leq i\leq B(k)}\{p_i^{\max} h_i / g_i\}},\end{aligned} \tag{13}$$

where

$$g_i = \frac{\gamma_i}{\gamma_i + G_i} \tag{14}$$

is defined as the power index of user i [26]. Relation (13) is the necessary and sufficient condition such that a power and rate solution is feasible under constraints (7) and (8) [25].

Let us regard $\sum_i g_i$ as the actual system load, which is the sum of power indices assigned to all backlogged users, while we assume that there is a target system load ψ. It should be noted that ψ here is not fixed but has value $0 \leq \psi < 1$. The meaning and motivation for the definition of the target system load ψ are that the system will attempt to provide the appropriate scheduling in order to make the actual system load $\sum g_i$ reach the target load (however, it serves as an upper bound and cannot be exceeded). For an arbitrarily selected ψ in the range of $0 < \psi < 1$, there exist two possible cases concerning the relationship between the actual system load $\sum g_i$ and the target system load. When considering small values for the target system load ψ, the system can easily make the actual system load $\sum g_i$ reach the target load under consideration, that is, $\sum g_i = \psi$. On the other hand, when ψ is large, especially when it approaches to 1, it may be impossible for the actual achievable system load $\sum g_i$ to reach ψ due to the limitation imposed by (13). Let us assume that in time slot k the maximum system load this system can achieve based on all users channel states and all possible user schedulings is $\psi^* = \max(\sum g_i)$. We will now consider the two cases mentioned above, that is, $0 < \psi \leq \psi^*$ and $\psi > \psi^*$.

2.2.1. Target Load Is Less than or Equal to Maximum System Load. If we assume $0 < \psi \leq \psi^*$, then the system load can achieve the target load, $\sum_i g_i = \psi$. Therefore, (13) can be rewritten as follows:

$$\begin{gathered}\min_{1\leq i\leq B(k)}\left\{\frac{p_i^{\max} h_i}{g_i}\right\} \geq \frac{\eta_0 W}{1-\psi}, \quad g_i \leq \psi, \\ \text{therefore } \frac{p_i^{\max} h_i}{g_i} \geq \frac{\eta_0 W}{1-\psi} \quad \forall i,\ 1 \leq i \leq B(k).\end{gathered} \tag{15}$$

For each individual user, there is a limitation on the maximum power index that it can reach, given by (15)

$$g_i \leq (1-\psi)\frac{p_i^{\max} h_i}{\eta_0 W}, \quad g_i \leq \psi. \tag{16}$$

2.2.2. Target Load Is Larger than Maximum System Load. If the target load is larger than the maximum system load, that is, $\psi > \psi^*$, it means there will be no feasible transmission power solution in (7) and (8) to achieve this target load and therefore the relationship in (15) does not hold any more. In this case, we simply apply the power index restriction of (16) to each user. The consequence is that the final achieved system load becomes $\sum_i g_i < \psi^* < \psi$ since $g_i \leq (1-\psi)p_i^{\max} h_i/\eta_0 W < (1-\psi^*)p_i^{\max} h_i/\eta_0 W$.

In fact, unless all possible transmission user sets are searched, it is unknown in advance whether or not the actual system load $\sum g_i$ can reach the chosen ψ. Therefore, applying (16) to the case $\psi > \psi^*$ unifies the definition of the power index range, within which a user can be assigned a feasible power index without knowing the value of ψ^*. One key principle and rule regarding the algorithm proposed in this paper is to assign to an individual user a power index that is less than or equal to its power index capacity. In the power index assignment algorithm described in Section 3.2, the situation where $\sum g_i < \psi$ may occur. However, it should be noted here that as proven by Theorem 1 later in the paper, the global optimal solution must be the one satisfying $\sum g_i = \psi$. The target load range where $\psi > \psi^*$ is then not possible to be the optimal solution. The intentionally introduced restriction of (16) in the case of $\psi > \psi^*$ allows the algorithm to rule out such values of ψ due to the fact that $\sum g_i < \psi$ in this case.

2.2.3. Definition of Power Index Capacity. Hence, given the system load ψ the maximum possible power index g_i a user can accept in (15) is determined by the maximum transmit power $p_i^{\max}$ and the channel gain h_i.

Definition 1. In a CDMA system with $B(k)$ backlogged users at time slot k, given the target system power index ψ, the maximum power index that does not violate (13) for a single user whose channel gain is h_i is defined as the power index capacity (PIC) $\pi_i(h_i, \psi)$ of this user.

From (15), it can be easily found that the PIC of user i is

$$\pi_i(h_i, \psi) = \min\left\{(1-\psi)\frac{p_i^{\max} h_i}{\eta_0 W}, \psi\right\}. \tag{17}$$

Note that in (17) the power index capacity is limited by the target system power index. This is reasonable since a power index capacity that is greater than ψ will have no practical meaning and application. Furthermore, since our focus in this paper is to find an optimal scheduling policy as well as the optimal system load ψ, the value of ψ in (17) is not determined in advance.

Intuitively, the power index represents the relationship between the transmission power and the corresponding interference that is caused to other users. If we considered that the total system power index is fixed to ψ, larger power index g_i for user i indicates that it has relatively higher signal-to-interference ratio compared to the other users with smaller power index, while at the same time it causes more interference to them. Accordingly, users with high-power indices may lower their transmission power to reduce the interference they may cause, which in turn means that they will have smaller power index to limit the intracell interference of the system, and therefore satisfy (13) that guarantees the existence of a feasible transmission power solution.

3. Problem Transformation and Optimal Solution

3.1. Problem Transformation. The corresponding constraints in terms of the power index can be represented as follows:

$$\max Z = \sum_{i=1}^{B(k)} w_i f_r(g_i, \gamma_i), \tag{18}$$

$$\sum_{i=1}^{B(k)} g_i \le \psi, \tag{19}$$

$$g_i \le \pi_i(h_i, \psi), \quad 1 \le i \le B(k), \tag{20}$$

$$0 \le \psi < 1. \tag{21}$$

Note that in the objective function we represent the rate $r_i = f_r(g_i, \gamma_i)$ as a function of power index g_i, where

$$f_r(g_i, \gamma_i) = \frac{g_i}{1-g_i}\frac{W}{\gamma_i}, \tag{22}$$

which converts the power index into transmission rate and can be easily derived from (14) by replacing G_i with W/r_i.

In the following, let $\mathbf{V} = \{v_1, v_2, \ldots, v_i, \ldots\}$ denote the set that contains all the power and rate vectors that satisfy constraints (7) and (8) and $v_i = \{p_{1,i}, p_{2,i}, \ldots, p_{B(k),i}, r_{1,i}, r_{2,i}, \ldots, r_{B(k),i}\}$. The elements $p_{j,i}$ and $r_{j,i}$ represent the transmit power and rate of the jth user in the ith vector. Similarly, we define another set $\mathbf{V}'$ containing the power and rate vectors v_i' that satisfy constraints (19), (20), and (21). By definition, it is obvious that any power and rate vector $v_i \in \mathbf{V}$ is feasible. However, since in constraint (21), ψ may take a value, that is, close to 1 , the required transmit power could also accordingly become larger than maximum allowable transmit power $p_i^{\max}$ if we simply look at the result from (15). The following proposition states that if perfect power control is assumed, for any rate (or power index) vector that satisfies constraints (19), (20), and (21), there always exists a feasible transmit power vector.

Proposition 2. *If the power index assignment for all $B(k)$ backlogged users satisfies constraints* (19), (20), *and* (21), *there always exists a feasible transmit power assignment, that is, $p_i < p_i^{\max}$ for $1 \le i \le B(k)$.*

Proof. Let vector $\mathbf{g} = \{g_1, g_2, \ldots, g_{B(k)}\}$ be the power index vector that satisfies constraints (19), (20), and (21). Denote $\psi = \sum_{i=1}^{B(k)} g_i$ the sum of all power indices in vector $\mathbf{g}$. From the definition of power index capacity, the power index capacity of each user is $\pi_i(h_i, \psi)$ and $g_i \le \pi_i(h_i, \psi)$. Based on Definition 1 and (17), we have the following relation:

$$\psi \le 1 - \frac{\eta_0 W \cdot \pi_i(h_i, \psi)}{p_i^{\max} h_i} \le 1 - \frac{\eta_0 W \cdot g_i}{p_i^{\max} h_i}. \tag{23}$$

Hence, for any user i, the transmit rate may be chosen within range

$$p_i^{\max}\frac{g_i}{\pi_i(h_i, \psi)} \le p_i \le p_i^{\max}, \tag{24}$$

which still satisfies the above inequality and proves this proposition. The power control of the CDMA system will choose the minimal transmit power, that meets the required SINR. □

The following proposition proves that the two sets **V** and **V**′ contain the same elements, which means that (19), (20), (21) and (7), (8) impose the same constraints over our problem.

Proposition 3. *Any vector $v_i \in \mathbf{V}$ is also included in set $\mathbf{V}'$, while any vector $v_i' \in \mathbf{V}'$ is also included in set $\mathbf{V}$.*

Proof. Suppose that $v_i \in \mathbf{V}$, and therefore it satisfies constraints (7), (8). It is apparent that $p_{j,i} \leq p_j^{\max}$. Since, as shown earlier, constraints (7), and (8) can also be represented by (13) [25], v_i also satisfies (13). Using function (22), we can convert the rate vector $\{r_{1,i}, r_{2,i}, \ldots, r_{B(k),i}\}$ into the corresponding power index vector $\{g_{1,i}, g_{2,i}, \ldots, g_{B(k),i}\}$. Let $\psi = \sum_{j=1}^{B(k)} g_{j,i}$. For a feasible power and rate vector, with known ψ ($0 \leq \psi < 1$ [25]), we can find each user power index capacity $\pi_j(h_j, \psi)$. Since v_i satisfies (13), based on Proposition 2 and the definition of power index capacity, we conclude that $g_{j,i} \leq \pi_j(h_j, \psi)$. That means that the assigned powers and rates in v_i also satisfy the constraints (19), (20), and (21). Therefore, $v_i \in \mathbf{V}'$.

Let us consider vector $v_i' = \{p_{1,i}', p_{2,i}', \ldots, p_{B(k),i}', r_{1,i}', r_{2,i}', \ldots, r_{B(k),i}'\} \in \mathbf{V}'$. As before, the rate vector part can be converted to corresponding power index vector $\{g_{1,i}', g_{2,i}', \ldots, g_{B(k),i}'\}$. Let $\psi = \sum_{j=1}^{B(k)} g_{j,i}'$ and hence $g_{j,i}' \leq \pi_j'(h_j, \psi)$ due to constraints (19), (20), and (21). Note that for the case where $\psi' > \sum_{j=1}^{B(k)} g_{j,i}'$, $\pi_j'(h_j, \psi) \geq \pi_j'(h_j, \psi')$. Based on the previous discussion, we can easily conclude that the power vector is feasible. Therefore,

$$\psi \leq 1 - \frac{\eta_0 W \cdot g_{j,i}'}{p_{j,i}' h_j}, \tag{25}$$

which satisfies (13), for user j, $1 \leq j \leq B(k)$. Therefore, $v_i' \in \mathbf{V}$. □

The above proposition shows that the optimal solution can also be obtained with the new constraints since they define the same solution set. Please note that, as mentioned before, the fairness constraints in the original problem are replaced by parameters $w_i's$. The choice of the proper values of $w_i's$ that maintain fairness is discussed in detail later in this paper.

Among the new constraints, the right-hand sides of inequalities (19) and (20) are not fixed values, but are functions of the selected target system load ψ. Hence, whether or not the final solution is feasible also depends on the choice of ψ. For any value of $\psi \in [0, 1)$, there could be many feasible solutions among which one will be the optimal. Moreover, there must exist an optimal system load ψ^* that can achieve the overall best solution. It is natural to regard the objective Z as the function of system load ψ, $Z = F(\psi)$, and thus Z is the local optimal result at some specific ψ. The maximum Z is achieved when $\psi = \psi^*$. The ultimate objective of the proposed method is to find this optimal ψ^* and the optimal power index assignment vector under it.

In Sections 3.2 and 3.3, we propose a two-step approach to solve the optimization problem (17)–(20). More specifically, in the first step (Section 3.2), we assume a fixed ψ and then given that fixed parameter ψ we propose a simple method (greedy algorithm) trying to find the optimal set of users to receive service. However, this optimality is not a global optimality. In general, as mentioned before, ψ could get any value within the interval $[0, 1)$. The global optimal solution can be obtained when parameter ψ is chosen to be the optimal one ψ^*. The actual objective of the second step of our approach (Section 3.3) is to find this optimal ψ^*, by which the global optimal set of users that will be scheduled to receive service can be identified.

3.2. Greedy Algorithm for a Given System Load. Before obtaining the best system load, we first discuss how to find the local best solution. Assuming that the value of $\psi \in [0, 1)$ is known, the right-hand sides of (19) and (20) can be determined. Combining the two constraints together, we can express the optimization problem (18) by replacing g_i with $\pi_i(h_i, \psi)x_i$, $0 \leq x_i \leq 1$ as follows:

$$\begin{gathered} \max Z = \sum_{i=1}^{B(k)} w_i f_r(\pi_i(h_i, \psi)x_i, \gamma_i), \\ \text{s.t. } \sum_{i=1}^{B(k)} \pi_i(h_i, \psi)x_i \leq \psi, \quad 0 \leq x_i \leq 1. \end{gathered} \tag{26}$$

Note that (26) is a nonlinear continuous knapsack problem with the x_i taking continuous values between 0 and 1. In general, solving this type of problem is proven to be difficult or even impossible in some cases [27]. However, Proposition 1 limits the transmit power of a user i, to either $p_i^{\max}$ or 0 for the optimal solution. This condition provides a possible method to solve the above nonlinear knapsack problem. Without loss of generality, we suppose that the optimal solution is when the first K users transmit at their maximum power, $p_i = p_i^{\max}$, $1 \leq i \leq K$. The optimal system load is $\psi^* = \sum_{i=1}^{K} g_i$. The following theorem states that the power index of an individual user is equal to its power index capacity under ψ^*, that is, $g_i = \pi(h_i, \psi^*)$.

Theorem 1. *Let the optimal solution allow K users to transmit at their maximum power and the system achieves the system load ψ^*. The power index that an individual user received in this case is equal to its power index capacity, that is, $g_i = \pi(h_i, \psi^*)$.*

Proof. For those users whose transmit powers are zero, the corresponding power index capacities are also zero. Therefore, their power indices are zero as well. Without loss of generality, we assume that the K users under consideration are identified as follows: $1 \leq i \leq K$. Based on Proposition 1, we have

$$\frac{h_i p_i^{\max} G_i}{\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0} = \gamma_i, \quad \text{for } 1 \le i \le K. \tag{27}$$

Performing some manipulations in these K equations, we have

$$\frac{h_i p_i^{\max}}{g_i}\left(1 - \sum_{i=1}^{K} g_i\right) = W\eta_0, \quad \text{for } 1 \le i \le K. \tag{28}$$

Letting $\psi^* = \sum_{i=1}^{K} g_i$, we obtain g_i as

$$g_i = \frac{h_i p_i^{\max}(1-\psi^*)}{W\eta_0}. \tag{29}$$

From the definition of power index capacity, we find that $g_i = \pi(h_i, \psi^*)$. □

With reference to the optimal solution of problem (26), we can prove the following theorem.

Theorem 2. *The optimal solution of the constrained optimization problem* (26) *can be obtained by solving the following linear 0-1 knapsack problem:*

$$\begin{aligned} \max Z &= \sum_{i=1}^{B(k)} w_i \frac{W}{\gamma_i} \frac{\pi_i(h_i,\psi)}{1-\pi_i(h_i,\psi)} x_i, \\ \text{s.t. } &\sum_{i=1}^{B(k)} \pi_i(h_i,\psi) x_i \le \psi, \quad x_i = \{0,1\}. \end{aligned} \tag{30}$$

Proof. Since $f_r(x,\gamma_i) = (W/\gamma_i)(x/(1-x))$ for user i, we present the objective function of (26) as follows:

$$\max Z' = \sum_{i=1}^{B(k)} w_i \frac{W}{\gamma_i} \frac{\pi_i(h_i,\psi)}{1-\pi_i(h_i,\psi)x_i} x_i. \tag{31}$$

Based on Proposition 1, we know that the optimal solution is achieved when the transmit power of a user i is either $p_i^{\max}$ or 0. According to Theorem 1, in terms of power index that means that users are assigned either their power index capacity or 0 for the chosen system load ψ. In the above relation (31), the solution for x_i is either 1 or 0. Therefore, we can modify (31) as follows without changing the final optimal solution:

$$\max Z = \sum_{i=1}^{B(k)} w_i \frac{W}{\gamma_i} \frac{\pi_i(h_i,\psi)}{1-\pi_i(h_i,\psi)} x_i, \tag{32}$$

where $x_i = \{0,1\}$. □

Instead of solving for the optimal solution of the above integer knapsack problem (30), which is in principle NP-hard, we utilize a greedy algorithm (GA) in order to obtain an approximate solution. Let Z_a denote the result achieved by the approximate solution, while Z and Z_c denote the corresponding results of the optimal solutions for the integer and continuous knapsack problems, respectively. It has been proven that $Z_a \le Z \le Z_c$ [28]. Furthermore, let

$$\alpha_i \triangleq \frac{W}{\gamma_i(1-\pi_i(h_i,\psi))}, \tag{33}$$

which is a constant value for an individual user. Let us further suppose that all backlogged users are sorted in descending order according to $w_i(k)\alpha_i$, that is, $w_i(k)\alpha_i \ge w_j(k)\alpha_j$, for $i < j$. If it is not the case, these values can be sorted in $O(n\log n)$ time through an efficient procedure. Thus, the optimal continuous solution of problem (30) is given by

$$\begin{aligned} x_i &= 1, \quad \text{for } i < s, \\ x_j &= 0, \quad \text{for } j > s, \\ x_s &= \frac{\psi - \sum_{i<s}\pi_i(h_i,\psi)}{\pi_s(h_s,\psi)}. \end{aligned} \tag{34}$$

An algorithm that finds the critical point s within $O(n)$ time in a system with n users is provided in [28]. Based on solution (34), the greedy algorithm (GA) obtains the approximate solution U as follows:

$$U = \max\{U_1, U_2\}, \tag{35}$$

where

$$\begin{aligned} U_1 &= \begin{cases} x_i = 1, & \text{for } i < s, \\ x_j = 0, & \text{for } j \ge s, \end{cases} \\ U_2 &= \begin{cases} x_i = 1, & \text{for } i = s, \\ x_j = 0, & \text{for } i \neq s. \end{cases} \end{aligned} \tag{36}$$

It has been shown in [28] that in worst case $Z_a/Z = 1/2$. Let Z represent the result that corresponds to the integer solution of (32) when ψ is assigned a value from $[0,1)$, and Z^* be the result when $\psi = \psi^*$. From the definition of ψ^*, we know that Z^* is the maximum value among all Z, that is, $Z^* = \max_\psi\{Z\}$. Based on Proposition 1 and the analysis in the previous subsection, it is easy to find that $\psi^* = \sum_i \pi_i(h_i,\psi^*)x_i$, $x_i = \{0,1\}$. Therefore, when the optimal system power index ψ^* is chosen, $Z_a = Z = Z_c = Z^*$. Since $Z_a \le Z \le Z^*$ and the equality $Z_a = Z^*$ holds only when $\psi = \psi^*$, and the optimal solution can be obtained.

3.3. Optimal System Load. As we discussed in the last subsection the optimal solution of problem (26) depends on the selected system load ψ. Relation (17) shows that the power index capacity increases as ψ decreases. At the first point when $\pi_i = \psi$, the power index capacity reaches its largest value and then it decreases linearly following the value of ψ. Although a smaller value of ψ may increase the single user power index capacity at some range, the finally achieved objective function could be low due to the small system load ψ. On the other hand, setting large ψ reduces the individual user power index capacity as (17) indicates. The consequence of smaller power index capacity is that more users are required to share ψ, and probably a small objective function should be used due to the concavity of function $f_r(x, \gamma_i)$ that converts the power index to throughput. Therefore, whether or not the objective function reaches its maximum value depends not only on the value of the system load ψ, but also on how it is shared among the candidate users. There must exist an optimal value of system load ψ^* that can achieve the maximum weighted rate.

Let the power index vector $\mathbf{g}$ denote the optimal solution, which can be found through the method described in the previous section for a given specific value of ψ. Apparently, $\mathbf{g}$ is a function of ψ. The objective function (18) is the sum of individual weighted rates that are obtained from $\mathbf{g}$ using function $f_r(x, \gamma_i)$. Therefore, Z can also be regarded as a function of ψ. Let $FZ(\psi)$ be the function that gives the maximum value of the sum of weighted rates at ψ. Then the original optimization problem can be rewritten as follows:

$$\begin{aligned} &\max Z = FZ(\psi), \\ &\text{s.t.} \quad 0 \leq \psi < 1. \end{aligned} \tag{37}$$

The optimal solution ψ^* of the above problem and its corresponding power index assignment by (34) with $\psi = \psi^*$ provides the final optimal solution of (18).

Problem (37) is a simple unconstrained maximization problem that searches for the maximum Z within the interval $[0, 1)$. The disadvantage of (37) is that it does not have an explicit expression. Hence, algorithms that rely on the first- or second-order derivatives will not be applicable in this case. Therefore, the searching process depends on the result of (34). Note that every time when a new value of ψ is chosen, the order of $w_i(k)\alpha_i$ may be different from that of previous ψ.

The time of calculating the best result for a newly chosen ψ, including the time of reordering the users (if needed), is easily obtained as $O(n \log n) + O(n) = O(n \log n)$ if n is assumed to be large enough. Moreover, there are many possible local maximum points within the range $0 \leq \psi < 1$. The final optimal ψ must be a global best value. Although in [29] many searching algorithms on how to locate the minimum/maximum solution within a range are described, to make these algorithms effective there must be only one extreme point in the specified range. However, in general it is not possible to know the range which contains only the global optimal value. Thus, an exhaustive search within $[0, 1)$ would be needed. However, the following proposition provides a lower bound ψ^0 with respect to the searching range instead of 0 in order to restrict the corresponding feasible searching range.

Proposition 4. *The lower bound of the feasible searching range is given by*

$$\psi^0 = \min_{1 \leq i \leq B(k)} \left(\frac{\zeta_i}{1 + \zeta_i} \right), \quad \textit{where } \zeta_i \triangleq p_i^{\max} \frac{h_i}{\eta_0} W. \tag{38}$$

Proof. With the decrease of the target system load ψ, the individual power index, provided by (14), will keep increasing till ψ reaches the point ψ_i for user i, that is $(1 - \psi_i)\varsigma_i = \psi_i$. With respect to user i, if $\psi \leq \psi_i$ its power index $\pi_i(h_i, \psi) = \psi$. ψ_i is given by $\psi_i = \varsigma_i/(1 + \varsigma_i)$, which varies with different users since their ς_i are not likely the same. Let ψ^0 be the minimum among all ψ_i's. Once $\psi < \psi^0$ all backlogged users will have the same power index capacities $\pi_i(h_i, \psi) = \psi$. Define a small increment $\Delta\psi$ and let $\psi' = \psi + \Delta\psi < \psi^0$. Apparently, for all users their power indices will all have small increment $\Delta\psi$ such that $\pi_i(h_i, \psi') = \psi + \Delta\psi$. Maintaining the previous power index assignment and giving $\Delta\psi$ to any backlogged user will help increase the objective function (18). We hence can keep adding $\Delta\psi$ to ψ till it reaches $\psi^0 = \Delta\psi + \psi$, which proves this proposition. □

Since the optimal ψ can reside between ψ^0 and 1, we need to calculate a series of sample values after every interval $\Delta\psi$. Apparently, the smaller the $\Delta\psi$, the more samples we get and thereby the more accurate is the obtained result. On the other hand, it also increases the required computational time and power.

Therefore, in practice we only use reasonably small $\Delta\psi$ in order to reduce the corresponding computational power and complexity, while still obtain a good approximation of the optimal solution. It should be noted though that in theory when $\Delta\psi$ becomes infinitely small the above methodology can be used to find the optimal solution. Specifically, there exists an algorithm with complexity of $O(n^4 \log n)$ that guarantees the finding of the optimal solution, however its high complexity limits its applicability for real-time computations and can be used only for benchmarking purposes. Let us assume that the order in (34) is known and fixed. Under this condition, there are only $B(k)$ possible results satisfying the optimal condition in Proposition 1, that is, try the maximum transmission power in the fixed order with number of users from 1 to $B(k)$. The maximum result is the optimal one. For any two users in the possible system load range from $(0, 1)$, their order of $w_i(k)\alpha_i$ will change at most three times. Therefore, there are totally $1.5B(k)(B(k) - 1)$ order changes for $B(k)$ users. Every order change requires first the sorting operation and then the comparison operation that have complexity of $O(n \log n)$ and $O(n)$, respectively, which makes the overall complexity of this method $O(n^4 \log n)$.

The optimal algorithm is described as follows.

(1) Find the m points of target system load, $x_1 < x_2 < \cdots < x_m$, between $[0, 1)$, where the users change their orders in $w_i(k)\alpha_i$. Such points represent actually any point that for any two users i and j, $w_i(k)\alpha_i = w_j(k)\alpha_j$, which is,

$$w_j(k)(1 - \pi_i(h_i, \psi)) = w_i(k)(1 - \pi_j(h_j, \psi)). \tag{39}$$

Based on the definition of power index capacity in (17), the above equation will have at most three solutions.

(2) Once the order is fixed, sort all $B(k)$ users by $w_i(k)\alpha_i$ in descending order. The value α_i can be calculated using any number between $[x_l, x_{l+1})$ since the order will be the same within this range.

(3) Perform $B(k)$ rounds of calculation of objective function (6). In round i, let the largest i users transmit with their largest transmit powers.

(4) Compare the result of round $(i + 1)$ to that of round i. If the result in round $(i + 1)$ is less than round i, then stop

the calculation. In that case, the result of round i is the best result in this order between x_l and x_{l+1}.

(5) The largest result obtained in step (4) is the global optimal solution.

Once the order is fixed in the range $[x_l, x_{l+1})$ at step (2), the method provided in Section 3.2 that finds the best local solution can be applied here, which will provide the largest n, $1 \leq n \leq B(k)$, users with this fixed order. The only difference is that the target system load is not provided directly by a specific known value ψ, but lies within a specific range. Based on Proposition 1, according to which the users allowed to transmit will use their maximum transmission power, we perform $B(k)$ rounds of calculation in step (3) and compare the results to find the optimal n users.

3.4. Fairness Conditions. As mentioned before, fairness is controlled by the vector $\mathbf{w} = \{w_1, w_2, \ldots, w_{B(k)}\}$. When changing the values of w_i, we are actually pursuing a set of optimal fixed values $\mathbf{w}^* = \{w_1^*, w_2^*, \ldots, w_{B(k)}^*\}$ that balance the rate of users with varying channel conditions and hence keep fairness. Since we do not know in advance the exact distribution of the channel conditions, and the number of users may also change, it is difficult to obtain vector $\mathbf{w}^*$ in advance. Therefore, a real-time algorithm is required that is capable of converging w_i toward w_i^*, while maintaining the asymptotic fairness. Stochastic approximation algorithm has been proven to be effective in estimating such parameters. Note that this algorithm has been implemented in [14, 15] in order to solve similar problems. Generally, the stochastic approximation algorithm is a recursive procedure for finding the root of a real-value function $f(x)$. In many practical cases, the form of function $f(x)$ is unknown. Therefore, the result with the input variable x cannot be obtained directly. Instead, the observations of the results, sometimes with noise, will be taken. It has been proven that the root of $f(x)$ can be estimated with the observation $Y_n = f(x_n)$ by the following procedure:

$$x_{n+1} = x_n - \varepsilon_n Y_n, \qquad (40)$$

where $\varepsilon_n > 0$, $\varepsilon_n \to 0$. We can simply let $\varepsilon_n = 1/n$. In most situations, the value of $f(x_n)$ may not be directly available, but instead the $f(x_n) + e_n$, where e_n is the observation noise. In this case, the above approximation approach still applies, with the observed value replaced by $Y_n = f(x_n) + e_n$. The convergence of x_n to the root requires $E(e_n) = 0$.

Here, we define our function $f(\mathbf{w}) = \{f(w_1), f(w_2), \ldots, f(w_{B(k)})\}$ as follows:

$$f(w_i) = \frac{E[r_i(n)]}{E\left[\sum_j r_j(n)\right]} - \frac{\phi_i}{\sum_j \phi_j}, \qquad (41)$$

whose root w_i^* will make $f(w_i) = 0$ which satisfies the fairness condition (3). The noise observation Y_n in our case is:

$$Y_n = \frac{r_i(n)}{E\left[\sum_j r_j(n)\right]} - \frac{\phi_i}{\sum_j \phi_j}. \qquad (42)$$

It is easy to prove that the mean of noise $E[e_n] = E[f(w_i) - Y_n] = 0$. Therefore, the value of w_i^* is then recursively obtained by

$$w_i(n+1) = w_i(n) - \frac{Y_n}{n}. \qquad (43)$$

However, Y_n need to know the mean of total system throughput $E[\sum_j r_j(n)]$. We use a smoothed value $\overline{R}(n)$ to approximate $E[\sum_j r_j(n)]$ and update $\overline{R}(n)$ as follows:

$$\overline{R}(n) = \overline{R}(n-1)\beta + (1-\beta)\sum_j r_j(n-1), \qquad (44)$$

where β is the smooth factor which determines how the estimated $\overline{R}(n)$ follows the change of actual achieved system throughput. In the remaining of the paper, throughout the performance evaluation of our approach, the value $\beta = 0.999$ is chosen. The numerical results presented in Sections 4.2.2 and 4.2.3, with respect to the convergence of w_i's and the achievable fairness, demonstrate that such a method is very effective in approximating the optimal values of w_i^* and therefore controlling and maintaining fairness.

4. Performance Evaluation

In this section, we evaluate the performance of the proposed method in terms of the achievable fairness and throughput, via modeling and simulation. Furthermore, to better understand the performance of the proposed scheduling algorithm-in the following we refer to as throughput maximization and fair scheduling (MAX-FAIR)—we compare it with the maximum throughput (MAX) scheme [16], which achieves the maximum total uplink throughput by allowing only the best k users in terms of their received power to transmit, and with the HDR algorithm [7, 9], which is a single user scheduling algorithm. The principles and operation of HDR basically refer to a proportional fair scheduling scheme, which can be used in the uplink scheduling to demonstrate the one-at-a-time proportional fair scheduling. Following the HDR principles the transmission of a single user at a given time slot is scheduled, with the data rates and slot lengths varying according to the specific channel condition. In the MAX scheme parameter, k is determined by iteratively comparing the throughput of best i users, $1 \leq i \leq N$, where N is the total number of users. The throughput achieved by MAX scheme is regarded as the upper bound throughput in the uplink CDMA scheduling. On the other hand, since HDR achieves temporal fairness, we consider it here to mainly observe the difference between temporal fairness and throughput fairness and their corresponding advantages in specific cases.

4.1. Model and Assumptions. Throughout our numerical study, we consider a single cell DS-CDMA multirate system with multiple active users. All active users are continuously backlogged during the simulation and generate packets with average size of 320 bytes. The maximum transmission power is assumed the same for all users, that is, $p_i^{\max} = 2$ Watts, while the system chip rate is $W = 1.2288 \times 10^6$chip/s as defined in IS-95 and the required SINR is $\gamma_i = 8$ dB for data service, the same for all users. The transmission time is divided into 1 millisecond equal length slots, which is the algorithm scheduling interval, while the simulation lasts for 1.7×10^5 slots.

To study the impact of the channel condition variations on the system throughput and fairness performance, we model the channels through an 8-state Markov-Rayleigh fading channel model [30]. According to this model, the channel has equal steady-state probabilities of being in any of the eight states. We also assume that the coherent time is much larger than the length of a time-slot, hence the channel state is assumed to be constant within a time slot. At the beginning of each time slot, the channel model decides to transit to a new state, which can only be itself or one of its neighbor states, that is, from state s to $s, s+1$, or $s-1$. Table 1 summarizes the state transition probabilities for all the eight states.

Furthermore, four different cases with respect to the ranges of the average SNRs that are assigned to the various users are considered. Specifically, Table 2 presents the corresponding ranges and lists the assignment of the average SNRs for each user for a seven-user scenario, under all these cases. The four different cases represent four different scenarios with respect to the SNR as follows (from top to bottom): large SNR range with low SNR users, low SNR, middle SNR, and high SNR. In the next subsection, we evaluate the performance of MAX-FAIR, MAX, and HDR methods under all four cases and compare their corresponding achieved throughput and fairness.

In most of the numerical results presented in the next subsection, unless otherwise is explicitly indicated, all users are assumed to have the same weight. Such a scenario allows us to better understand and compare the achievable performances of the various scheduling schemes, when users have different channel conditions. However, the operation and effectiveness of the proposed MAX-FAIR policy is also demonstrated in an environment, where users present different weights.

4.2. Numerical Results and Discussion. The numerical results presented in Sections 4.2.1 and 4.2.2 refer mainly to the impact of some of the parameters associated with the proposed MAX-FAIR algorithm on its operation and achievable performance and allow us to obtain a better understanding of its operational characteristics and properties. Then in Sections 4.2.3 and 4.2.4, comparative results about the achievable throughput and fairness of the MAX-FAIR, MAX and HDR algorithms are presented.

4.2.1. Finite System Power Index Samples. Figure 1 shows the sensitivity of the weighted throughput achieved by the MAX-FAIR algorithm as a function of the number of samples used to obtain these values. The last point in the horizontal axis corresponds to the optimal value. It should be noted that in the vertical axis, the depicted weighted throughputs are normalized over the optimal value. Moreover, the different curves provided in this figure correspond to different combinations of the SNR ranges and the number of active users. As can be seen, the more samples we choose, the closer is the obtained maximum value to the optimal value, which clearly presents the tradeoff between the accuracy and the required computational power, as discussed before in Section 3.3. For instance, we observe that in the cases with small SNR range (e.g., [0,1] dB), even 20 samples are sufficient to get satisfactory results, while for the cases with larger SNR range (e.g., [−3,3] dB), more samples may be required.

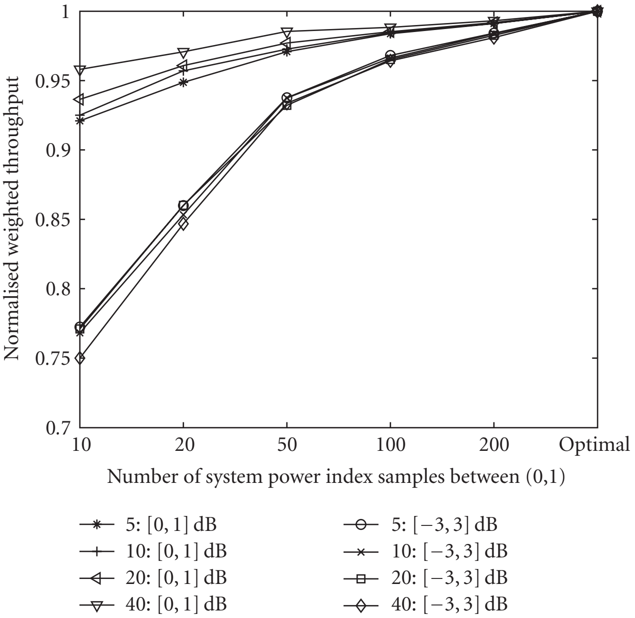


Figure 1: The impact of number of samples on the weighted throughput (MAX-FAIR).

Furthermore, as it can be observed from this figure, for the case of [0,1] dB, the larger the number of active users in the system, the less sensitive is the achievable maximum result to the number of samples (i.e., the slope of the corresponding curve becomes smoother as the number of active users increases). On the other hand, when there are users with high SNR values (e.g., [−3,3] dB), the increasing number of active users makes the achieved throughput drop slightly for small number of samples. This difference in the system behavior is closely related to a different number of simultaneously served users, under different SNR ranges and channel conditions, as depicted by the different observed service patterns in Figure 2.

Specifically, in Figure 2, we present the probabilities of the number of simultaneously served users in each scheduling cycle. For this experiment, we consider 40 backlogged users in the system and perform 200 trials. In each trial, users are randomly assigned the SNRs in the designated SNR range, following the 8-state model [30] described in Section 4.1. We observe that when there are users having high SNR values, for example, in the cases of [−3,3] dB and [2,4] dB, only a small number of users (at most 2 in this experiment), are served concurrently. However, in the case

Table 1: Channel state transition probability.

	s = 1	s = 2	s = 3	s = 4	s = 5	s = 6	s = 7	s = 8
$p_{s,s}$	0.9304	0.8419	0.8170	0.8216	0.8349	0.8590	0.8945	0.9616
$p_{s,s-1}$	0	0.069	0.0879	0.0894	0.0876	0.0777	0.0637	0.0384
$p_{s,s+1}$	0.0696	0.0891	0.0951	0.089	0.0775	0.0633	0.0418	0

Table 2: Simulation cases with different SNR(dB) distribution.

	1	2	3	4	5	6	7
Case: $[-3,3]$	−3	−3	−3	0	0	0	3
Case: $[-4,-2]$	−4	−4	−4	−3	−3	−3	−2
Case: $[0,1]$	0	0	0	1	1	1	1
Case: $[2,4]$	2	2	2	3	3	3	4

that all users have small SNR values, for example, in the case of [−4,−2] dB, the number of simultaneously served users increases significantly (it is distributed between 4 and 17 in our case as can be seen by Figure 2). Such user distribution indicates that in the case that a single user cannot consume all the system resources (e.g., the case where users have low SNR values), more users will be scheduled simultaneously in order to achieve a more efficient resource utilization and as a result increase the total system throughput. This also demonstrates the advantage of our proposed scheduling algorithm over the one-by-one scheduling algorithms that have been proposed in literature. As a result, with respect to Figure 1, for the case of [0,1] dB, multiple users are scheduled to reach the maximal throughput. Increasing the number of active users enables the system to schedule more available candidates to achieve higher throughput, and therefore the achievable result is less sensitive to the number of samples. However, for the case [−3,3] dB at most only 1 or 2 users are scheduled for simultaneous transmission. In the following experiments and numerical results, we adopt the accuracy of 100 samples, which is sufficient to reach 95% of the optimal-weighted throughput.

4.2.2. Parameter Convergence by Stochastic Approximation. As described in Sections 2.1 and 3.4, parameters w_i's are used to represent the fairness constraints in our optimization problem formulation. Figure 3 shows the dynamic change of parameters w_i's as the system and time evolve , for two different cases that correspond to two different SNR ranges. A seven-user scenario is considered, while for demonstration purposes for each case the corresponding values of only two representative users are presented—one user with strong channel and one user with weak channel. As mentioned before, all the users are assigned the same weight in order to more clearly demonstrate the influence of the channel conditions on w_i's. It can be seen by this figure that the converged values of w_i's have the effect of compensating users with the weak channels and reducing the priority of users with strong channels in the scheduling policy. In fact, the converged values of w_i's will make both users (weak and strong) to gain proper system resources and therefore achieve fair throughput. Please note that it is the relative values of w_i's that control the priority of accessing the system resources, and not their absolute values. Furthermore, it should be noted that the lower the average SNR of a weak user, the larger the gap between the weak user and a strong user, which has negative impact on the achievable system throughput, as we will see in the following subsection.

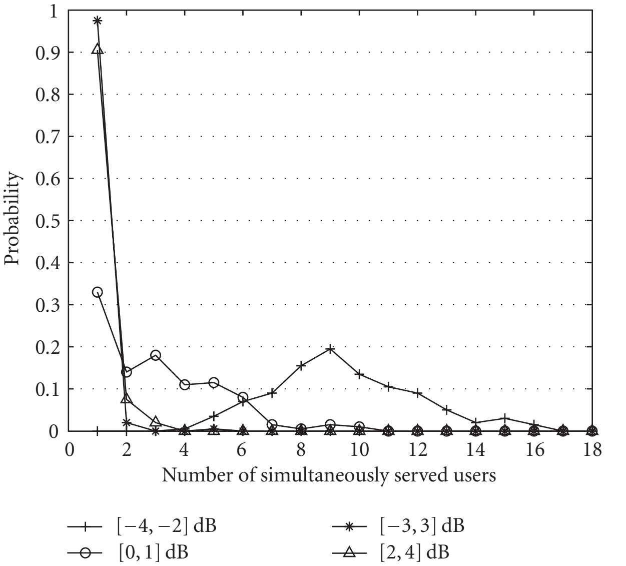


Figure 2: The service pattern under different channel conditions (i.e., SNRs) (MAX-FAIR).

4.2.3. Throughput and Fairness Performance. Figure 4 shows the average throughputs of all the users under the MAX-FAIR, MAX, and HDR methods, for a seven-user scenario where the average SNR range is [−3,3] dB and the corresponding average SNR assignments to the seven users are as shown in Table 2. In order to better demonstrate the tradeoff between the computational complexity and the achievable throughput of MAX-FAIR approach, we obtained the corresponding results under two different cases with respect to the number of power index samples (i.e., 20 and 100 samples). As observed in this figure the MAX-FAIR with 100 power index samples achieves slightly higher throughput, however it requires five times the computational power of the MAX-FAIR with 20 power index samples.

When compared to other two scheduling schemes, MAX-FAIR presents the best throughput-fairness performance (balances the achievable throughput of all users) despite the variable channel conditions of the different users, which indicates that the fairness is well maintained under the proposed scheduling algorithm. As mentioned before in the paper, the main objective of HDR is to achieve

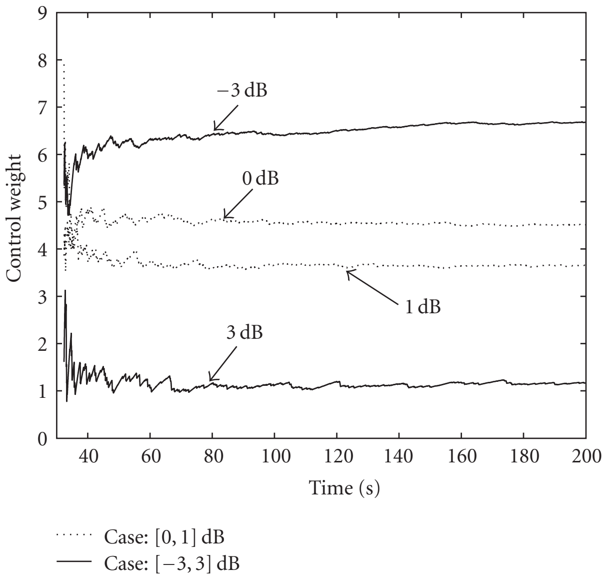


Figure 3: The convergence of w_i's for different users and different SNR ranges (MAX-FAIR).

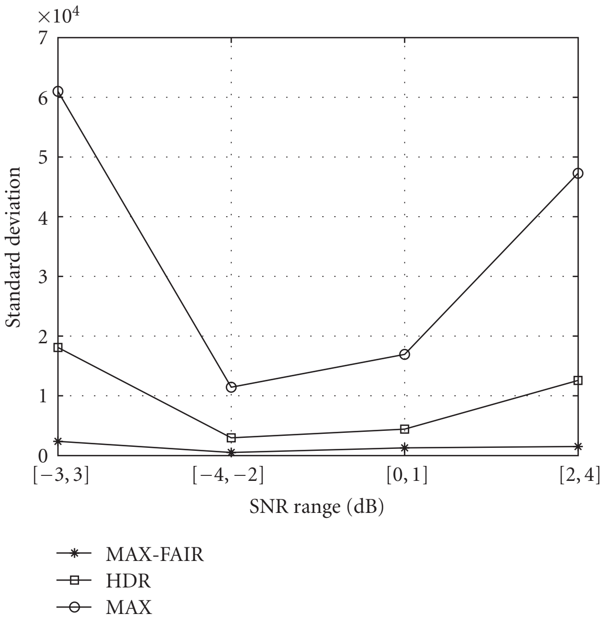


Figure 5: Standard deviation of achievable average throughputs.

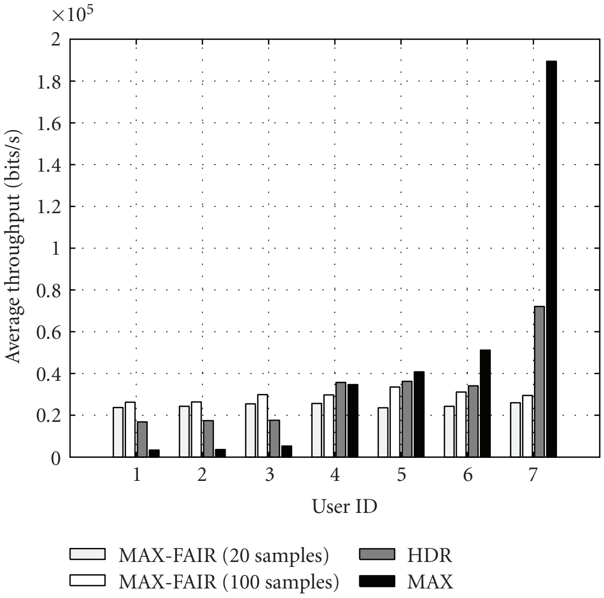


Figure 4: Average throughput for the [−3,3] dB case.

temporal fairness. Therefore, under HDR scheduling each user throughput is closely related to its channel conditions. That is why in Figure 4 we observe that users 1, 2, and 3 have smaller throughput than users 4, 5, and 6, while user 7 has the largest throughput under the HDR scheme. Under the MAX algorithm, user 7 consumes most of the system resources and achieves much higher throughput than the rest of the users due to the fact that the objective of MAX algorithm is to achieve the highest possible total system throughput, without however considering the fairness issue. In Figure 5, we further measure and evaluate the fairness performance by the standard deviation of the average throughput under all the four different SNR cases. Among the three algorithms, MAX-FAIR algorithm has the smallest deviation for all the different cases under consideration, while the corresponding values change only slightly from case to case. We also find that in general the standard deviation increases as the SNRs become higher. This happens because small fluctuation of w_i results in larger throughput change, if all the users have higher SNR levels.

Figure 6 compares the corresponding average system throughputs of the three algorithms under evaluation, for the different SNR ranges (cases). As we expected, MAX-FAIR outperforms HDR in most cases due to the simultaneous scheduling of multiple users, as has been demonstrated in Figure 2, and consequently results in higher resource utilization. However, in the case of SNR range of [−3,3] dB, MAX-FAIR achieves slightly lower throughput than the HDR. The reason of that resides in the different fairness criteria considered and satisfied in these two algorithms, namely, the throughput fairness and temporal fairness. If we examine again Figure 3, we notice that users that have low average SNR (−3 dB) (e.g., users 1, 2, and 3) finally converge to a high w_i, which enables them to have equal opportunity to transmit under the MAX-FAIR scheduling policy. Due to their weak channel conditions, their average throughputs will be low and hence the total system throughput will become lower because of the satisfaction of the throughput fairness constraint. However, as explained before since access time is not the only resource to be shared among the users in these systems, considering throughput fairness instead of temporal fairness is more meaningful in these systems and environments, despite the slightly lower total throughput that can be achieved in some cases under this consideration. One possible alternative solution is to relax the fairness constraint if the QoS permits it. Our experiments have demonstrated

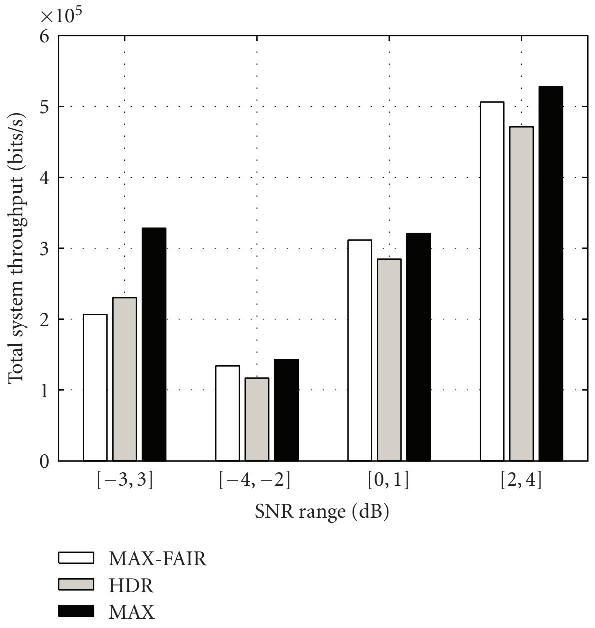


Figure 6: Achieved system throughput under different SNR ranges.

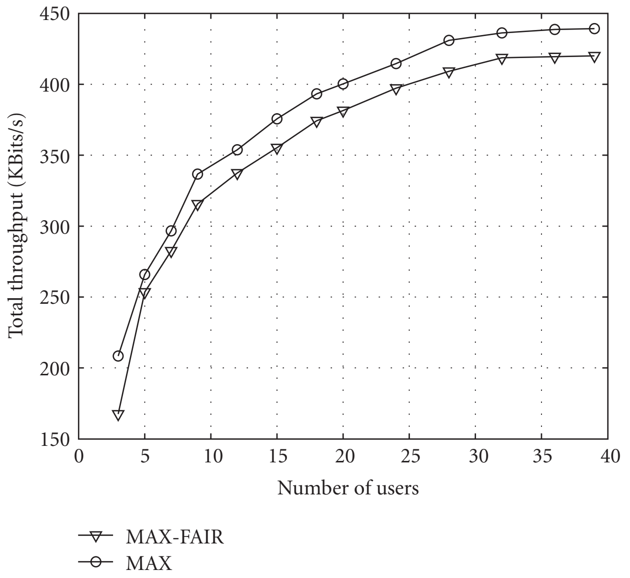


Figure 8: System throughput as a function of the number of backlogged users.

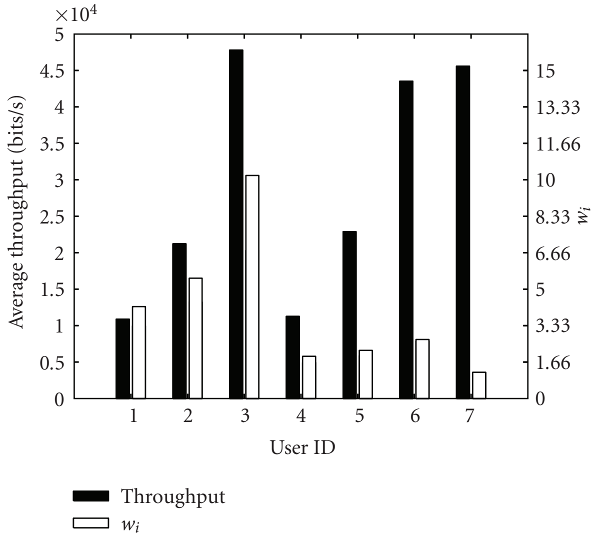


Figure 7: Average throughput under different QoS requirements (weights) by MAX-FAIR.

that after relaxing fairness to 85% of its original requirement, the MAX-FAIR catches up and outperforms the HDR.

In order to obtain a more in-depth understanding of the MAX-FAIR fairness operation, in Figure 7, we present the achieved average throughputs for all the seven users under MAX-FAIR scheme, for a scenario where the SNR range is assumed to be [−3,3] dB, and the users are assigned different weights. The different weights can be considered as the mapping of different QoS requirements. In this scenario, users 1 and 4 have weight 1, users 2 and 5 have weight 2, while users 3, 6, and 7 have weight 4. Figure 7 demonstrates that the MAX-FAIR successfully schedules the transmissions and distributes the resources so that the various users achieve throughput according to their corresponding assigned weights. Specifically users with weights 2 and 4 obtain, respectively, two times and four times the throughput achieved by users with weight 1. In this figure, we also present (on the right-hand side vertical axis) the converged values of parameters w_i's. Here, the different values of w_i's reflect both the channel condition variations and the weight differences. Please note that the relationship between w_i and weight is not linear due to the nonlinearity between the allocated resources and throughput.

4.2.4. Number of Users. Figure 8 shows the achieved total system throughput under MAX and MAX-FAIR algorithms as a function of the number of backlogged users, for the case where the users SNRs are located within [0,1] dB range. Please note that as mentioned before MAX algorithm provides the maximum uplink transmission throughput without considering the fairness property, and therefore is assumed to provide the upper bound throughput in uplink scheduling. From this figure, we can clearly observe the great advantage of the proposed MAX-FAIR approach and its ability to achieve very high throughput, while still maintaining fairness. When the number of backlogged users reaches a certain level, for example, 35 in this experiment, the throughput becomes flat for both MAX-FAIR and MAX, which means that the chances of improving the throughput by opportunistic scheduling with multiple users have been fully utilized.

5. Conclusions

In this paper, the CDMA uplink throughput maximization problem, while maintaining throughput fairness among the various users, was considered. It was shown that such a problem can be expressed as a weighted throughput maximization problem, under certain power and QoS requirements, where the weights are the control parameters that reflect the fairness constraints. A stochastic approximation method was presented in order to effectively identify the required control parameters. The numerical results presented in the paper, with respect to the convergence of the control parameters and the achievable fairness, demonstrated that this method is very effective in approximating the optimal values and therefore controlling and maintaining fairness. Furthermore, the concept of power index capacity was used to represent all the corresponding constraints by the users power index capacities at some certain system power index. Based on this, the optimization problem under consideration was converted into a binary knapsack problem, where the optimal solution can be obtained through a global search within a specific range.

The performance of the proposed policy in terms of the achievable fairness and throughput was obtained via modeling and simulation and was compared with the performances of other scheduling algorithms. The corresponding results revealed the advantages of the proposed policy over other existing scheduling schemes and demonstrated that it achieves very high throughput, while satisfies the QoS requirements and maintains fairness among the users, under different channel conditions and requirements.

Acknowledgment

This work has been partially supported by EC EFIPSANS Project (INFSO-ICT-215549).

References

[1] F. Adachi, M. Sawahashi, and H. Suda, "Wideband DS-CDMA for next-generation mobile communications systems," *IEEE Communications Magazine*, vol. 36, no. 9, pp. 56–69, 1998.

[2] K. D. Wong and V. K. Varma, "Supporting real-time IP multimedia services in UMTS," *IEEE Communications Magazine*, vol. 41, no. 11, pp. 148–155, 2003.

[3] R. Berezdivin, R. Breinig, and R. Topp, "Next generation wireless communications concepts and technologies," *IEEE Communications Magazine*, vol. 40, no. 3, pp. 108–116, 2002.

[4] N. R. Sollenberger, N. Seshadri, and R. Cox, "The evolution of IS-136 TDMA for third-generation wireless services," *IEEE Personal Communications*, vol. 6, no. 3, pp. 8–18, 1999.

[5] J. Ramis, L. Carrasco, G. Femenias, and F. Riera-Palou, "Scheduling algorithms for 4G wireless networks," *IFIP International Federation for Information Processing*, vol. 245, pp. 264–276, 2007.

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

[7] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Communications Magazine*, vol. 38, no. 7, pp. 70–77, 2000.

[8] F. Berggren, S.-L. Kim, R. Jäntti, and J. Zander, "Joint power control and intracell scheduling of DS-CDMA nonreal time data," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 1860–1870, 2001.

[9] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proceedings of the 51st IEEE Vehicular Technology Conference (VTC '00)*, vol. 3, pp. 1854–1858, Tokyo, Japan, May 2000.

[10] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "CDMA data QoS scheduling on the forward link with variable channel conditions," Bell Labs Technical Memorandum 10009626-000404-05TM, Bell Labs, Paris, France, April 2000.

[11] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–154, 2001.

[12] S. Shakkottai and A. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: the exponential rule," Tech. Rep., Bell Labs, Paris, France, 2000.

[13] X. Liu, E. K. P. Chong, and N. B. Shroff, "Transmission scheduling for efficient wireless utilization," in *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '01)*, vol. 2, pp. 776–785, Anchorage, Alaska, USA, April 2001.

[14] X. Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, no. 4, pp. 451–474, 2003.

[15] Y. Liu and E. Knightly, "Opportunistic fair scheduling over multiple wireless channels," in *Proceedings of the 22nd Annual Joint Conference on the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 2, pp. 1106–1115, San Francisco, Calif, USA, March 2003.

[16] S. A. Jafar and A. Goldsmith, "Adaptive multirate CDMA for uplink throughput maximization," *IEEE Transactions on Wireless Communications*, vol. 2, no. 2, pp. 218–228, 2003.

[17] L. Xu, X. Shen, and J. W. Mark, "Dynamic bandwidth allocation with fair scheduling for WCDMA systems," *IEEE Wireless Communications*, vol. 9, no. 2, pp. 26–32, 2002.

[18] C. Li and S. Papavassiliou, "Fair channel-adaptive rate scheduling in wireless networks with multirate multimedia services," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, pp. 1604–1614, 2003.

[19] J. Cho and D. Hong, "Tradeoff analysis of throughput and fairness on CDMA packet downlinks with location-dependent QoS," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 1, pp. 259–271, 2005.

[20] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling policies for wireless systems with short term fairness constraints," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '03)*, vol. 1, pp. 533–537, San Francisco, Calif, USA, December 2003.

[21] C. Li and S. Papavassiliou, "Opportunistic scheduling with short term fairness in wireless communication systems," in *Proceedings of the Conference on Information Sciences and Systems (CISS '04)*, pp. 167–172, Princeton, NJ, USA, March 2004.

[22] S. Ramakrishna and J. M. Holtzman, "A scheme for throughput maximization in a dual-class CDMA system," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 6, pp. 830–844, 1998.

[23] T. Ottosson and A. Svensson, "On schemes for multirate support in DS-CDMA systems," *Wireless Personal Communications*, vol. 6, no. 3, pp. 265–287, 1998.

[24] S. Borst and P. Whiting, "Dynamic rate control algorithms for HDR throughput optimization," in *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '01)*, vol. 2, pp. 976–985, Anchorage, Alaska, USA, April 2001.

[25] A. Sampath, P. S. Kumar, and J. M. Holtzman, "Power control and resource management for a multimedia CDMA wireless system," in *Proceedings of the 6th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '95)*, vol. 1, pp. 21–25, Toronto, Canada, September 1995.

[26] C. Li and S. Papavassiliou, "Joint throughput maximization and fair scheduling in uplink DS-CDMA systems," in *Proceedings of IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communication*, pp. 193–196, Princeton, NJ, USA, April 2004.

[27] D. S. Hochbaum, "A nonlinear Knapsack problem," *Operations Research Letters*, vol. 17, no. 3, pp. 103–110, 1995.

[28] S. Martello and P. Toth, *Knapsack Problems: Algorithms and Computer Implementations*, John Wiley & Sons, New York, NY, USA, 1990.

[29] R. Miller, *Optimization: Foundations and Applications.*, John Wiley & Sons, New York, NY, USA, 2000.

[30] H. Wang and N. Moayeri, "Finite-state Markov channel—a useful model for radio communication channels," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 1, pp. 163–171, 1995.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 682368, 12 pages
doi:10.1155/2009/682368

Research Article

Spatial and Temporal Fairness in Heterogeneous HSDPA-Enabled UMTS Networks

Andreas Mäder[1] and Dirk Staehle[2]

[1] *Department of Distributed Systems, University of Wuerzburg, Sanderring 2, 97070 Würzburg, Germany*
[2] *NEC Laboratories Europe, Kurfuersten-Anlage 36, 69115 Heidelberg, Germany*

Correspondence should be addressed to Andreas Mäder, maeder@informatik.uni-wuerzburg.de

Received 15 July 2008; Revised 27 November 2008; Accepted 29 December 2008

Recommended by Ekram Hossain

The system performance of an integrated UMTS network with both High-Speed Downlink Packet Access users and Release '99 QoS users depends on many factors like user location, number of users, interference, multipath propagation profile, and radio resource sharing schemes. Additionally, the user behavior is an important factor; users of Internet best-effort applications tend to follow a volume-based behavior, meaning they stay in the system until the requested data is completely transmitted. In conjunction with the opportunistic transmission scheme implemented in HSDPA, this has implications to the spatial distribution of active users as well as to the time-average user and cell throughput. We investigate the relation between throughput, volume-based user behavior and traffic dynamics with a simulation framework which allows the efficient modeling of large UMTS networks with both HSDPA and Release '99 users. The framework comprises an HSDPA MAC/physical layer abstraction model and takes network aspects like radio resource sharing and other-cell interference into account.

1. Introduction

Mobile network operators continue to deploy the High-Speed Downlink Packet Access (HSDPA) service in their existing Universal Mobile Telecommunication System (UMTS) networks. From the users perspective, the HSDPA promises high data rates (up to 14.4 Mbps with Release 5) and low latency. From the perspective of an operator, HSDPA is hoped to play a key role for the much longed for breakthrough of high-quality mobile data services. From a technical perspective, HSDPA introduces a new paradigm to UMTS; instead of adapting the transmit power to the radio channel condition in order to ensure constant link quality, HSDPA adapts the link quality to the radio channel conditions. This enables a more efficient use of scarce resources like transmit power, channelization codes, and also hardware components.

The basic principle of the HSDPA is to adapt the link to the instantaneous radio channel condition using adaptive modulation and coding (AMC). HSDPA employs a shared channel, the High-Speed Downlink Shared channel (HS-DSCH), which is used by all HSDPA users. With a shared channel, radio resources are occupied only if a transmission occurs, which enables a more efficient transport of bursty traffic. In each transport time interval (TTI), the scheduler located in the NodeB decides about the users to be scheduled and about their data rate. The scheduling decision can be either on behalf of channel quality indicator (CQI) reports from the user equipments (UE) to enable opportunistic scheduling schemes which use the air interface more efficiently, or simple nonopportunistic schemes like round-robin can be used which shares the resources time fair among the users.

An important aspect of HSDPA systems is the perceived fairness of the connection metrics between the users. This is in contrast to pure UMTS Release '99, where the circuit-switched design of the radio bearers guarantees equal Quality of Service (QoS) properties of all users of the same service class [1]. However, since in HSDPA the theoretically achievable data rate depends on the channel condition, the actual achieved data rates depend on user location, number of users, interference, scheduling discipline, and in integrated networks also on the number of dedicated channel (DCH) connections. In this work, we distinguish between two fairness aspects. *Spatial fairness* refers to the spatial distribution of the perceived data rates within a cell or sector. *Temporal fairness* refers to the long-term time-average user throughput [2].

Our contribution is twofold: first, we propose a flow-level simulation framework which takes on the one hand physical layer aspects, scheduling disciplines, interference, and radio resource management schemes into account, but also allows for simulation of large networks due to its analytical approach. Second, we investigate the impact of three well-known scheduling disciplines, namely round-robin, proportional fair, and Max C/I on the spatial user distribution and on the system and user performance. One of our main findings is that Max C/I scheduling, although providing sum-rate optimal rate allocations in *static* system scenarios, performs worse than proportional fair scheduling if traffic dynamics are considered.

The remaining of this article is organized as follows: in the next section, we motivate our work and give an overview of the current literature. In Section 3, we give a brief overview of the HSDPA. In Section 4, we explain radio resource sharing between DCH and HSDPA connections and formulate a model for the calculation of NodeB transmit powers. In Section 6, a physical layer abstraction model for the HSDPA is proposed which enables the calculation of the average throughputs per flow for different scheduling disciplines. Simulation scenarios and numerical results are presented in Section 7, followed by a conclusion in Section 8.

2. Motivation and Related Work

The focus of this work is the impact of elastic flows on the system performance. We have to distinguish between QoS flows which require a fixed bandwidth, as for voice calls over DCH transport channels, and "best-effort" or elastic flows which adapt their bandwidth requirements to the currently available bandwidth. Such a flow may be an FTP transfer or the combined elements of a web page including inline objects such as embedded videos, that may be transmitted in parallel TCP connections. A flow can be loosely defined as a coherent stream of data packets with the same destination address [3]. An important distinction between the two types of flows is that QoS flows typically follow a time-based traffic model, which means that the user wants to keep the connection for a certain time span. In contrast, elastic flows are volume-based, that is, the user is satisfied as soon as a certain data volume is transmitted. An effect in this context which is that of *spatial inhomogeneity*, which has been mentioned in [4] for systems without AMC, and has been further investigated in [5, 6] for pure single-cell HSDPA systems. Users with bad radio conditions experience lower data rates than users with better radio conditions, leading to a spatial unfairness, which we define as the discrepancy between location-dependent user arrival probabilities and the observed residence probabilities in steady state. We investigate this effect in Section 7.1 for different scheduling disciplines in a multicell scenario, that is, with consideration of other-cell interference, and with location-dependent arrival rates.

A related point is the system performance and fairness of the perceived data rates under different scheduling regimes. In the literature, a large number of fundamental works investigate the tradeoff between fairness and system capacity in a wireless systems with opportunistic scheduling. Examples can be found in [2, 7–10], where in [7] the concept of multiuser diversity (MUD) in downlink direction has been investigated, motivated by the findings in [11] for the uplink direction. For HSDPA systems, research mainly concentrated on variations of the proportional fair scheduler developed for the 1xEV-DO system [12]. Different approaches exist to include QoS constraints on delay or data rate into the scheduling decision [13–17]. The fairness of different schedulers in HSDPA systems is investigated in [18, 19]. Both works conclude that Max C/I provides the highest system throughput. We compare user and system throughput for round-robin, Max C/I, and proportional fair scheduling. The results show that on the one hand, as expected the two channel-aware schemes clearly outperform round-robin scheduling, but on the other hand, proportional fair scheduling leads to a higher time-average throughput than Max C/I scheduling. We discuss this result in detail in Section 7.2.

Statistically valid results for integrated UMTS networks require long simulation runs or analytical approaches. An intuitive example is the DCH blocking probability; a DCH user which is located far from the antenna is subject to strong interference from surrounding NodeBs, he may therefore require a very high transmit power. If this user additionally has a long call time, the influence on the blocking probability is significant. Since such events occur not very often with reasonable loads, long simulation runs are required. The results in this work are therefore generated with a simulation framework based on [20, 21], that uses analytic methods to approximate the effects of the physical layer and the scheduling discipline on flow level. This allows for accurate and time-efficient simulations of large UMTS networks.

3. System Description

We consider a UMTS network where HSDPA and DCH connections share the same radio resources, namely transmit power and channelization codes. The core of the HSDPA is the HS-DSCH, which uses up to 15 codes with spreading factor (SF) 16 in parallel. The HS-DSCH enables two types of multiplexing; time multiplex by scheduling the subframes to different users, and code multiplex by assigning each user a nonoverlapping subset of the available codes. The latter requires the configuration of additional High-Speed Shared Control Channels (HS-SCCHs). Throughout this work we assume that only one HS-SCCH is present, hence consider time multiplex only.

In contrast to dedicated channels, where the transmit power is adapted to the propagation loss with fast power control and thus enabling a more or less constant bit rate, the HS-DSCH adapts the channel to the propagation loss

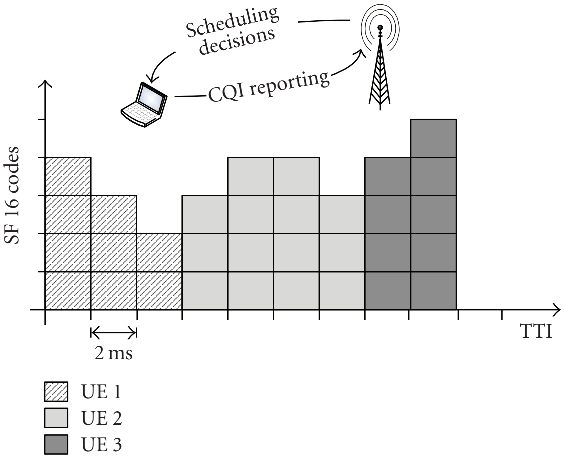


Figure 1: Schematic view of the HSDPA transport channel.

with AMC. The UE sends CQI values to the NodeB. The CQI is a discretization of the received signal-to-interference ratio (SIR) at the UE and ranges from 0 (no transmission possible) to 30 (best quality). The scheduler in the NodeB then chooses a transport format combination (TFC) such that a predefined target BLER, which is often chosen as 10%, is fullfilled if possible. The TFC contains information about the modulation (QPSK or 16QAM), the number of used codes (from 1 to 15), and the coding rate resulting in a certain transport block size (TBS) that defines the information bits transmitted during a TTI. A number of tables in [22] define a unique mapping between CQI and TFC. This means that with an increasing CQI, the demand on code resources is also increasing. This leads to cases where a high CQI is reported to the NodeB, but the scheduler has to select a lower TBS due to lacking code resources. A schematic view of the HSDPA functionality is shown in Figure 1.

4. Sharing Code and Power Resources between HSDPA and DCH

A key issue of the radio resource management in HSDPA enhanced UMTS networks is the sharing of code and power resources between DCHs, signaling channels, common channels, and finally channels required for the HSDPA, namely, the HS-DSCH and the HS-SCCH. The signaling channels and common channels mostly require a fixed channelization code and a fixed power as for the pilot channel (CPICH) or the forward access channel (FACH). The DCHs are subject to fast power control which means that their power consumption depends on the cell or system load that determines the interference at the UE. The general level of power consumption depends on the processing gain and the required target bit-energy-to-noise ratio (E_b/N_0) of the radio access bearer (RAB).

The HSDPA requires code and power resources. Codes are the channelization codes that are generated according to the orthogonal variable spreading factor (OVSF) code tree. The number of codes that is available for a certain spreading factor (SF) is equal to the spreading factor itself. A 384 kbps DCH occupies an SF 8 channelization code. Accordingly, the maximum number of parallel 384 kbps users per sector is theoretically 8. In practice, only 7 parallel 384 kbps users are possible since the signaling and common channels also require some code resources. Let us introduce an SF 512 code as the basic code unit. Then, a DCH i with SF k occupies $c_i = 512/k$ code resources. An HSDPA code with SF 16 requires $c_{\mathrm{HS}} = 32$ code resources. Let C_{DCH} be the total code resources occupied by all DCHs, C_{CCH} be the resources occupied by signaling and common channels, and, $C_{\mathrm{HS}} = n_{\mathrm{HS}} \cdot c_{\mathrm{HS}}$ be the total number of code resources used by the HSDPA where n_{HS} is the number of SF 16 codes allocated to the HS-DSCH. The total number of code resources is equal to $C_{\mathrm{tot}} = 512$. We consider *adaptive* code allocation [23, 24], which is illustrated in a simplified view (pilot and control channels are omitted) in Figure 2 for both transmit power and channelization codes. We further assume that the codes are always optimally arranged in the code tree, and that no code tree fragmentation occurs. The number of codes available for the HSDPA is then

$$n_{\mathrm{HS}} = \left\lfloor \frac{C_{\mathrm{tot}} - C_{\mathrm{CCH}} - C_{\mathrm{DCH}}}{c_{\mathrm{HS}}} \right\rfloor. \tag{1}$$

Accordingly, the transmit power $T_{x,\mathrm{tot}}$ consists of a constant part T_{CCH} for common and signaling channels, a part T_{DCH} for DCHs, and a part T_{HS} for the HS-DSCH. Let T^* be the target transmit power at the NodeB. Then, the HS-DSCH power with adaptive power allocation is

$$T_{\mathrm{HS}} = T^* - T_{\mathrm{CCH}} - \overline{T}_{\mathrm{DCH}}, \tag{2}$$

where T^*_{HS} is the power reserved for the HS-DSCH, and $\overline{T}_{\mathrm{DCH}}$ is the total DCH power averaged over some period of time.

5. Calculation of Downlink Transmit Powers

We define a UMTS network as a set $\mathcal{L}$ of NodeBs with associated UEs, $\mathcal{M}_x$. A DCH connection k corresponds to a radio bearer at NodeB $x \in \mathcal{L}$ with data rate R_k and code resource requirements c_k. Since the power consumed by the DCH connection is subject to power control, the received E_b/N_0 ε_k fluctuates around a target-E_b/N_0 value $\overline{\varepsilon}^*_k$, which is adjusted by the outer-loop power control such that the negotiated QoS parameters like frame error rate are fulfilled. A common approximation for the average E_b/N_0 value is

$$\overline{\varepsilon}_k = \frac{W}{R_k} \cdot \frac{T_{k,x} \cdot d_{k,x}}{W \cdot N_0 + I_{k,\mathrm{oc}} + \alpha_i \cdot T_{x,\mathrm{tot}} \cdot d_{k,x}}, \tag{3}$$

where the orthogonality α_k describes the impact of the multipath profile for DCH k, $d_{k,x}$ is the average path gain between NodeB x and UE k, W is the system chip rate, and N_0 is the thermal noise density. We assume perfect power control, that is, the mean E_b/N_0 value meets exactly the target-E_b/N_0 such that $\overline{\varepsilon}_k = \overline{\varepsilon}^*_k$. The mean transmit power requirement of a DCH connection follows then as

$$T_{k,x} = \frac{\overline{\varepsilon}^*_k \cdot R_k}{W} \cdot \left(\frac{W \cdot N_0 + I_{k,\mathrm{oc}}}{d_{k,x}} + \alpha_k \cdot T_{x,\mathrm{tot}} \right). \tag{4}$$

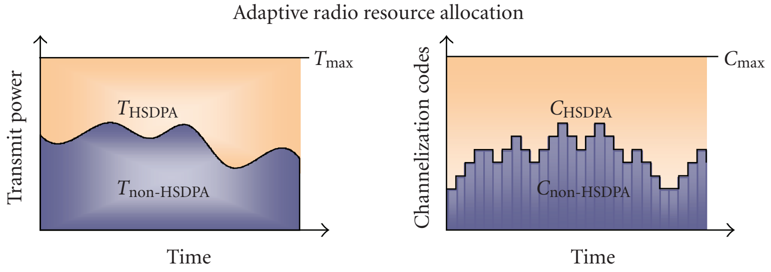


Figure 2: Adaptive radio resource management scheme.

The average other-cell interference comprises the received powers of surrounding NodeBs such that $I_{k,\text{oc}} = \sum_{y\in\mathcal{L}\setminus x} T_{y,\text{tot}} \cdot d_{k,y}$. The total NodeB transmit powers can be calculated with an equation system over all NodeBs. For that reason, we follow [25] and define the load of NodeB x with respect to NodeB y as

$$\eta_{x,y} = \sum_{k\in\mathcal{M}_x} \omega_{k,y}, \\ \text{with } \omega_{k,y} = \frac{\bar{\varepsilon}_k^* \cdot R_k}{W} \cdot \begin{cases} \alpha, & \text{if } \mathcal{L}(k) = y, \\ \dfrac{d_{k,y}}{d_{\mathcal{L}(k)}, k}, & \text{if } \mathcal{L}(k) \neq y. \end{cases} \tag{5}$$

After some algebraic modifications, this allows us to formulate the total DCH transmit power in a compact form as

$$T_{x,\text{DCH}} = \sum_{y\in\mathcal{L}} \eta_{x,y} \cdot T_{y,\text{tot}}. \tag{6}$$

In this equation, we neglect the thermal noise since in a reasonable designed network its impact on the transmit power requirements is minimal. Note also that the equation includes the case $y = x$ for the own-cell interference. For the total transmit power we introduce the boolean variable $\delta_{y,\text{HS}}$ indicating whether at least one HSDPA flow is active in cell x. The total transmit power at NodeB x is then

$$T_{x,\text{tot}} = \delta_{x,\text{HS}} \cdot T_x^* + (1 - \delta_{x,\text{HS}}) \\ \cdot \left(T_{x,\text{CCH}} + \sum_{y\in\mathcal{L}} \eta_{x,y} \cdot T_{y,\text{tot}} \right). \tag{7}$$

This equation states that if the HS-DSCH is active, the total transmit power is equal to the target power. Otherwise, it consist only of the DCH transmit power and the transmit power for common channels. Introducing the vectors

$$V[x] = \delta_{x,\text{HS}} \cdot T_x^* + (1 - \delta_{x,\text{HS}}) \cdot T_{x,\text{CCH}}, \tag{8}$$

and matrix

$$M[x,y] = (1 - \delta_{x,\text{HS}}) \cdot \eta_{x,y} \tag{9}$$

leads to the matrix equation

$$T = V + M \cdot T \Longleftrightarrow T = (I - M)^{-1} \cdot V, \tag{10}$$

which provides the transmit powers of all NodeBs in the system. The matrix I is the identity matrix, and T is the vector of NodeB transmit powers T_x. The DCH and HSDPA transmit powers are then calculated with (6) and (2).

6. HSDPA Physical Layer Model

Consider an HS-DSCH with power $T_{\text{HS}} = \Delta_{\text{HS}} \cdot T_{\text{tot}}$ and n_{HS} parallel codes allocated to the HS-DSCH. Accordingly, the SIR at UE i for a RAKE receiver with perfect maximum ratio combining is equal to

$$\gamma_i = \Delta_{\text{HS}} \cdot \sum_{p\in\mathcal{P}} \frac{T_{\text{tot}} \cdot d_{i,p,x}}{W \cdot N_0 + I_{\text{oc},i} + \sum_{r\in\mathcal{P}\setminus p} T_{x,\text{tot}} \cdot d_{i,r,x}}, \tag{11}$$

where $d_{i,p,x}$ is the instantaneous propagation gain of signal path $p \in \mathcal{P}$. The UE measures the SIR and maps it to the maximum CQI with a transmission format that achieves a frame error rate of 10%. In [26] the following relation of SIR and CQI q is given:

$$q = \max\left(0, \min\left(30, \left\lfloor \frac{\text{SIR[dB]}}{1.02} + 16.62 \right\rfloor\right)\right). \tag{12}$$

The CQI-value q defines the maximum possible TBS $v(q)$, that can be transmitted in one TTI. It also defines the number of required parallel codes $n_{\text{HS}}(q)$. If the number of available codes n_{HS} is less than $n_{\text{HS}}(q)$, the scheduler selects the maximum possible TBS value according to n_{HS}. This means that an optimal usage of resources is only possible if the transmission format according to the reported CQI utilizes all available codes. If too few code resources are available, power resources are wasted, and if too few power resources are available, the CQI is too small to utilize all available codes. The reported CQI value depends essentially on the multipath profile, the users' location, the available HS-DSCH power, and the other-cell power. The number of codes required for a certain CQI value depends on the CQI category.

Above equations give the CQI and TBS for a concrete instance of the propagation gains in particular of the multipath component power. For a simplified simulation and evaluation of the HSDPA performance, an approximate model for the HSDPA bandwidth similar to the orthogonality factor model for DCH is required. The orthogonality factor [27] is used to determine the signal-to-interference ratio for a DCH i as

$$\gamma_i = \frac{W}{R_i} \cdot \frac{T_x \cdot d_{x,i}}{I_{i,\text{other}} + \alpha \cdot I_{i,\text{own}}}, \tag{13}$$

where W/R_k is the processing gain, $I_{i,\text{other}}$ is the other-cell interference, and $I_{i,\text{own}} = T_{x,\text{tot}} \cdot d_{x,i}$ is the own-cell interference. The orthogonality factor α specifies the part of the power received from the own cell that contributes to the interference due to multipath propagation. It captures the impact of the multipath profile in a single value between 0.05 and 0.4 depending on the multipath profile. For a deeper discussion of the orthogonality factor model please refer to [28–30] and the references therein.

Actually, the values γ_k, I_{own}, and I_{other} are mean values averaged over the short-term fading. More precisely, we should write (13) as

$$\begin{aligned} E[\gamma_i] &= \frac{W}{R_i}\cdot\frac{T_{x,i}\cdot d_{x,i}}{E[I_{i,\text{other}}]+\alpha\cdot E[I_{i,\text{own}}]} \\ &= \frac{W}{R_i}\cdot\frac{T_{x,i}}{T_{x,\text{tot}}}\cdot\frac{1}{E[I_{i,\text{other}}]/E[I_{i,\text{own}}]+\alpha}. \end{aligned} \qquad (14)$$

The orthogonality factor model is not applicable to the HSDPA since it only yields the mean SIR. However, for the evaluation of the average HSDPA data rate of a UE at a certain location, the distribution of the reported CQI values is required. The essential assumption of the orthogonality factor model is that the mean normalized SIR, that is, the last fraction in (14), is a function of the ratio Σ of average other-cell received power and average own-cell received power (or short other-to-own-cell power ratio)

$$\Sigma_i = \frac{E[I_{i,\text{other}}]}{E[I_{i,\text{own}}]} = \frac{\sum_{y\neq x} T_{y,\text{tot}}\cdot d_{y,i}}{T_{x,\text{tot}}\cdot d_{x,i}}. \qquad (15)$$

In [20], the orthogonality factor model is enhanced to yield not only the mean but also the standard deviation of the SIR in decibel scale as a function of Σ_i. Assuming that the distribution of the SIR follows a normal distribution that is entirely characterized by its mean and standard deviation, the distribution of the reported CQI values, $p_{\text{CQI}}(q)$, is obtained from the cumulative density function (CDF) of the distribution of the SIR. Truncating the CQI distribution according to the available codes for the HS-DSCH yields the distribution of the TBS as

$$p_{\text{TBS}}(\nu) = \begin{cases} p_{\text{CQI}}(\nu(q)), & \text{if } \nu(q) < \nu^*, \\ \sum_{q=\nu^*}^{30} p_{\text{CQI}}(q), & \text{else}, \end{cases} \qquad (16)$$

where ν^* is the maximum allowed TBS according to the available code resources. Accordingly, we denote the CDF of the CQI and TBS values with $P_{\text{CQI}}(q)$ and $P_{\text{TBS}}(\nu)$.

The physical layer abstraction model gives also insights into the impact of system parameters like multipath channel profile, number of available codes and, UE category. Figure 3 shows the gross data rate, that is, the throughput a single UE would achieve, depending on the other-to-own-interference ratio for the ITU Vehicular A, Pedestrian A, and Vehicular B multipath propagation models. A profile with a strong dominating path, like in Pedestrian A, enables indeed very high data rates up to 13 Mbps. In contrast, profiles with a relatively strong second path, like Vehicular A and Vehicular B, lead to significantly lower data rates due to a higher intersymbol interference. In fact, with these two models, it is sufficient to provide five SF 16 codes for the HS-DSCH. Figure 4 shows the gross data rates for different UE categories, which reflect the capability for 16QAM, number of parallel codes and, interscheduling time. Interesting is that UEs without QAM 16 support (categories 11 and 12) have significantly lower data rates than UEs with QAM 16, although the transport block sizes are identically (categories 1–6).

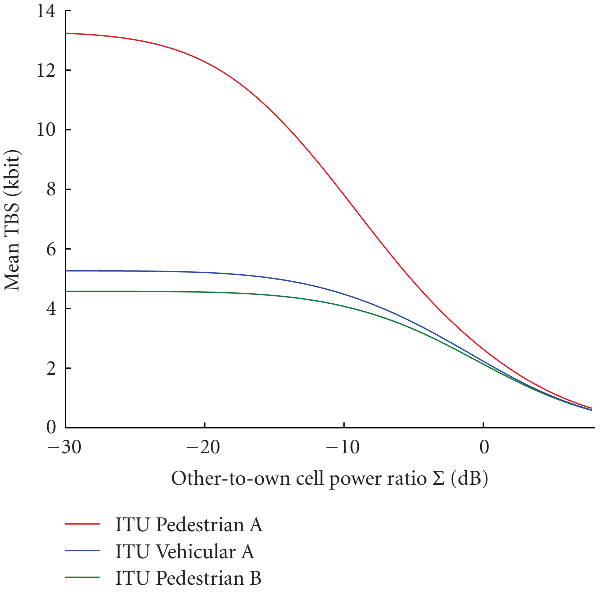


FIGURE 3: Gross data rate for different channel profiles.

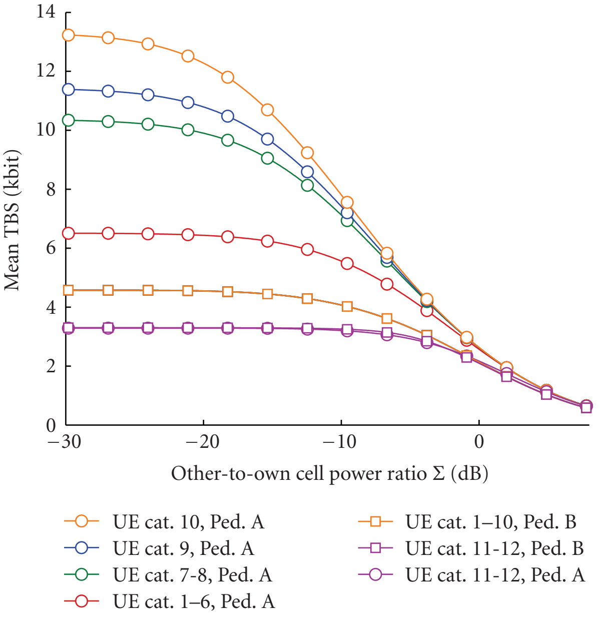


FIGURE 4: Gross data rates for different UE categories.

6.1. Scheduling. The scheduler in the NodeB has a large influence on the user-level and system-level performance of the HSDPA. Several proposals exist for HSDPA scheduling, from which we considered three of the most common schemes. The channel-blind round-robin scheme selects users consecutively for transmission. The MaxTBS-scheduler chooses always the user with the currently best possible TBS, including restrictions due to code resources. Finally, the proportional fair scheduler selects the user which has the proportionally best TBS in relation to its past throughput.

Channel-aware schedulers like MaxTBS and proportional fair benefit from multiuser diversity [7]. With an increasing number of users in a cell, the probability to see at least one user with good radio conditions also increases. If "strong" users are favored by the scheduler, the aggregated cell throughput increases. Exploitation of multiuser diversity is therefore in the end beneficial for the overall system capacity, also because reduced transmission times for volume-based users leads to longer time periods where the HS-DSCH is switched off—which in turn reduces interference.

6.1.1. Round-Robin Scheduling. The *round-robin* scheduler selects the users consecutively for transmission. In a sufficiently long time interval, the probability that a user k is selected is therefore approximately $1/|\mathcal{M}|$. Round-robin is a channel-blind scheduling discipline, which means that the average throughput of each mobile depends only on its channel condition and the number of users in the cell, but not on the channel conditions of other users. Consequently, the cell throughput does not benefit from multiuser diversity. However, round-robin is robust and does not suffer from any convergence issues like proportional fair scheduling in some cases [31], and it is easy to implement due to its simple principle. Round-robin is an allocation-fair scheduling discipline in the sense that, to every user, the same amount of radio resources in terms of codes and power are allocated. This approach is often sufficient to prevent starvation of users at the cell edge.

6.1.2. MaxTBS Scheduling. With *MaxTBS* (or Max C/I) scheduling, the user with the currently best TBS is scheduled. This scheduling discipline maximizes the sum-rate capacity (in our context the cell throughput) given the saturated case, that is, all users have at least one packet to transmit [32, 33]. If two or more users have the maximum possible TBS, a random user out of this set is selected with equal probability. In contrast to round-robin scheduling, the throughput of a user depends not only on its own location, but also on the location of the other users. In [6], this scheduling discipline is modeled as a priority queue, where locations closer to the NodeB have higher priority than locations farther away. However, it is also possible to calculate the average throughput directly from the TBS distributions of the users. In this work we use the formulation we developed in [21]. MaxTBS strongly favors the user with the best channel quality. This implicates that users with weak radio conditions are penalized and perceive on average very low data rates, leading to unfair rate allocations. We show in the next section how this behavior negatively affects the average throughput if traffic dynamics are considered.

6.1.3. Proportional Fair Scheduling. Proportional fair (PF) scheduling is a scheduling discipline which has been developed for the 1xEv-DO-system in the downlink [12]. The basic principle is to allocate each user proportional to its link quality and its past throughput. This is achieved by selecting the user that has the best instantaneous relative throughput over its past throughput, which is often calculated with a sliding window approach. However, different versions of PF scheduling exist. The most fundamental difference is the way how the past throughput is calculated. The first variant updates the past throughput every scheduling period regardless whether the user has been scheduled or not, the second variant updates the past throughput only if the user is indeed chosen for transmission. The difference between both versions is that in the first case the mean throughput of a user is proportional to its channel quality only, while in the second case it is also related to the generated traffic. In [31, 34] it is argued that both variants approximately lead to the same results in case of statistically identical fades and infinite backlogs. The second assumption is reasonable during the interevent time, while the first assumption is contradicted by the fact that the shape of the CQI distribution depends on the level of received other-cell interference. A direct formulation of the flow-average throughput and a comparison between both variants can be found in [21].

7. Flow-Level Performance Results

UMTS networks are dynamic systems because of the mutual dependency among the transmit powers of different cells. This means that a well-designed performance evaluation has to consider networks with a reasonable size in order to capture these effects and their impact on flow-level performance properly. We consider two different types of networks: a 19-NodeB hexagonal layout with a NodeB distance of 1.2 km, and an irregular layout with 22 NodeBs which is generated from a Voronoi tessellation. The network areas are partitioned into area elements with an edge length of 25 m. Figure 5 shows the irregular network with antenna locations (dots) and arrival cluster centers (stars). In the hexagonal layout, user arrive according to a homogeneous Poisson process such that arrival rates are equal for all area elements. In the irregular network, users arrive according to a clustered Poisson process as described in [25] and shown in Figure 6; the total arrival rate λ_f in an area element f results from the superposition of circular clusters with constant arrival rates. In the irregular network therefore not only the layout but also the arrival process is heterogeneous.

Results are generated with a time-dynamic simulation which considers the HSDPA data traffic of a user as a flow with a certain data volume. The network area is discretized into a set of area elements with an edge length of 25 m. The time axis is divided in interevent times. We assume that between two events the users stay roughly within an area element.

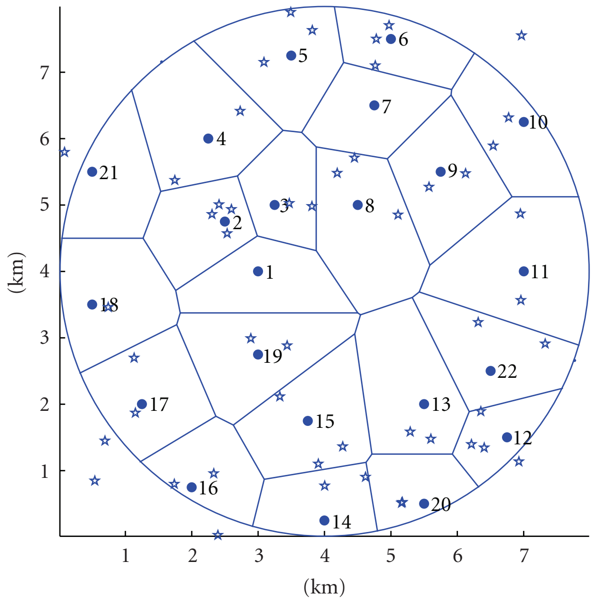


Figure 5: Irregular network layout. Dots indicate NodeB (antenna) locations, stars mark cluster centers.

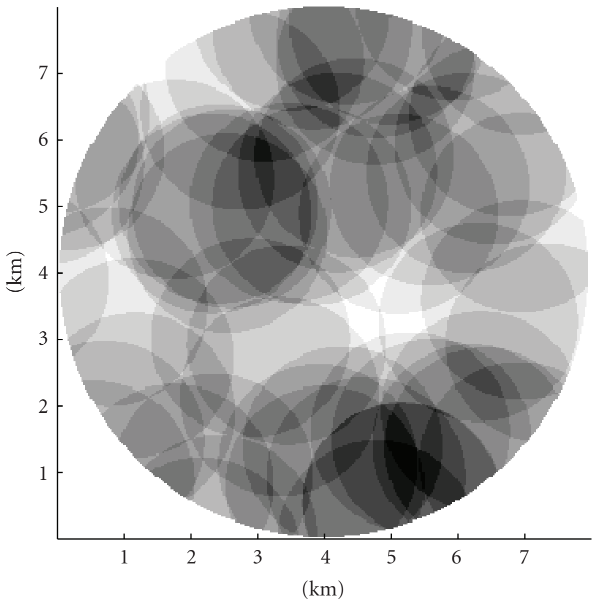


Figure 6: Inhomogeneous arrival densities. Darker colors indicate higher probability of arrival.

We consider two types of events: arrival events, that is, the arrival of a new user into the system, and departure events, which may occur if an HSDPA user has received all its data or if the call time of DCH user is reached. On arrival of a new user, admission control for DCH and HSDPA is performed. The admission control for DCH connections is threshold-based. An incoming connection is blocked if the total transmit power including the new connection exceeds the target transmit power, or if the available code resources are not sufficient. For this purpose, the required transmit power is calculated at the serving NodeB under the worst-case assumption that all NodeBs transmit with the target power in order to prevent possible outage. For the HSDPA, we assume a count-based admission control which restricts the maximum number of concurrent connections to a fixed value. If the incoming connection is admitted into the system, the call time or the data volume, depending on the user type, is calculated according to the respective distribution parameters. We assume exponentially distributed call times with mean $E[T] = 120$ s for DCH users and exponentially distributed flow sizes with mean volume $E[V] = 100$ KB for HSDPA users. The arrival rate of the DCH users is determined from the offered DCH code load defined as

$$\rho_c = \sum_{s\in\mathcal{S}} \frac{\lambda_s}{\mu_s}\cdot\frac{c_s}{C_{\text{tot}}}, \qquad (17)$$

where $\mu_s = 1/E[T_s]$, and the index s denotes the service class of the radio bearer.

On each event, the system variables are recalculated if necessary. If the event is generated by a DCH arrival or departure, HSDPA code resources in the relevant cells are decreased or increased according to the DCH code requirements. Additionally, the total transmit powers are updated for all NodeBs in order to capture the new interference situation. Transmit power recalculation is also done if the HS-DSCH is switched on or off because of HSDPA user arrivals or departures. In all cases, the data volume transmitted by HSDPA users within the past interevent time is subtracted from their remaining data volumes. New HSDPA data rates are calculated, taking the new radio resource and interference situation into account. Finally, the expected departure times of the HSDPA users are updated according to the remaining data volumes and data rates.

7.1. Volume-Based Traffic Model and Spatial Fairness. As mentioned before, an important distinction between QoS and elastic flows is that QoS flows typically follow a time-based traffic model, which means that the user wants to keep the connection a certain time span, for example, for the time of a conversation. In contrast, elastic flows are volume-based, that is, the user leaves the system as soon as a certain data volume is transmitted. In reality, the user behavior is a mixture between both models, depending on factors like user satisfaction, pricing models, type of content. However, the two models can be seen as the extremes of the actual user behavior.

A time-based traffic model implicates that the number of currently active users is independent of the perceived data rates. Moreover, the *spatial distribution* of the number of users is corresponding to the spatial *arrival process*; if users arrive with arrival rate λ, the number of concurrently active users in steady-state follows according to Little as λ/μ, if no blocking occurs.

A volume-based traffic model means that users stay in the system until their service demands are fullfilled. Therefore, the number of active users depends on the assigned data rates. In HSDPA systems, the data rate depends on the channel quality, which means that users with low average channel qualities stay longer in the system than those with good channel qualities. Since the average channel quality is dominated by the other-cell interference, users at the cell edges stay longer in the system than users in the center of the cell. This implies that the spatial arrival process and the spatial steady state distribution are not directly related anymore, a fact that complicates planning of HSPDA networks significantly. One reason is that Monte Carlo methods [35] now have to estimate the spatial user population for every snapshot, which is difficult without knowledge of the the currently ongoing flows. With round-robin scheduling, a direct formulation of the mean transfer time was found in [5, 24], since in that case the data rates of the users only depend on the number of users and their position, but are otherwise independent of each other.

We now clarify the effect of spatial heterogeneity with some example scenarios. Figure 7 shows the arrival probability and the residency probability versus the distance to the antenna for cell number 2 from the irregular scenario. The arrival probability describes the probability that a user arrives in this cell at a certain point, while the residence probability reflects the spatial distribution of the users in the cell in steady state. The spiky shape of the curves is due to the discretization of the cell area into area elements. It is obvious that arrival and residence probabilities are not equal, and that the magnitude of the deviation depends on the scheduling discipline. MaxTBS scheduling shows the highest deviation, since users close to the antenna leave the system much earlier than users farther away. An interesting result is that residence probabilities with proportional fair scheduling fir slightly better to the arrival probabilities if compared to round-robin scheduling. We will see later that this effect comes from the fact that the proportional fair scheduler favors users on the cell edges.

Figure 8 shows the corresponding ratio between arrival and residence probability in the same cell. With time-based users, the ratio would be equal to one at all distances. With volume-based users and MaxTBS-scheduling, the probability to meet a user at the cell edge is four times higher than the arrival probability at the same location.

The deviation of arrival and residence probabilities is the result of spatial unfairness regarding the data rate allocation. This is demonstrated in Figure 9, which shows the average user throughput depending on the distance to the antenna. MaxTBS-scheduling favors strongly user in the cell center, and thus shows the highest degree of unfairness. Proportional fair and round-robin scheduling lead to more balanced results. The difference between round-robin and proportional fair reflects the scheduling gain due to multiuser diversity. Note that the gain of the proportional fair scheduler over the round-robin scheduler is nearly independent of the distance.

Finally, in Figure 10, the same statistic for the center cell of the homogeneous scenario is shown, but in a scenario with a higher DCH load of $\rho_c = 0.6$. Here, the lack of resources leads to low throughputs, such that the aforementioned favoring of user at the cell edge with proportional fair scheduling is clearly visible. This is caused by the higher variance of the TBS distribution of users which experience more other-cell interference than users close to the antenna, see also [36] for a discussion of this effect.

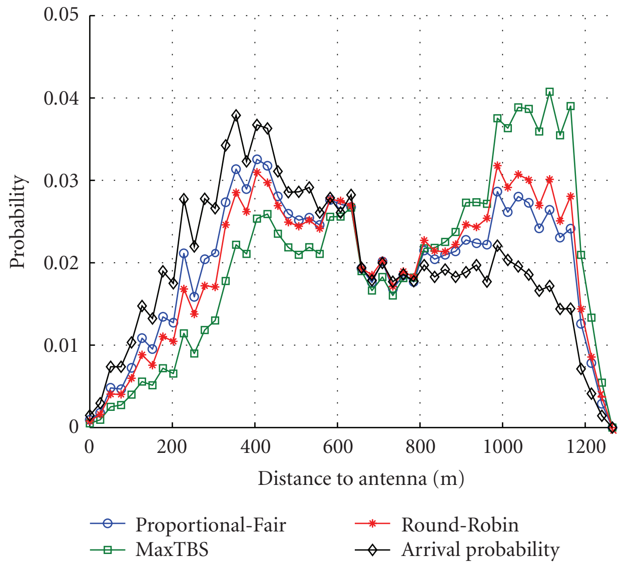


Figure 7: Arrival and residence probabilities for cell 2 in the irregular network with inhomogeneous user arrivals and DCH offered load $\rho_c = 0.4$. The black line with diamond markers indicates the user arrival probability.

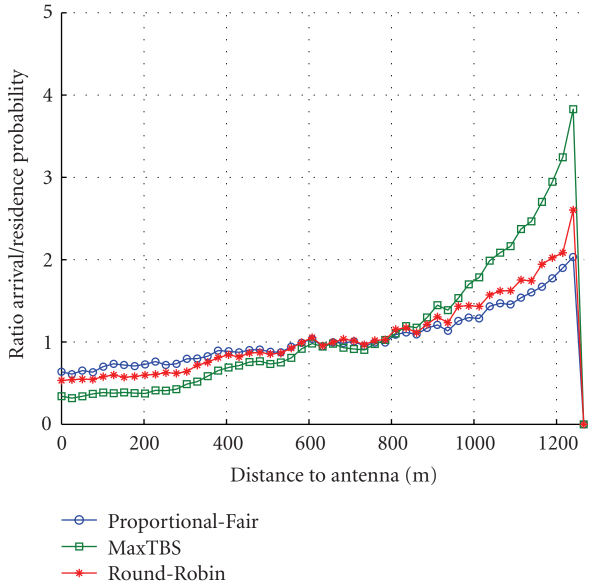


Figure 8: Ratio between arrival and residence probabilities. MaxTBS-scheduling leads to the highest inhomogeneity.

7.2. Impact of Scheduling Disciplines. We now investigate the impact of different scheduling disciplines on the overall performance of the network. We consider the homogeneous scenario with hexagonal cell layout and increase the offered DCH load from 0.1 to 0.8.

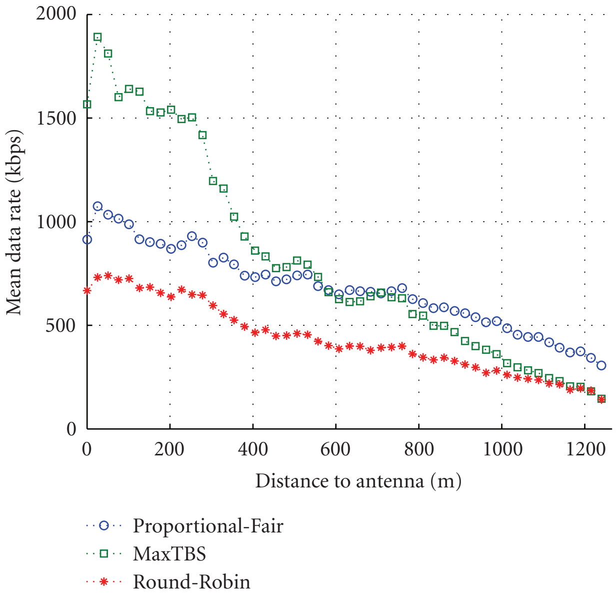


Figure 9: Mean throughput versus distance to antenna with offered DCH load $\rho_c = 0.4$ for cell 2 of the irregular scenario.

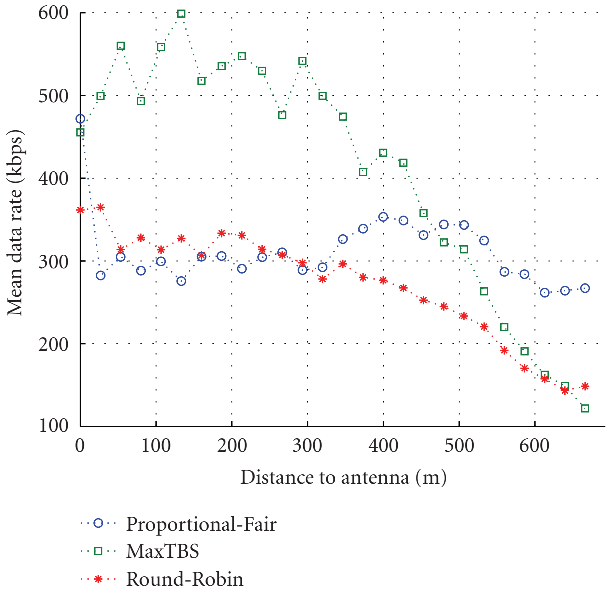


Figure 10: Mean throughput versus distance to antenna for the center cell of the hexagonal scenario with offered DCH load $\rho_c = 0.6$.

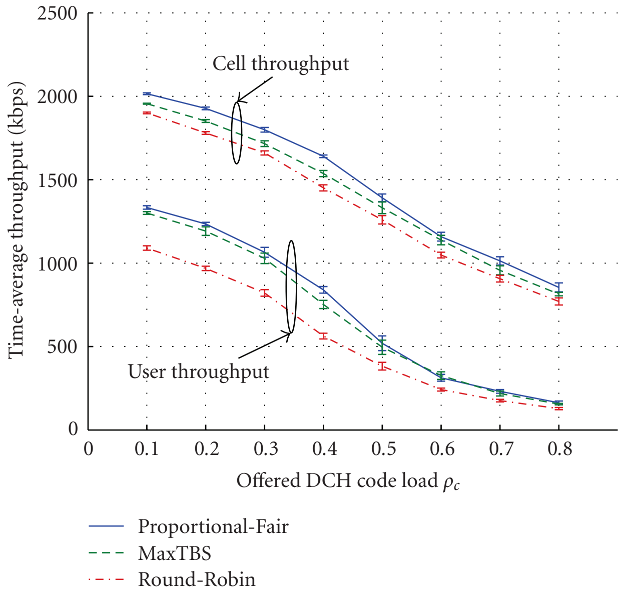


Figure 11: Time-average user and cell throughput versus offered DCH load for different scheduling disciplines.

Figure 11 shows the resulting time-average cell and user throughput versus the offered DCH load. As expected, the channel-aware scheduling disciplines lead to better results than the channel-blind round-robin discipline, regardless of the DCH load. However, with higher DCH load, the difference between the scheduling disciplines becomes smaller, since the lack of code resources prevents an efficient exploitation of multiuser diversity. An interesting result is that proportional-fair scheduling leads to higher throughput curves than MaxTBS-scheduling, which is at a first glance counter intuitive. MaxTBS-scheduling maximizes cumulated data rates (the sum-rate) for a static scenario, that is, for a fixed number of ongoing flows and consequently also during any interevent time [32]. This also means that MaxTBS-scheduling always leads to a higher cell throughput than proportional-fair scheduling if we consider the same snapshot for both schedulers, reflecting the well known tradeoff between system capacity (defined as cell throughput) and fairness of data rate allocation (see, e.g., [10]).

However, this unfairness means that in cases where the differences between the average channel conditions are large, the MaxTBS scheduler has a strong tendency to overproportionally favor the best user, such that the data rates of the remaining UEs are very low. These users stay very long in the system which is then reflected in the time-average cell and user throughput. With proportional-fair scheduling the data rate of users with good channel conditions is lower, however this is compensated with lower sojourn times of users with bad channel conditions. Note that in principle this also holds for round-robin scheduling, but channel-blindness overweights this effect such that the average throughput is indeed lower.

In the literature, some numerical results seem to contradict the results presented here. In [37, 38], the system throughput for round-robin, proportional fair and Max C/I (i.e., MaxTBS) is shown, and it is concluded that Max C/I scheduling provides the highest average cell throughput. However, the results apply to static scenarios with persistent data flows for a fixed number of users. In such a scenario, MaxTBS scheduling is optimal, but it is not comparable with the flow-level throughput in system with traffic dynamics. In [19], users arrive according to a Poisson process and request 100 KB of data, which is incidentally the same average amount of data as in our scenario. However, users are dropped from the system if they stay longer than 12.5 seconds in the system, such that the time-average user sojourn time is reduced. So, in fact this study employs a mixture between time-and volume-based traffic model. Consequently, the results show a small performance gain for Max C/I scheduling. Similarly, in [18] users are dropped from the system if their throughput is lower than 9.6 kbps. It is not clear over which time span the throughput is measured, but the dropping of low-bandwidth users skews the time-average throughput to the benefit of the Max C/I scheduler.

Figure 12 shows the CDF of the user and cell throughputs for an offered DCH load of $\rho_c = 0.4$. The CDF of the MaxTBS scheduler confirms the time-average throughput curves; a large portion of the probability weight is on very low data rates, but in the same time the higher quantiles, for example, for 0.8, are higher than for proportional fair and round-robin scheduling. In terms of fairness, it is remarkable that the shape of the curves for Round-robin and proportional-fair are similar with exception of a small peak for low data rates for the proportional fair scheduler. Also note the stair-like shape of cell-throughput CDF for low data rates, which is caused by preemption from DCH connections.

Figure 13 exemplarily demonstrates the behavior of the three schedulers for scenario with three users which have fixed data volumes and Σ-values of −20 dB, −10 dB, and 0 dB. The figure shows the remaining total data volume versus time. Figure 14 shows the corresponding data rates. With MaxTBS scheduling, the first and second users leave the system faster than with the other disciplines (indicated by the vertical dashed lines), but the remaining data volume of the “worst” user with $\Sigma = 0$ dB is so large that in total, the proportional-fair scheduler needs less time to transport the whole data volume. Note that it depends on channel profile and cell layout how large the advantage of the proportional-fair scheduler is and whether it exists at all.

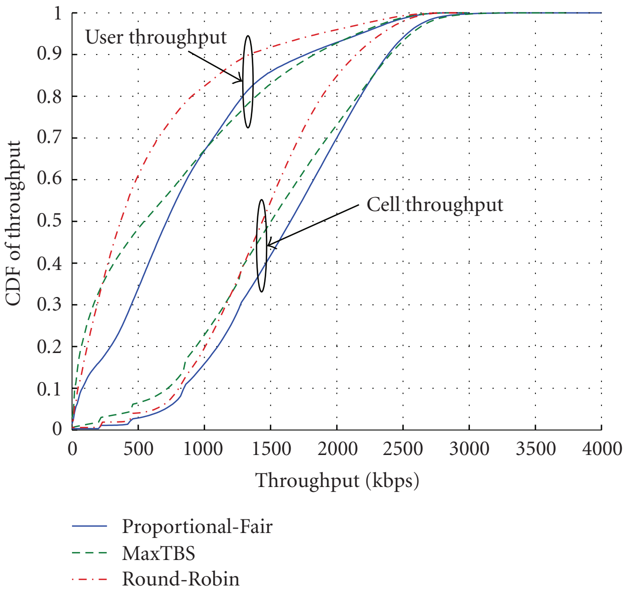


Figure 12: CDF of user and cell throughput for an offered DCH load of $\rho_c = 0.4$.

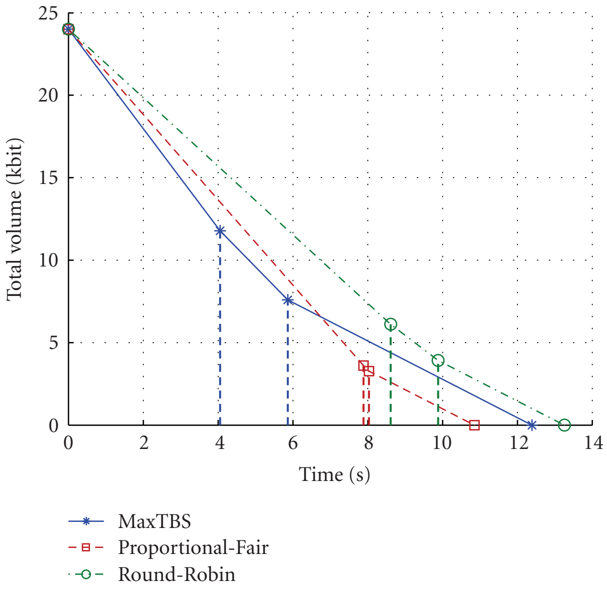


Figure 13: Total remaining data volume versus time for a three-user scenario with fixed data volume. Vertical dashed lines indicate departures.

8. Conclusion and Outlook

We investigated spatial and temporal fairness aspects of integrated HSDPA-enhanced UMTS networks on flow level. Results have been generated with a flow-level simulation which considers the network-wide interference situation and its impact on DCH transmit powers and HSDPA data rates. The latter are calculated with a physical layer abstraction model which considers code resources, multipath-propagation, HS-DSCH transmit power, and different scheduling disciplines.

The numerical results have been generated within two-network scenarios: a homogeneous scenario with hexagonal cells and equal arrival rates over the whole space, and an inhomogeneous scenario with irregular-shaped cells and location-dependent arrival densities. An expected result is that the shared-bandwidth approach of the HSDPA transport channel leads to spatial user residence probabilities which are different to the corresponding arrival probabilities. The degree of unfairness depends on the employed scheduling discipline; “greedy” scheduling disciplines like MaxTBS lead to a high unfairness, while channel-blind round-robin scheduling and proportional fair scheduling show similar results. However, proportional-fair scheduling has a nearly constant relative gain in terms of throughput over round-robin scheduling independent of the distance to the antenna and of the arrival densities.

A further objective of this paper is to understand the flow-level performance of different scheduling disciplines.

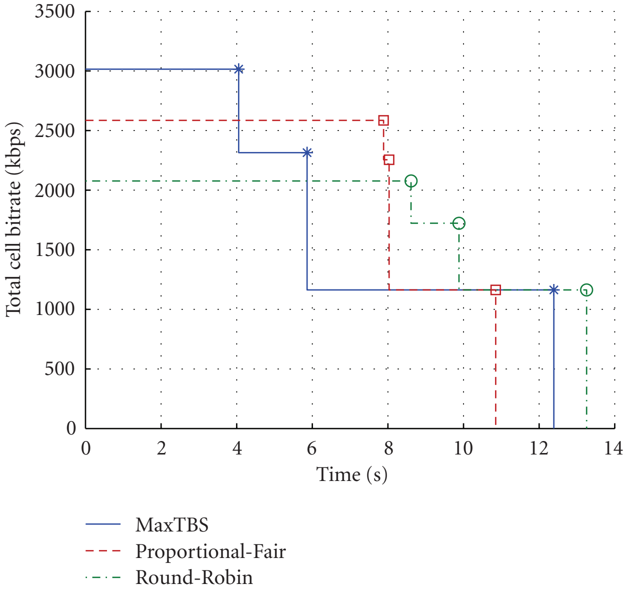


FIGURE 14: Corresponding cell throughput versus time.

The comparison between round-robin, proportional fair, and MaxTBS scheduling showed that, remarkably, proportional fair scheduling has a slight performance gain in terms of average cell and user throughput. The reason is that although MaxTBS-scheduling maximizes the sum rate within a static scenario, traffic dynamics, and the high unfairness of the data rate allocation with MaxTBS favors in the end proportional fair scheduling. This shows that the consideration of traffic dynamics is a crucial point of the performance evaluation of shared bandwidth systems, and it encourages further investigations of the relation between physical layer parameters and flow-level performance.

References

[1] "Quality of service (QoS) concept and architecture," Tech. Rep. TS 23.107 V6.1.0, 3GPP, Valbonne, France, March 2004.

[2] X. Liu, E. K. P. Chong, and N. B. Shroff, "Optimal opportunistic scheduling in wireless networks," in *Proceedings of the 58th IEEE Vehicular Technology Conference (VTC '03)*, vol. 3, pp. 1417–1421, Orlando, Fla, USA, October 2003.

[3] J. W. Roberts, "A survey on statistical bandwidth sharing," *Computer Networks*, vol. 45, no. 3, pp. 319–332, 2004.

[4] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Transactions on Networking*, vol. 7, no. 4, pp. 473–489, 1999.

[5] R. Litjens, J. van den Berg, and M. Fleuren, "Spatial traffic heterogeneity in HSDPA networks and its impact on network planning," in *Proceedings of the 19th International Teletraffic Congress (ITC '05)*, pp. 653–666, Bejing, China, August-September 2005.

[6] H. van den Berg, R. Litjens, and J. Laverman, "HSDPA flow level performance: the impact of key system and traffic aspects," in *Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '04)*, pp. 283–292, Venice, Italy, October 2004.

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

[8] X. Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, no. 4, pp. 451–474, 2003.

[9] Y. Liu and E. Knightly, "Opportunistic fair scheduling over multiple wireless channels," in *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 2, pp. 1106–1115, San Francisco, Calif, USA, March-April 2003.

[10] L. Yang, M. Kang, and M.-S. Alouini, "On the capacity-fairness tradeoff in multiuser diversity systems," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 4, part 1, pp. 1901–1907, 2007.

[11] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proceedings of IEEE International Conference on Communications (ICC '95)*, vol. 1, pp. 331–335, Seattle, Wash, USA, June 1995.

[12] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR: a high efficiency-high data rate personal communication wireless system," in *Proceedings of the 51st IEEE Vehicular Technology Conference (VTC '00)*, vol. 3, pp. 1854–1858, Tokyo, Japan, May 2000.

[13] P. Ameigeiras, J. Wigard, and P. Mogensen, "Performance of the M-LWDF scheduling algorithm for streaming services in HSDPA," in *Proceedings of the 60th IEEE Vehicular Technology Conference (VTC '04)*, vol. 2, pp. 999–1003, Los Angeles, Calif, USA, September 2004.

[14] M. Lundevall, B. Olin, J. Olsson, et al., "Streaming applications over HSDPA in mixed service scenarios," in *Proceedings of the 60th IEEE Vehicular Technology Conference (VTC '04)*, vol. 2, pp. 841–845, Los Angeles, Calif, USA, September 2004.

[15] A. K. F. Khattab and K. M. F. Elsayed, "Channel-quality dependent earliest deadline due fair scheduling schemes for wireless multimedia networks," in *Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '04)*, pp. 31–38, Venice, Italy, October 2004.

[16] M. C. Necker, "A comparison of scheduling mechanisms for service class differentiation in HSDPA networks," *AEU - International Journal of Electronics and Communications*, vol. 60, no. 2, pp. 136–141, 2006.

[17] T. E. Kolding, "QoS-aware proportional fair packet scheduling with required activity detection," in *Proceedings of the 64th IEEE Vehicular Technology Conference (VTC '06)*, pp. 1–5, Montreal, Canada, September 2006.

[18] P. Ameigeiras, J. Wigard, and P. Mogensen, "Performance of packet scheduling methods with different degree of fairness in HSDPA," in *Proceedings of the 60th IEEE Vehicular Technology Conference (VTC '04)*, vol. 2, pp. 860–864, Los Angeles, Calif, USA, September 2004.

[19] T. E. Kolding, F. Frederiksen, and P. E. Mogensen, "Performance aspects of WCDMA systems with high speed downlink packet access (HSDPA)," in *Proceedings of the 56th IEEE Vehicular Technology Conference (VTC '02)*, vol. 1, pp. 477–481, Vancouver, Canada, September 2002.

[20] D. Staehle and A. Mäder, "A model for time-efficient HSDPA simulations," in *Proceedings of the 66th IEEE Vehicular Technology Conference (VTC '07)*, pp. 819–823, Baltimore, Md, USA, September-October 2007.

[21] A. Mäder and D. Staehle, "A flow-level simulation framework for HSDPA-enabled UMTS networks," in *Proceedings of the*

10th ACM Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems (MSWiM '07), pp. 269–278, Chania, Greece, October 2007.

[22] "Medium Access Control (MAC) protocol specification," Tech. Rep. TS 25.321 V6.6.0, 3GPP, Valbonne, France, September 2005.

[23] H. Holma and A. Toskala, Eds., *HSDPA/HSUPA for UMTS: High Speed Radio Access for Mobile Communications*, John Wiley & Sons, New York, NY, USA, 1st edition, 2006.

[24] A. Mäder, D. Staehle, and M. Spahn, "Impact of HSDPA radio resource allocation schemes on the system performance of UMTS networks," in *Proceedings of the 66th IEEE Vehicular Technology Conference (VTC '07)*, pp. 315–319, Baltimore, Md, USA, September-October 2007.

[25] D. Staehle and A. Mäder, "An analytic model for deriving the node-B transmit power in heterogeneous UMTS networks," in *Proceedings of the 59th IEEE Vehicular Technology Conference (VTC '04)*, vol. 4, pp. 2399–2403, Milan, Italy, May 2004.

[26] F. Brouwer, I. de Bruin, J. C. Silva, N. Souto, F. Cercas, and A. Correia, "Usage of link-level performance indicators for HSDPA network-level simulations in E-UMTS," in *Proceedings of the 8th IEEE International Symposium on Spread Spectrum Techniques and Applications (ISSSTA '04)*, pp. 844–848, Sydney, Australia, August-September 2004.

[27] H. J. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Transactions on Wireless Communications*, vol. 3, no. 4, pp. 1250–1259, 2004.

[28] D. N. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," in *Proceedings of IEEE International Symposium on Information Theory (ISIT '97)*, p. 27, Ulm, Germany, June-July 1997.

[29] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels—I. Ergodic capacity," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1083–1102, 2001.

[30] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *IEEE/ACM Transactions on Networking*, vol. 13, no. 3, pp. 636–647, 2005.

[31] U. Türke, M. Koonert, R. Schelb, and C. Görg, "HSDPA performance analysis in UMTS radio network planning simulations," in *Proceedings of the 59th IEEE Vehicular Technology Conference (VTC '04)*, vol. 5, pp. 2555–2559, Milan, Italy, May 2004.

[32] J. M. Holtzman, "CDMA forward link waterfilling power control," in *Proceedings of the 51st IEEE Vehicular Technology Conference (VTC '00)*, vol. 3, pp. 1663–1667, Tokyo, Japan, May 2000.

[33] A. Furuskär, S. Parkvall, M. Persson, and M. Samuelsson, "Performance of WCDMA high speed packet data," in *Proceedings of the 55th IEEE Vehicular Technology Conference (VTC '02)*, vol. 3, pp. 1116–1120, Birmingham, Ala, USA, May 2002.

[34] A. Haider, R. Harris, and H. Sirisena, "Simulation-based performance analysis of HSDPA for UMTS networks," in *Proceedings of the Australian Telecommunication Networks and Applications Conference (ATNAC '06)*, Melbourne, Australia, December 2006.

[35] U. Türke, M. Koonert, R. Schelb, and C. Görg, "HSDPA performance analysis in UMTS radio network planning simulations," in *Proceedings of the 59th IEEE Vehicular Technology Conference (VTC '04)*, vol. 5, pp. 2555–2559, Milan, Italy, May 2004.

[36] J. M. Holtzman, "CDMA forward link waterfilling power control," in *Proceedings of the 51st IEEE Vehicular Technology Conference (VTC '00)*, vol. 3, pp. 1663–1667, Tokyo, Japan, May 2000.

[37] A. Furuskär, S. Parkvall, M. Persson, and M. Samuelsson, "Performance of WCDMA high speed packet data," in *Proceedings of the 55th IEEE Vehicular Technology Conference (VTC '02)*, vol. 3, pp. 1116–1120, Birmingham, Ala, USA, May 2002.

[38] A. Haider, R. Harris, and H. Sirisena, "Simulation-based performance analysis of HSDPA for UMTS networks," in *Proceedings of the Australian Telecommunication Networks and Applications Conference (ATNAC '06)*, Melbourne, Australia, December 2006.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 412837, 10 pages
doi:10.1155/2009/412837

Research Article

Outage Probability versus Fairness Trade-off in Opportunistic Relay Selection with Outdated CSI

Jose Lopez Vicario, Albert Bel, Antoni Morell, and Gonzalo Seco-Granados

Group of Signal Processing for Communications and Navigations (SPCOMNAV), Autonomous University of Barcelona, 08193 Bellaterra (Cerdanyola del Valles), Barcelona, Spain

Correspondence should be addressed to Jose Lopez Vicario, jose.vicario@uab.es

Received 1 July 2008; Revised 18 November 2008; Accepted 20 January 2009

Recommended by Alagan Anpalagan

We analyze the existing trade-offs in terms of system performance versus fairness of a cooperative system based on opportunistic relay selection (ORS) and with outdated channel state information (CSI). In particular, system performance is analytically evaluated in terms of outage probability, and the fairness behavior is assessed based on the power consumption at the different relays. In order to improve the fairness behavior of ORS while keeping the selection diversity gain, we propose a relay selection mechanism where the relay with the highest normalized signal-to-noise ratio (SNR) is selected for relaying the source's information. The proposed strategy is compared with existing relay selection strategies by adopting a novel graphical representation inspired by expected profit versus risk plots used in modern portfolio theory. As shown in the paper, this strategy allows operating the system in more favorable points of the outage versus fairness region.

1. Introduction

Cooperative diversity has been shown to be an efficient way to combat wireless impairments using low-complexity terminals [1–4]. Basically, these schemes allow for the exploitation of spatial diversity gains without the need of multiantenna technology. Different spatial paths are provided by sending/receiving the information to/from a set of cooperating terminals working as relays. By doing so, most of the advantages of multiple-input multiple-output (MIMO) techniques [5] can be extracted while keeping the complexity of the individual terminals reduced. Indeed, the benefits captured by cooperative communications are well extended in the research community, and standardization groups are considering the inclusion of cooperative techniques in practical systems. For instance, the IEEE 802.16j Relay Task Group [6] is involved in the incorporation of relaying mechanisms in the standard adopted by the new wireless system WiMAX [7].

Among the set of cooperative techniques, opportunistic relay selection (ORS) is a useful strategy for practical implementation [8]. This is because ORS is a low-complexity strategy consisting only in activating the best relay (in accordance to a given performance metric). Apart from the inherent simplicity of the proposed technique, this strategy avoids the need of synchronization (needed by most distributed space-time coding schemes) and reduces the power consumption of the terminals.

When ORS is implemented in a real system, however, there may exist a delay between the instants when the selection process is encompassed and the actual transmission of data from the selected relay takes place. In other words, the channel state of the selected relay considered at the selection decision can substantially differ from the actual one and, as a result, system performance is affected.

Besides, in an ORS scheme only the best relay is allowed to cooperate with the source. If channel conditions are not statistically equal for all relays, ORS may be unfair among relays. That is, relays with the worst channel conditions are never selected, and all the cooperation is performed by a reduced set of relays. This can induce a negative effect in the network behavior as one (or more) relay(s) can waste all the battery energy for the sake of cooperation.

Contributions. In this paper, we concentrate our efforts on the analytical study of the behavior of ORS based on decode and forward protocol in a realistic situation where the channel state information (CSI) available at the selection procedure is outdated. More specifically, we derive the exact expression for the outage probability, which is defined as the probability where the instantaneous capacity is below a target value. In order to improve the fairness of ORS, we adopt a fair relay selection strategy where the relay with the largest normalized SNR is selected for relaying the source's information. Furthermore, we explore the existing trade-offs in terms of system performance versus fairness among relays when different relay selection strategies are adopted. To do so, we propose an analysis tool inspired by mean versus standard deviation plots adopted in modern *portfolio* theory [9, 10]. In particular, we adapt such representation to the proposed ORS scenario by illustrating the gain in terms of system performance versus the difference among relays in terms of power consumption. As shown in the paper, this kind of representation is quite useful to quantify what the performance versus fairness trade-off of the proposed relaying strategy is.

Relation to Prior Work. The study of the impact of outdated CSI on ORS has been addressed by few works. For instance, it was shown in [11] that a selection relaying mechanism based on localization knowledge can outperform an opportunistic scheme with instantaneous information. Although it was not explicitly discussed, the reason for that is that available CSI was subject to delays. As a consequence, the selection scheme proposed in [11] may work better when decisions are made based on location information instead of instantaneous but outdated CSI (localization variations are considerably slower than those induced by the wireless channel). In this work, we shed some light into this issue by providing an analytical study of the behavior of ORS when CSI is outdated.

Concerning the fairness analysis of cooperative strategies, some studies deal with this topic in literature. In [12, 13] cooperation protocols based on power rewards were proposed for energy-constrained ad hoc networks in order to attain a fair situation where all the nodes run out of energy simultaneously. With the same objective in mind, a relay set selection protocol was proposed in [14]. In particular, the authors of that work proposed a multistate energy allocation method, where in each state a different set of relays are selected until these relays run out of energy. The fairness nature of the proposed strategy comes from the fact that the same energy is allocated to all the nodes of the active set, being this energy optimized with the aim of minimizing outage probability. In [15–17], cooperative schemes based on ORS with amplify and forward were adopted. The authors in [15] focused the study on the comparison of round robin with centralized and distributed ORS-based selection strategies. Clearly, better performance was achieved with the ORS strategies while preserving fairness in the temporal domain. In that case, nonetheless, fairness was assured due to the i.i.d channel modeling of the proposed scenario. In [16], a power saving technique was proposed, where transmit power at the relays was minimized according to SNR constraints. By doing so, a good balance between the diversity gain and fairness of battery usage was obtained but complexity and signaling requirements of the system were increased with the proposed power allocation method. On the other hand, the authors in [17] proposed a selection scheme based on the selection of the relay with the best weighted SNR aimed at improving the fair behavior of ORS (measured by the percentage of power consumption). In our work, we also consider a selection scheme based on weighted SNR but, as discussed later, different considerations must be adopted in the proposed scenario based on decode and forward protocol, and different conclusions are drawn. Besides, we propose a fairness analysis tool inspired in *portfolio* theory to facilitate the study of the existing trade-offs in terms of system performance versus fairness among relays in a realistic scenario where available CSI is subject to delays.

Organization. The corresponding system model is presented in Section 2. In Section 3, a closed-form expression for the outage probability of the proposed relay selection mechanism is derived, and some numerical results are provided to evaluate the performance of different relay selection schemes. After that, the fairness of the different relaying strategies is illustrated in Section 4 by using outage probability versus standard deviation of the power consumption plots. Finally, in Section 5, the summary and conclusions of this paper are presented.

2. System Model

Consider a wireless network where one mobile unit (source) sends information to the base station (destination). In order to improve system performance, a cooperative mechanism is considered. In particular, an ORS strategy is adopted in a scenario with K mobile units of the network working as relays. In Figure 1, we present an example of the proposed scenario. Notice that we have considered a parallel relay topology [18] where relays are linearly placed halfway between the source and the destination, in a segment of length d, where d is also the distance of the source-destination link. It is worth noting, however, that the main results obtained in this paper depend on the relay selection mechanisms but not on the specific relay arrangement.

2.1. Signal Model. For the sake of notation simplicity, we define an arbitrary link A-B between two nodes A and B. Node A can be the source ($A = S$) or the kth relay ($A = k$), while node B can correspond to the kth relay ($B = k$) or to the destination ($B = D$). With this model in mind, the received signal in the link A-B can be written as follows:

$$r_B = h_{A,B} x_A + n_B, \tag{1}$$

where $x_A \in \mathbb{C}$ is the transmitted symbol from node A with power $P_A = \mathbb{E}[|x_A|^2]$, $n_B \in \mathbb{C}$ is AWGN noise with zero mean and variance σ_n^2 (independent of the value of B), $h_{A,B} \in \mathbb{C}$ is the channel response between nodes A and B modeled as $h_{A,B} \sim CN(0, \sigma_{A,B}^2)$ (Rayleigh fading), being $\sigma_{A,B}^2$ the channel strength depending on the simplified path-loss model [19],

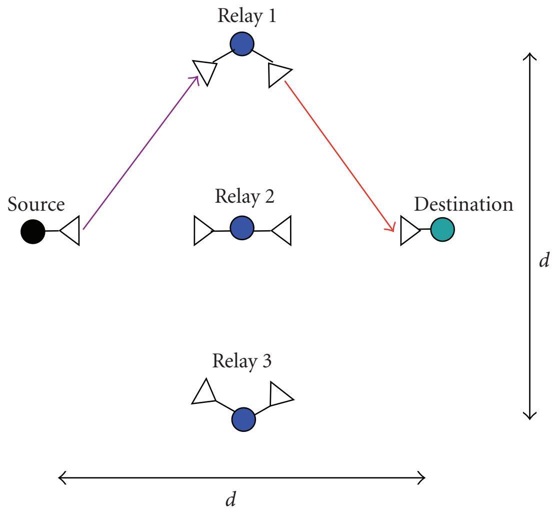


Figure 1: Scheme of the proposed relaying strategy.

$\sigma^2_{A,B} = (\lambda_c/4\pi d_o)^2(d_{A,B}/d_o)^{-\mu}$, with λ_c standing for the carrier wave-length, d_o is a reference distance, $d_{A,B}$ is the distance of the link A-B, and μ is the path-loss coefficient (being $\mu = 3$ in this work). We assume a block-fading channel where the channel response remains constant during one time-slot and that the different channels (for changing A or B) are independently distributed. Concerning power allocation, we consider that total transmit power of the system, P, is evenly distributed among the source and the selected relay, k^*, that is, $P_S = P_{k^*} = 0.5P$. We denote by $\gamma_{A,B} = P_A|h_{A,B}|^2/\sigma_n^2$ the instantaneous signal-to-noise ratio (SNR) experienced in the link A-B in a given time-slot and by $\overline{\gamma}_{A,B} = P_A\sigma^2_{A,B}/\sigma_n^2$ its long-term average. Also, we define $\widehat{\gamma}_{A,B}$ as the SNR employed by the relay selection mechanism, which can differ from the actual SNR SNR $\gamma_{A,B}$ but both of them have the same long-term average $\mathbb{E}[\widehat{\gamma}_{A,B}] = \mathbb{E}[\gamma_{A,B}] = \overline{\gamma}_{A,B}$ (further details can be found in Section 2.3).

Finally, it is worth pointing out that one of the main scopes of this work is to show the impact of outdated CSI on relay selection decisions, and, for the sake of mathematical tractability, we will be considering the capacity of a single carrier system. The study can be easily extended to OFDM by applying the same analysis to each subcarrier simultaneously, and, hence, it is applicable to WiMAX on a subcarrier per subcarrier basis.

2.2. Relaying Mechanism. In this work, we consider a half-duplex two-hop decode and forward (DF) protocol as relaying strategy. When using half-duplex DF, the transmission is divided in two time-slots. In the first time-slot, the source transmits the information to the relays, which attempt to demodulate and decode this information. In the second time-slot, the relays encode again the information and retransmit it to the destination [4]. In an ORS scheme, only the best relay is allowed to cooperate with the source. More specifically, the subset of relays able to decode the information is named as the decoding subset $\mathcal{DS}$, and, from that subset, the relay with the best relay-destination channel quality retransmits the information (see Figure 2).

Unlike other approaches, the scheme proposed in this work selects the relay with the largest normalized SNR instead of the largest absolute SNR because of practical considerations. In other words, the selected relay k^* is such that:

$$k^* = \arg\max_{k\in\mathcal{DS}}\left\{\frac{\widehat{\gamma}_{k,D}}{\mathbb{E}[\widehat{\gamma}_{k,D}]}\right\} = \arg\max_{k\in\mathcal{DS}}\left\{\frac{\widehat{\gamma}_{k,D}}{\overline{\gamma}_{k,D}}\right\}. \quad (2)$$

The reason why we propose this selection strategy is due to the fairness introduced in the selection procedure as all relays will be chosen with the same probability. Thus, the power consumption of the different terminals is uniformly distributed, while diversity gains can still be efficiently extracted. This can help to improve the acceptance by the different users of cooperation mechanism since all of them contribute to common welfare with the same amount of battery. Notice that this strategy was also presented in [17]. In that paper, however, it was shown that the benefits provided by the largest normalized SNR in terms of fairness were not significant. It is then worth recalling that a different scenario based on amplify and forward was presented, and, for that reason, different conclusions were drawn (further details in Section 4.1). If the selection were based on the absolute SNR, some users may be reluctant to participate since they may experience battery consumption faster than the average.

Notice that the relay selection approach makes its decision based on the estimated version of the SNR, $\widehat{\gamma}_{k,D}$. Concerning the accuracy of this estimate, it will depend on the way that CSI is provided. Here, we discuss two methodologies according to the adopted duplexing mode, that is, frequency (FDD) or time (TDD) division duplexing.

(i) FDD: since uplink and downlink channels operate at different frequency bands, feedback mechanisms are required. First of all, relays belonging to the decoding subset send a signalling message to the destination (i.e., BS) indicating that they are able to relay the message. This signalling message can be, for instance, a pilot sequence used by the BS to estimate the instantaneous SNRs of the different relays. Once the different SNRs are estimated, the BS selects the relay with the best quality and broadcasts this decision via a selection command (only $\log_2 K$ bits required).

(ii) TDD: in the case that channel reciprocity between the uplink and downlink holds, each of the relays is able to know its own CSI. TDD: in the case that channel reciprocity between the uplink and downlink holds, each of the relays is able to know its own CSI. With this information, a possible selection strategy is that proposed in [20]. Those relays belonging to the decoding subset start a timer. The timer of each relay adopts as initial value a parameter inversely proportional to its instantaneous SNR. Then, the timer that first expires is that belonging to the best relay. In order to avoid collision, this relay signals its presence to the rest of relays via a flag packet

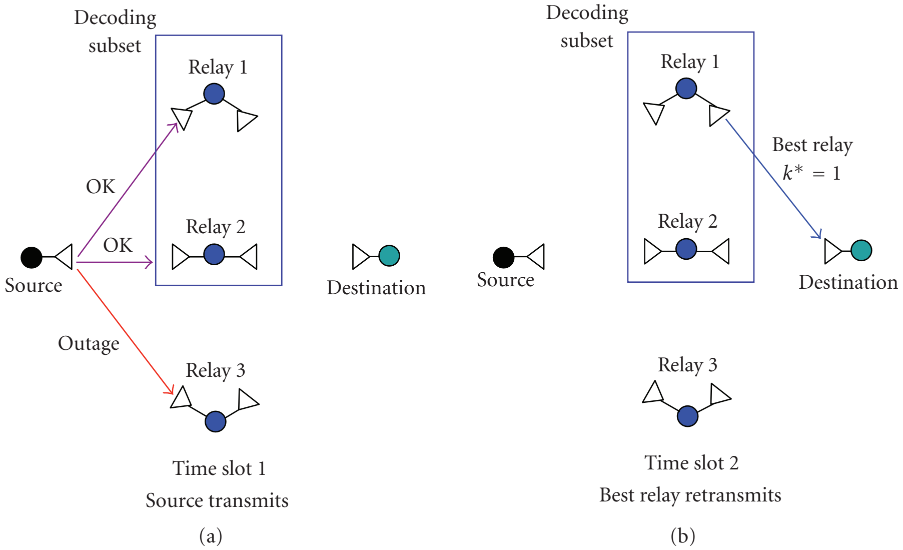


Figure 2: Cooperative communications scheme based on ORS with DF.

before the relaying procedure is started (further details about strategies to avoid collision can be found in [20]). Clearly, channel reciprocity holds in TDD when the time coherence of the channel is higher than the time difference between uplink and downlink time slots. In the opposite case, the methodology adopted for the FDD case should be considered as well. With this information, a possible selection strategy is that proposed in [20]. Those relays belonging to the decoding subset start a timer. The timer of each relay adopts as initial value a parameter inversely proportional to its instantaneous SNR. Then, the timer that first expires is that belonging to the best relay. In order to avoid collision, this relay signals its presence to the rest of relays via a flag packet before the relaying procedure is started (further details about strategies to avoid collision can be found in [20]).

As can be observed in both strategies, there exists a time delay, T_D, between decision and relay transmission instants that may affect system performance.

2.3. Modeling of CSI Delay. We consider that the SNR estimates available at the selection procedure were obtained from a channel state, $\hat{h}_{k,D}$, which differs from the actual channel response at the relay retransmission instant, $h_{k,D}$, due to the effect commented above. Indeed $\hat{h}_{k,D}$ is an outdated version of $h_{k,D}$, that is, these two random variables are samples of the same Gaussian process. Then, $h_{k,D}$ conditioned on $\hat{h}_{k,D}$ follows a Gaussian distribution [21]:

$$h_{k,D} \mid \hat{h}_{k,D} \sim \mathcal{CN}\left(\rho_k \hat{h}_{k,D}, \left(1-\rho_k^2\right)\sigma_{k,D}^2\right), \tag{3}$$

where parameter ρ_k (with $0 \leq \rho_k \leq 1$) is the correlation coefficient between $\hat{h}_{k,D}$ and $h_{k,D}$ (degree of CSI accuracy), having different expressions according to the channel model. Under the assumption of Jakes' model, for instance, the correlation coefficient takes the value $\rho_k = J_o(2\pi f_{d_k} T_{D_k})$, where f_{d_k} stands for the Doppler frequency, T_{D_k} is the delay mentioned in the previous subsection, and $J_o(\cdot)$ denotes the zero-order Bessel function of the first kind.

From the above discussion, it is straightforward to show that the actual SNR, $\gamma_{k,D}$, conditioned on its estimate, $\hat{\gamma}_{k,D} = P_k|\hat{h}_{k,D}|^2/\sigma_n^2$, follows a noncentral chi-square distribution with 2 degrees of freedom, whose probability density function (pdf) takes the following expression [21]:

$$\begin{aligned} &f_{\gamma_{k,D}|\hat{\gamma}_{k,D}}\left(\gamma_{k,D} \mid \hat{\gamma}_{k,D}\right) \\ &\quad = \frac{1}{\overline{\gamma}_{k,D}\left(1-\rho_k^2\right)} e^{-(\gamma_{k,D}+\rho_k^2\hat{\gamma}_{k,D})/\overline{\gamma}_{k,D}(1-\rho_k^2)} I_0\left(\frac{2\sqrt{\rho_k^2\gamma_{k,D}\hat{\gamma}_{k,D}}}{\overline{\gamma}_{k,D}\left(1-\rho_k^2\right)}\right), \end{aligned} \tag{4}$$

where $I_0(\cdot)$ stands for the zero-order-modified Bessel function of the first kind, and one should take into consideration that the long-term average of $\hat{\gamma}_{k,D}$ is equal to $\mathbb{E}[\hat{\gamma}_{k,D}] = \mathbb{E}[|\hat{h}_{k,D}|^2]P_k/\sigma_n^2 = \mathbb{E}[|h_{k,D}|^2]P_k/\sigma_n^2 = \overline{\gamma}_{k,D}$.

3. Outage Probability Analysis

In this section, we analyze the behavior of the proposed relay selection strategy in terms of outage probability. To do so, we first obtain an analytical expression for the outage probability. After that, we show some numerical examples where the proposed fair strategy is compared to other existing relay selection strategies.

3.1. Analytical Expression of the Outage Probability. The outage probability is defined as the probability where the instantaneous capacity of the system is below a predefined value R. Since we consider a two-hop DF scenario, we should start the analysis by studying the decoding subset $\mathcal{DS}$, that is, the subset of relays that are not in outage in the source-to-relay link:

$$\mathcal{DS} = \{k : \log_2(1+\gamma_{S,k}) \geq 2R\} = \{k : \gamma_{S,k} \geq 2^{2R}-1\}. \quad (5)$$

Note that we have considered that outage in the first hop occurs when instantaneous capacity is lower than $2R$ (as it will do in the relay-to-destination link). By doing so, the resulting end-to-end spectral efficiency is R as the proposed two-hop scheme requires two time-slots to transmit the information from the source to the destination.

By defining now $\mathcal{DS}_l$ as an arbitrary decoding subset with l relays, we can easily compute its probability as [8] follows:

$$\begin{aligned}\operatorname{Prob}(\mathcal{DS}_l) &= \prod_{i\in\mathcal{DS}_l}\operatorname{Prob}(\gamma_{S,i}\geq y)\prod_{j\notin\mathcal{DS}_l}\operatorname{Prob}(\gamma_{S,j}<y)\\ &= \prod_{i\in\mathcal{DS}_l}\exp\left(-\frac{y}{\overline{\gamma}_{S,i}}\right)\prod_{j\notin\mathcal{DS}_l}\left(1-\exp\left(-\frac{y}{\overline{\gamma}_{S,j}}\right)\right),\end{aligned} \quad (6)$$

where the second equality comes from the Rayleigh fading assumption, and y has been defined as $y = 2^{2R}-1$ for the sake of notation simplicity. With this last expression, the outage probability of ORS can be written as follows [8]:

$$P_{\text{out}}(y) = \sum_{l=0}^{K}\sum_{\mathcal{DS}_l}\operatorname{Prob}(\text{outage}\mid\mathcal{DS}_l)\operatorname{Prob}(\mathcal{DS}_l), \quad (7)$$

where the second summation is over all the possible decoding subsets $\mathcal{DS}_l$ (i.e., the $\binom{K}{l}$ possible subsets of l relays taken from the K relays). As for Prob(outage | $\mathcal{DS}_l$), this is the probability where the selected relay is in outage conditioned on the fact that the decoding subset is $\mathcal{DS}_l$. In [8], this probability was solved by assuming an ideal scenario with an absolute SNR selection. Our contribution here is to adapt the outage expression to a (realistic) scenario with outdated CSI and a max-normalized SNR strategy. Indeed, the only term in (7) affected by these two particularities is Prob(outage | $\mathcal{DS}_l$). This is because a node belongs to the decoding subset if it has perfectly decoded the information, which is independent of CSI delays and relay selection decisions. Conversely, Prob(outage | $\mathcal{DS}_l$) depends on the relay selection accuracy, and this clearly depends on both ρ_k and how the relay has been selected. When $l = 0$, that probability is clearly equal to 1 as there are no active nodes to relay the transmission. For $l > 0$, we should first define $\mathcal{A}_{k,\mathcal{DS}_l}$ as the event where relay k is selected (i.e., $k^* = k$) under the assumption that the decoding subset is $\mathcal{DS}_l$. By doing so, we can re-rewrite Prob(outage | $\mathcal{DS}_l$) as follows:

$$\begin{aligned}&\operatorname{Prob}(\text{outage}\mid\mathcal{DS}_l)\\ &\quad= \sum_{k\in\mathcal{DS}_l}\operatorname{Prob}(\gamma_{k,D}<y\mid\mathcal{A}_{k,\mathcal{DS}_l})\operatorname{Prob}(\mathcal{A}_{k,\mathcal{DS}_l})\\ &\quad= \sum_{k\in\mathcal{DS}_l}\int_0^\infty F_{\gamma_{k,D}\mid\hat{\gamma}_{k,D}}(y\mid\hat{\gamma}_{k,D})\\ &\qquad\times f_{\hat{\gamma}_{k,D}\mid\mathcal{A}_{k,\mathcal{DS}_l}}(\hat{\gamma}_{k,D}\mid\mathcal{A}_{k,\mathcal{DS}_l})\,d\hat{\gamma}_{k,D}\operatorname{Prob}(\mathcal{A}_{k,\mathcal{DS}_l})\\ &\quad= \frac{1}{l}\sum_{k\in\mathcal{DS}_l}\int_{\gamma_{k,D}=0}^{y}\int_{\hat{\gamma}_{k,D}=0}^{\infty}f_{\gamma_{k,D}\mid\hat{\gamma}_{k,D}}(\gamma_{k,D}\mid\hat{\gamma}_{k,D})\\ &\qquad\times f_{\hat{\gamma}_{k,D}\mid\mathcal{A}_{k,\mathcal{DS}_l}}(\hat{\gamma}_{k,D}\mid\mathcal{A}_{k,\mathcal{DS}_l})\,d\gamma_{k,D}\,d\hat{\gamma}_{k,D},\end{aligned} \quad (8)$$

where $F(\cdot)$ stands for the cumulative density function (CDF), $\operatorname{Prob}(\mathcal{A}_{k,\mathcal{DS}_l})$ is equal to $1/l$ due to the fairness property of the proposed relay selection strategy (i.e., all the normalized estimated SNRs have the same statistics), and $f_{\gamma_{k,D}\mid\hat{\gamma}_{k,D}}(\gamma_{k,D}\mid\hat{\gamma}_{k,D})$ is given by (4). Note that $f_{\hat{\gamma}_{k,D}\mid\mathcal{A}_{k,\mathcal{DS}_l}}(\hat{\gamma}_{k,D}\mid\mathcal{A}_{k,\mathcal{DS}_l})$ can be easily computed since this relay selection problem is statistically equivalent to the scheduling problem observed in a multiuser broadcast channel with independently distributed Rayleigh fading channels and a max-normalized SNR scheduler. More specifically, the following equation can be obtained [22]:

$$\begin{aligned}&f_{\hat{\gamma}_{k,D}\mid\mathcal{A}_{k,\mathcal{DS}_l}}(\hat{\gamma}_{k,D}\mid\mathcal{A}_{k,\mathcal{DS}_l})\\ &\quad= l\frac{\exp\left(-\hat{\gamma}_{k,D}/\overline{\gamma}_{k,D}\right)}{\overline{\gamma}_{k,D}}\left(1-\exp\left(-\frac{\hat{\gamma}_{k,D}}{\overline{\gamma}_{k,D}}\right)\right)^{l-1}.\end{aligned} \quad (9)$$

By plugging (9) and (4) into (8), we obtain an integral equation already solved in a previous work by the authors related with multiuser diversity and delayed CSI [21] (details are omitted for brevity):

$$\begin{aligned}&\operatorname{Prob}(\text{outage}\mid\mathcal{DS}_l)\\ &\quad= \sum_{k\in\mathcal{DS}_l}\sum_{m=0}^{l-1}\binom{l-1}{m}\frac{(-1)^m}{m+1}\\ &\qquad\times\left(1-\exp\left(-\frac{y(m+1)}{\overline{\gamma}_{k,D}(1+(1-\rho_k^2)m)}\right)\right).\end{aligned} \quad (10)$$

Finally, by introducing (10) along with (6) in (7), the outage probability can be written as follows:

$$\begin{aligned}P_{\text{out}}(y) &= \prod_{j=1}^{K}\left(1-\exp\left(-\frac{y}{\overline{\gamma}_{S,j}}\right)\right)\\ &\quad+\sum_{l=1}^{K}\sum_{\mathcal{DS}_l}\sum_{k\in\mathcal{DS}_l}\sum_{m=0}^{l-1}\binom{l-1}{m}\frac{(-1)^m}{m+1}\\ &\qquad\times\left(1-\exp\left(-\frac{y(m+1)}{\overline{\gamma}_{k,D}(1+(1-\rho_k^2)m)}\right)\right)\\ &\qquad\times\prod_{i\in\mathcal{DS}_l}\exp\left(-\frac{y}{\overline{\gamma}_{S,i}}\right)\prod_{j\notin\mathcal{DS}_l}\left(1-\exp\left(-\frac{y}{\overline{\gamma}_{S,j}}\right)\right),\end{aligned} \quad (11)$$

where the first term is related to the case that the decoding subset is an empty set (i.e., $l = 0$).

Finally, it is worth noting that although the analysis has been carried out from an information theoretic point of view, it can be readily extended to a practical scheme with adaptive coding and modulation (e.g., a WiMAX system). Notice that the expression derived in this section evaluates the probability of having instantaneous SNR lower than a specified value given by the Shannon capacity, y, and this value can be set equal to the different SNR thresholds of the adaptive coding and modulation modes.

3.2. Numerical Evaluation. As far as numerical evaluation is concerned, special attention has been paid to carry out a fair comparison in a realistic scenario. It has been considered the wireless scenario presented in Section 2 with a parallel relay topology as shown in Figure 1, where the distance of the source-to-destination link is $d = 100$ meters, the carrier frequency is set to $f_c = 3.5\,\text{GHz}$ (in close alignment with the commercial WiMAX equipments deployed in the European Community), the target rate is $R = 1$ bits/seg/Hz, and the number of relays is $K = 5$. In order to obtain the outage probability of the proposed system, we adopt Monte Carlo simulation, where in each realization the different channels ($h_{S,k}$, $h_{k,D}$, and $\widehat{h}_{k,D}$) are modeled as described in Sections 2.1 and 2.3. Finally, we define system SNR as the average received SNR of the single-hop scheme. For each value of system SNR, the cooperative schemes use the same total power P as that needed by the single-hop scenario to achieve this SNR value. By doing so, we are fairly evaluating the advantage of using cooperation as the total transmit power of the system is kept constant. Besides, for the sake of benchmarking, we compare the outage probability of the proposed cooperative scheme with that obtained without cooperation and the following relay selection strategies.

(i) *Round robin*. This strategy is theoretically the fairest strategy as it is based on iteratively selecting the different relays of the decoding subset.

(ii) *Conventional ORS (max SNR)*. Clearly this technique does not care about fairness among relays as it selects the relay with the maximum absolute SNR.

As observed in Figure 3, the outage probability expression derived in the previous subsection completely agrees with the simulation results. It is also observed that the proposed max-normalized SNR strategy is able to extract the diversity gains of the cooperative system as results corresponding with $\rho = 1$ are quite overlapped with those obtained with the (outage optimal) max SNR scheme. However, performance of both strategies is quite sensitive to the value of ρ. Outage performance is significantly affected when ρ moves away from 1. In particular, one can observe that only a slight improvement can be obtained by using ORS-based cooperation with respect to a direct transmission strategy when $\rho = 0.5$. Apart from that, it is also observed that the gap between the max-normalized and max SNR strategies becomes wider for decreasing values of ρ. This is because the higher SNR peaks generation capability of the conventional ORS strategy compensates more efficiently the CSI uncertainties.

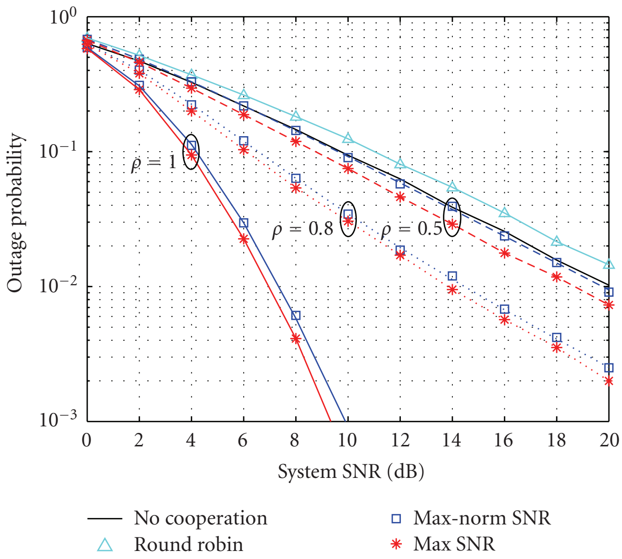


Figure 3: Outage probability versus system SNR for the different communication strategies and values of ρ. For the max-normalized SNR strategy, symbols are associated with the simulated results whereas lines correspond to the theoretical expression. ($K = 5$ relays, $R = 1$ bit/s/Hz, $d = 100$ m).

As for the round-robin strategy, it is clearly observed that this is not a useful technique in terms of outage probability as better performance can be obtained without cooperation. This is mainly due to the fact that better results can be obtained by concentrating total power and transmission time in a single-hop communication instead of dividing them between the source and a relay terminal that has been selected (data link layer) without CSI (physical layer) considerations. It is then emphasized the need of adopting cross-layer strategies in the design of cooperative communication systems.

4. Fairness Analysis

In the previous section, we have explored the performance of the different transmission techniques in terms of outage probability. Nonetheless, this analysis has been performed without considering the fairness among selected relays; this last issue is important to improve the acceptance by the different users of cooperation mechanisms. In this section, we concentrate our efforts on the study of the fairness behavior of the different relay selection mechanisms, and we show that there exists a trade-off in terms of system performance versus fairness among relays. To do so, we use a graphical representation based on modern *portfolio* theory that helps to easily quantify such trade-off.

4.1. Fairness Criterion. In this work, we measure the fairness among relays in terms of the percentage of power consumption used for relaying purposes. This metric was also adopted in [17] but, here, some differences are observed as we consider a scenario based on decode and forward where the power used by the selected relays remains constant. In the proposed scenario, in particular, the power consumption destined to cooperation purposes is originated by the following mechanisms.

(1) *Receiving procedure*. In the first time-slot of the decode and forward procedure, the receiver circuitry of each relay consumes power to receive the signal and to measure the SNR in order to estimate if the relay is able to decode signal.

(2) *Relay selection mechanisms*. According to the relay selection strategies presented in Section 2.2, relays belonging to the decoding subset dedicate battery power to the following actions:

   (i) *FDD*: battery power is mainly used to transmit the signaling message to the destination indicating that the relay is able to retransmit the information.

   (ii) *TDD*: power consumption is mainly caused by the internal timing procedure and, in the case of the best relay, by the transmission of the flag packet to the rest of relays.

(3) *Decoding and retransmission procedure*. Once the relay selection procedure is finished, the selected relay decodes/encodes the source's information and retransmits it to the destination. Clearly, this is the most power demanding mechanism where the fair behavior of the relay selection strategy plays a crucial role.

As will be commented in the next subsection, we study the fairness by analyzing the standard deviation of power consumption among relays (adopting a similar approach than that presented in [17]). Therefore, mechanism (1) described above does not affect the standard deviation measure as all the relays perform that procedure. Basically, differences among relays will be observed in mechanisms (2) and (3). However, because mechanism (2) is carried out by all the relays in the decoding subset and the involved power consumption can be neglected in comparison with that destined to (3), we focus our study in the analysis of the *decoding and retransmission procedure*. In such a procedure, a fix amount of power is consumed when it is executed. On one hand, decoding and encoding the source's message always need the same power budget. On the other hand, the proposed scenario considers that selected relays transmit with the assigned constant power $P_k^* = 0.5P$. As a result, computing the amount of percentage of power allocated to each relay is equivalent to obtaining the percentage of time where each relay is active. In such circumstances, *the standard deviation of the percentage of power consumption of the different relays is obtained in this work by computing the standard deviation of the fraction of time periods where relays are activated for relaying the source's information*. For that reason, we propose the use of the max-normalized SNR strategy as all the relays in the decoding subset will be chosen with the same probability. As commented previously, the behavior of the proposed strategy could be quite different when a different relaying protocol is adopted (see, e.g., [17]).

4.2. System Performance versus Fairness Trade-offs Representation. As observed in Section 3.2, the fair behavior provided by the max-normalized SNR and round-robin strategies penalizes system performance (specially for decreasing values of ρ in the former case). Therefore, it seems that there exists a trade-off in terms of the degree of fairness among the different relays and its impact in terms of system performance. In this section, we are devoted to show the existence of such a trade-off with the help of an analysis tool inspired by means versus standard deviation plots adopted in modern *portfolio* theory [9, 10]. This kind of representation is used in financial market theory with the aim of assessing the existing trade-offs in terms of the expected profit (mean) versus the possible risk (standard deviation) when a possible investment is considered. In this work, we adapt such representation to the proposed wireless scenario based on cooperative communications by illustrating the gain in terms of system performance (outage probability) versus the difference among relays in terms of power consumption (standard deviation of the percentage of power consumption). By doing so, we can easily quantify what the performance versus fairness trade-off of the different relaying strategies is.

Before analyzing the behavior of the different relaying schemes, it is worth mentioning that this portfolio-based representation is also adopted in several works related with the design of resource allocation mechanisms in wireless networks. More specifically, Bartolome introduced this methodology in the wireless communications community to study the degree of fairness of the MIMO Broadcast Channel with zero-forcing transmit beamforming when different bit allocation techniques are adopted [23]. By using the mean versus standard deviation plots, trade-offs in terms of global rate versus fairness among users were easily showed. Then, it was proved that this approach facilitates the design and comparison of different resource allocation algorithms according to the desired degree of fairness. This technique can also be found in studies about the comparison of optimum versus zero-forcing beamforming [24], design of fair algorithms in a context where an orthogonal linear precoding is adopted [25, 26], and the study of the robustness of multiuser systems against CSI imperfections [27].

In Figure 4, the outage probability versus the standard deviation of the power consumption of the different relays is represented for the relay selection mechanisms discussed in the previous section, where each point in the plot of the ORS-based cooperation mechanisms (max-norm SNR and max SNR) is related with a different ρ (with $\rho = \{0.1, 0.5, 0.8, 1\}$). We start the analysis by considering a scenario with system SNR equal to 10 dB. Although the consideration of the direct transmission could not make sense here, we have included the outage probability of this case in order to assess if system performance gain obtained with a cooperative strategy justifies the battery consumption of the terminals for relaying purposes. Notice that the standard deviation of the direct transmission case is set equal to 0. Besides, it is also worth noting that the standard deviation of the ORS-based mechanisms does not depend on parameter ρ as relay selection decisions are independent of the level of

CSI inaccuracy. In other words, the standard deviation of the power consumption depends on the degree of fairness applied by the ORS-based schemes on the relay selection procedure, but for a given degree of fairness, it is only the outage probability that depends on the quality of the channel estimate but not the power consumption distribution.

As observed in the figure, the highest standard deviation is obtained with the max SNR strategy. Clearly, it is observed how the good performance results of the conventional ORS strategy are attained at the expense of a considerable reduction in terms of fairness. Indeed, the standard deviation observed in that case amounts to approximately 13%, resulting in a faster battery consumption of those relays with better channel conditions. Concerning the max-normalized SNR and round robin strategies, the fairer behavior of these strategies is reflected by the lower standard deviation obtained in these cases (1.6% and 2%, resp.).

Surprisingly, the fairest cooperative strategy is the max-normalized SNR strategy instead of the round robin one. The round robin scheme iteratively selects the different relays of the decoding subset. In the case of low and medium system SNRs, the probability that the decoding subset has all the relays of the system is reduced. In these circumstances, relays closer to the source have a higher probability to be able to retransmit the signal and, thus, to belong to the decoding subset. Then, the power consumption of these relays in relaying procedures is higher than that used by the rest of relays. When the rest of relays are in the decoding subset, the relay selection mechanism selects them iteratively without taking into account that these relays have not been activated for too long, and some actions should be adopted in order to compensate this situation. In the max-normalized SNR strategy, however, relays are selected when their SNRs are in their own peaks, and, then, some compensation actions are implicitly carried out by the selection strategy.

The origin of this last effect is clarified by analyzing in Figure 4 results corresponding to a scenario with system SNR equal to 20 dB. As observed, the standard deviation of both the round robin and max-normalized SNR strategies is quite similar. In that case, the decoding subset has the K relays of the system with a high probability, and, then, the problems reducing the fair behavior of these strategies are alleviated. In the figure, one can also observe that the conventional ORS strategy is less fair when the system SNR is increased. This is because in the low- and mid-SNR regimes situations where the decoding subset is only formed by the *worst* relays can happen. In those cases, the relay selection mechanism will activate a subset of relays that never will be chosen when all the relays of the system are in the decoding subset. In order to extend such analysis, we also present a graphical representation where the SNR dependance of the system is clearly reflected (see Figure 5). As observed in the figure, when the SNR of the system is increased, the fairness of the round robin and max-normalized SNR strategies is improved, whereas the system becomes less fair in the max SNR case due to the reasoning discussed above.

As for the existing trade-offs in terms of system performance versus fairness, one can easily assess the behavior of the different strategies thanks to the proposed representation. More specifically the following conclusions can be drawn.

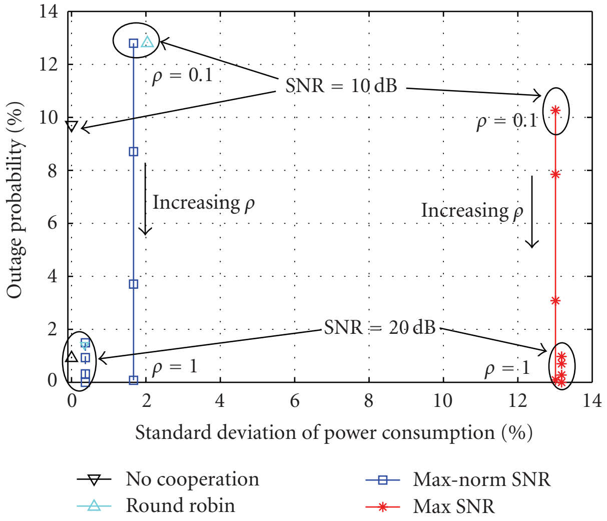


FIGURE 4: Outage probability versus standard deviation of the power consumption of the different relay selection mechanisms for different values of ρ and System SNR ($\rho = \{0.1, 0.5, 0.8, 1\}$, $K = 5$ relays, $R = 1$ bit/s/Hz, $d = 100$ m. Solid line: System SNR = 10 dB, dashed line: System SNR = 20 dB).

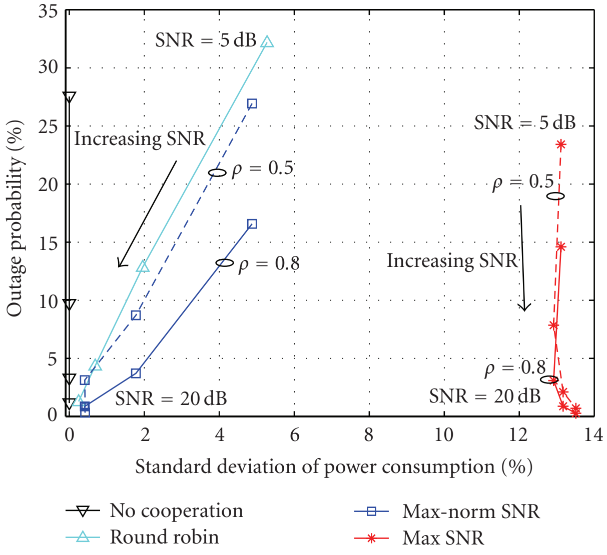


FIGURE 5: Outage probability versus standard deviation of the power consumption of the different relay selection mechanisms for different values of ρ and System SNR (System SNR = $\{5, 10, 15, 20\}$ dB, $K = 5$ relays, $R = 1$ bit/s/Hz, $d = 100$ m. Solid line: $\rho = 0.8$, dashed line: $\rho = 0.5$).

(i) The best performance results are obtained with the conventional ORS strategy. However, the fairness of the system is considerably penalized.

(ii) An appropriate strategy to exploit cooperative diversity while keeping a good performance versus fairness trade-off is the max-normalized SNR strategy.

Indeed, it is shown that this strategy can present a better fairness behavior than that provided by round robin.

(iii) For low ρ values and high system restrictions in terms of outage probability, conventional ORS strategy could be the most appropriate strategy. For high ρ values, however, it is clear that more benefits are obtained with max-normalized SNR as similar results are obtained in terms of outage probability but the fairness among relays is substantially improved.

(iv) The round robin strategy is not useful for exploiting cooperation benefits.

Finally, one can also notice that the proposed representation helps to assess the viability of using a cooperative technique as direct transmission results have also been included in the figures. In particular, one can observe in Figures 4 and 5 that it could be better to use a direct transmission when the SNR is high and/or CSI is not accurate enough (low ρ values). This is because, similar outage probability results can be obtained without destining battery power to cooperation purposes.

5. Conclusions

In this work, we have studied the impact of outdated CSI in cooperative systems. The analysis has been carried out in terms of the trade-off of outage probability versus fairness of the system. To do so, an analytical expression has been obtained for the outage probability of an ORS scenario, whereas the difference among relays in terms of power consumption has been considered as a fairness measure and obtained by means of simulations. In order to assure a good balance in terms of performance versus fairness, we have proposed a relay selection strategy based on the max-normalized SNR criterion. The proposed strategy has been compared with existing relay selection strategies with the help of an analysis plot inspired in modern portfolio theory. In particular, we have represented the existing trade-offs of the different relaying mechanisms by plotting the outage probability versus the standard deviation of the power consumption. It has been shown that the max-normalized SNR guarantees a good performance versus fairness trade-off when available CSI is sufficiently accurate. When CSI is not accurate enough, however, direct transmission could be a better strategy.

Acknowledgement

This work was supported by the Spanish Government Project TEC2008-06305/TEC.

References

[1] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—part I: system description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1938, 2003.

[2] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—part II: implementation aspects and performance analysis," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1939–1948, 2003.

[3] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

[4] R. U. Nabar, H. Bölcskei, and F. W. Kneubühler, "Fading relay channels: performance limits and space-time signal design," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 1099–1109, 2004.

[5] D. Gesbert, M. Shafi, D.-S. Shiu, P. J. Smith, and A. Naguib, "From theory to practice: an overview of MIMO space-time coded wireless systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 3, pp. 281–302, 2003.

[6] IEEE 802.16's Relay Task Group, http://wirelessman.org/relay.

[7] A. Ghosh, D. R. Wolter, J. G. Andrews, and R. Chen, "Broadband wireless access with WiMax/802.16: current performance benchmarks, and future potential," *IEEE Communications Magazine*, vol. 43, no. 2, pp. 129–136, 2005.

[8] A. Bletsas, H. Shin, and M. Z. Win, "Cooperative communications with outage-optimal opportunistic relaying," *IEEE Transactions on Wireless Communications*, vol. 6, no. 9, pp. 3450–3460, 2007.

[9] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[10] H. Markowitz, "Foundations of portfolio theory," *The Journal of Finance*, vol. 46, no. 2, pp. 469–477, 1991.

[11] B. Zhao and M. C. Valenti, "Practical relay networks: a generalization of hybrid-ARQ," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 7–18, 2005.

[12] L. Dai and L. J. Cimini Jr., "Improved fairness in energy-constrained cooperative ad-hoc networks," in *Proceedings of the 40th Annual Conference on Information Sciences and Systems (CISS '06)*, pp. 734–738, Princeton, NJ, USA, March 2006.

[13] L. Dai, W. Chen, K. B. Letaief, and Z. Cao, "A fair multiuser cooperation protocol for increasing the throughput in energy-constrained ad-hoc networks," in *Proceedings of IEEE International Conference on Communications (ICC '06)*, vol. 8, pp. 3633–3638, Istanbul, Turkey, June 2006.

[14] W. Chen, L. Dai, K. B. Letaief, and Z. Cao, "Fair and efficient resource allocation for cooperative diversity in ad-hoc wireless networks," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '07)*, pp. 4099–4104, Hong Kong, March 2007.

[15] I. Krikidis and J. C. Belfiore, "Three scheduling schemes for amplify-and-forward relay environments," *IEEE Communications Letters*, vol. 11, no. 5, pp. 414–416, 2007.

[16] W.-J. Huang, F.-H. Chiu, C.-C. J. Kuo, and Y.-W. Hong, "Comparison of power control schemes for relay sensor networks," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, vol. 3, pp. 477–480, Honolulu, Hawaii, USA, April 2007.

[17] D. S. Michalopoulos and G. K. Karagiannidis, "PHY-layer fairness in amplify and forward cooperative diversity systems," *IEEE Transactions on Wireless Communications*, vol. 7, no. 3, pp. 1073–1083, 2008.

[18] J. N. Laneman, *Cooperative diversity in wireless networks: algorithms and architectures*, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Mass, USA, August 2002.

[19] A. Goldsmith, *Wireless Communications*, Cambridge University Press, Cambridge, UK, 2005.

[20] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 659–672, 2006.

[21] J. L. Vicario and C. Antón-Haro, "Analytical assessment of multi-user vs. spatial diversity trade-offs with delayed channel state information," *IEEE Communications Letters*, vol. 10, no. 8, pp. 588–590, 2006.

[22] L. Yang and M.-S. Alouini, "Performance analysis of multiuser selection diversity," in *Proceedings of IEEE International Conference on Communications (ICC '04)*, vol. 5, pp. 3066–3070, Paris, France, June 2004.

[23] D. Bartolomé, *Fairness analysis of wireless beamforming schedulers*, Ph.D. dissertation, Universitat Politecnica de Catalunya, Barcelona, Spain, January 2005.

[24] M. Bengtsson, D. Bartolomé, J. L. Vicario, and C. Antón-Haro, "Beamforming and bit-loading strategies for multi-user SDMA with admission control," in *Proceedings of the 16th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '05)*, vol. 2, pp. 842–846, Berlin, Germany, September 2005.

[25] R. Bosisio, G. Primolevo, O. Simeone, and U. Spagnolini, "Fair scheduling and orthogonal linear precoding/decoding in broadcast MIMO systems," in *Proceedings of the 16th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '05)*, vol. 2, pp. 832–836, Berlin, Germany, September 2005.

[26] A. Zagami, *Beamforming and bit-loading strategies for multiuser MIMO systems*, M.S. thesis, Universitat Politecnica de Catalunya and Politecnico di Torino, Barcelona, Spain, June 2006.

[27] J. L. Vicario and C. Antón-Haro, "A unified approach to the analytical assessment of multi-user diversity with imperfect channel state information: ergodic capacity and robustness analysis," *European Transactions on Telecommunications*, vol. 18, no. 6, pp. 573–582, 2007.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 467182, 16 pages
doi:10.1155/2009/467182

Research Article

Cross-Layer Optimal Rate Allocation for Heterogeneous Wireless Multicast

Amr Mohamed[1] **and Hussein Alnuweiri**[2]

[1] *Department of Computer Engineering, Qatar University, P.O. Box 2317, Doha, Qatar*
[2] *Department of Electrical Engineering, Texas A&M University at Qatar, P.O. Box 23874, Doha, Qatar*

Correspondence should be addressed to Amr Mohamed, amrm@ece.ubc.ca

Received 30 March 2008; Revised 5 December 2008; Accepted 12 January 2009

Recommended by Ekram Hossain

Heterogeneous multicast is an efficient communication scheme especially for multimedia applications running over multihop networks. The term heterogeneous refers to the phenomenon when multicast receivers in the same session require service at different rates commensurate with their capabilities. In this paper, we address the problem of resource allocation for a set of heterogeneous multicast sessions over multihop wireless networks. We propose an iterative algorithm that achieves the optimal rates for a set of heterogeneous multicast sessions such that the aggregate utility for all sessions is maximized. We present the formulation of the multicast resource allocation problem as a nonlinear optimization model and highlight the cross-layer framework that can solve this problem in a distributed ad hoc network environment with asynchronous computations. Our simulations show that the algorithm achieves optimal resource utilization, guarantees fairness among multicast sessions, provides flexibility in allocating rates over different parts of the multicast sessions, and adapts to changing conditions such as dynamic channel capacity and node mobility. Our results show that the proposed algorithm not only provides flexibility in allocating resources across multicast sessions, but also increases the aggregate system utility and improves the overall system throughput by almost 30% compared to homogeneous multicast.

1. Introduction

The one-hop broadcast characteristic of the MAC layer in wireless ad hoc networks has triggered the use of multicast communication scheme as one of the natural strategies that can multiply the overall network throughput with very limited overhead. This is because multicast packets are forwarded once to reach all the multicast members in the neighborhood using a single transmission, and this effect increases even more in multihop ad hoc networks.

Heterogeneous multicast, also called *multirate* multicast, is an efficient mode of data delivery for many multimedia applications, especially those operating in real time such as audio/video teleconferencing and TV broadcasting. In multirate multicast, the receivers of a multicast group are offered service at different rates commensurate with their capabilities (e.g., processing power limitations) or based on their local network conditions (e.g., surrounding wireless link states). Therefore, multirate schemes have a great advantage over unirate multicast (or homogeneous multicast) in adapting to diverse receiver requirements and heterogeneous network conditions.

The simplest way of attaining multirate multicast is by frame dropping. In this approach, intermediate nodes in a multicast tree may drop data frames to lower the rate for the downstream nodes. Another way is by hierarchical encoding or layered streaming which is particularly suitable for audio/video traffic. In this approach, the sender provides data in several layers organized in a hierarchy. Receivers subscribe to the layers cumulatively to provide progressive refinement [1]. This means that the receiver can only choose from a discrete set of data rates on each link. Another method of attaining multirate multicast which is particularly suitable for overlay multicast [2] is stream adaptation through transcoding [3] using intermediate media gateways, thus allowing the receivers to choose their streaming rates from a broader continuous range. We assume that the network has one or more of these capabilities.

In this paper, we present an optimal resource allocation algorithm for heterogeneous multicast over wireless ad hoc networks. Multirate multicast has a distinct advantage compared to unirate multicast especially for optimal resource allocation. This is because unirate multicast techniques are often unable to efficiently allocate network resources for multicast groups that have some congested group members (receivers). For such multicast groups, unirate multicast techniques tend to allocate rates based on the most congested receivers potentially wasting significant network resources. On the other hand, multirate multicast allows the rate to change for designated tree members to accommodate the congested receivers downstream. Hence, it provides more flexibility in allocating rates across the multicast tree such that the overall network resource utilization is maximized (see the example in Section 2). Our heterogeneous multicast solution has the following key features.

(i) It guarantees optimal resource utilization while providing system-wide fairness for end-to-end multirate multicast flows.

(ii) It guarantees steering the *entire* network toward the optimal point in real time, and hence reacts robustly to network conditions (e.g., mobility and route changes) as they occur.

(iii) It is based on *primal-dual* and *pricing* methods which facilitate the decomposition of the resource optimization problem into subproblems that are easier to solve in a modular structure.

(iv) For network deployment, we design a cross-layer framework that utilizes a measurement-based technique for MAC-layer channel capacity estimation, and a light-weight network HELLO protocol for constructing contention domains to allow for allocating resources across end-to-end multicast sessions.

(v) This cross-layer solution also works in a truly distributed network environment, with limited overhead, and with no synchronization requirements between node calculations.

The problem of resource allocation for unicast flows has been investigated before in [4–6]. In these works, common pricing mechanism has been used whereby each network resource calculates a *price* that represents the relationship between the load on the network resource and the capacity that it can offer. Resource allocation for multirate multicast in wired networks has been studied in [7, 8]. An iterative algorithm based on subgradient techniques [9] has been employed to account for the nondifferentiability of the primal problem. The authors in [2] proposed an overlay strategy for allocating resources over a multirate multicast tree by considering each link as a point-to-point unicast session. Rates are then allocated across each unicast session such that the aggregate utility across all unicast sessions is maximized. The problem of optimal and fair resource allocation has been widely studied in the context of wired networks. Among these studies (e.g., [2, 4, 5, 7, 8]), price-based methods have shown to be effective in achieving a decentralized solution for rate allocation. The location-based contention, time-varying wireless channel characteristics, and multirate multicast in one-hop broadcast wireless medium represent both challenges and opportunities which we addressed in our model.

The remainder of this paper is organized as follows. In Section 2, we explain the terminology used for heterogeneous multicast and formulate the optimization problem. The approach for multirate multicast is presented in Section 3. We present our distributed asynchronous algorithm for heterogeneous multicast in Section 4. We provide the simulation results in Section 5 and finally, we conclude this paper in Section 6.

2. Model and Problem Formulation

2.1. Model and Notations. Table 1 highlights the notations used by the model. We consider a wireless ad hoc network consisting of a set of nodes V spread over a wireless space, each with a specific transmission range and interference range. We exploit the protocol model explained in [10] and leveraged in [6] for wireless packet transmission. In this model, the transmission from node i is successfully received by node j $(i, j \in V)$ if (1) the distance between the two nodes is no more than a certain range (i.e., transmission range), and (2) for all other nodes $k \in V$ simultaneously transmitting over the same channel, the distance between j and k is more than a specific range (i.e., the interference range). For some protocols which require acknowledgment from j to i (e.g., IEEE 802.11 MAC), node i is also required to be interference free at the time of sending the acknowledgment.

We model the wireless ad hoc network as a *directed* graph $G = (V, E)$, where E is the set of wireless "virtual" links produced as a result of nodes located within the transmission range of each other. Each wireless link $e \in E$ has two end nodes i and j (i.e., $e = \{i, j\}$). The network is shared by a set of M *end-to-end* multicast groups. Each multicast group m has a unique source node s_m, a set of receivers $R_m = \{r_{m1}, r_{m2}, \ldots\}$, and uses a subset of wireless links E_m and a subset of nodes V_m for either receiving or relaying traffic. Note that $R_m \subseteq V_m$.

We further divide the multicast tree nodes into *gateway nodes* and *relay nodes* as shown in Figure 1. Gateway nodes are the nodes that have rate control capabilities through one of the methods explained before, such as layered transmission, transcoding and frame dropping. Relay nodes on the other hand merely forward data frames without performing any rate change. We use v_i to denote a gateway at node i. If v_i is a member of multicast tree m, hence denoted by v_{mi}, then v_i can control the rate of the downstream nodes.

A fundamental difference between the unicast and multicast cases is the fact that one-hop broadcast may be used to transfer traffic from one sending node to one or more receiving nodes. To capture this notion, the one-hop data transmission from one sending node i to a set of receiving nodes $J \subseteq V_m$ within the multicast flow m along *one* or

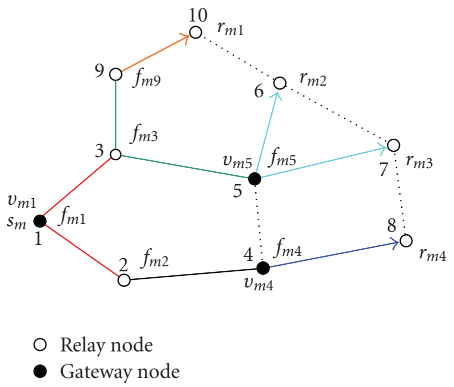


Figure 1: Multirate multicast network model.

more wireless links (branches) is referred to as a multicast subflow of m or f_{mi}. Each multicast subflow uses one or more branches b_{mij} from one sending node i to a set of receiving nodes $J \subseteq V_m$, that is, $f_{mi} = \{b_{mij} : \forall b_{mij} = \{i, j\}\ j \in J\}$, with a cardinality K_{mi} equal to the number of branches of f_{mi}. We also define an active wireless link $a_{ij} \in E$ to be the wireless link that carries traffic from at least one multicast group, and $A \subseteq E$ is the set of all active links. Also, a_{iJ} refers to the *aggregated* multicast subflow from node i and is represented by the set of active links $a_{ij}\ \forall j \in J$ that are used by one or more multicast subflows $f_{mi}\ \forall m \in M$ simultaneously.

For simplicity, we will assume that a packet is successfully transmitted over a multicast subflow f_{mi} if (1) the packet reaches all receiving nodes J on all the branches b_{mij}; (2) acknowledgments (using the notation of IEEE 802.11 MAC standards) have been transmitted successfully from all these receiving nodes back to the sending node i [11]. Based on this assumption, the protocol model can be extended for multicast subflows as follows: the traffic from two different subflows on a group of active wireless links contend if either the sending node or *any* of the receiving nodes of one subflow are within the interference range of the sending node or *any* of the receiving nodes of the other subflow.

The multicast subtree starting from gateway node v and *ending* at either a terminal node or another gateway node is denoted by T_v. In Figure 1, $T_{v_{m1}}$ starts at node 1 and ends at the set of nodes $\{4, 5, 10\}$, and $T_{v_{m5}}$ starts at node 5 and ends at the set of nodes $\{6, 7\}$. This set of terminal nodes for subtree T_v is denoted by $\mathfrak{I}_v$.

Each multicast subtree T_v has a rate x_v (expressed in bits/s) which is allowed to vary within the rate interval $I_v = [w_v, W_v]$ [5], and I is the set of all such intervals. F_v denotes the set of multicast subflows that belong to subtree T_v. $\Upsilon_m = \{v_{m1}, v_{m2}, \ldots\}$ is the set of all gateway nodes that are members of group m, and Υ is the set of all gateway nodes on all multicast trees $\forall m \in M$. Each multicast group has at least one gateway node (i.e., group source is considered a gateway node) to control the rate to the downstream nodes. We use the notation $\pi_m(v)$ to indicate the parent gateway node of gateway node v by going upstream toward the source of group m (e.g., $v_{m1} = \pi(v_{m4})$). Note that the source node has no parent gateway node (i.e., $\pi(v_{m1}) = \varnothing$).

Also, note that for one multicast group m, any gateway node in the path between the source node and any receiver node may control the transmission by reducing the rate on this path to improve the overall network resource utilization (see the example in Section 2). Therefore, the rate that a gateway node is using for transmission at any given time must be greater than or equal to the maximum rate of all downstream gateway/receiver nodes. For example, in Figure 1, the rate used by v_{m1} for transmission must be greater than or equal to the maximum rate used by any of the gateway nodes v_{m4}, or v_{m5}. This adds a set of new constraints to the resource allocation problem which can be formulated by the following linear inequalities:

$$x_v \leq x_{\pi_m(v)} \quad \forall v \in \Upsilon_m : \pi_m(v) \neq \varnothing \quad \forall m \in M, \qquad (1)$$

where x_v is the rate used by the gateway node v, and $x_{\pi_m(v)}$ is the rate used by parent gateway of gateway node v across the multicast group m.

To model the contention between the active wireless links, we use a contention domain mechanism [12] by forming a logical contention graph $G_c = (V_c, E_c)$. Each vertex in G_c corresponds to the aggregated multicast subflow a_{iJ} which carries the traffic from *one* or *more* subflows simultaneously. A link between two vertices indicates that the traffic on the two aggregated subflows contend with each other. A complete subgraph in G_c is referred to as a *clique*. A *maximal clique* is the clique that is not part of any other clique. This clique represents the maximal set of active wireless links that contend with each other. This means that only one "subflow" within this clique may transmit a packet at a time [6]. Therefore, the sum of the rates over the maximal clique cannot exceed the channel capacity achieved by using a particular scheduling mechanism in the MAC layer (e.g., IEEE 802.11 DCF). The following inequality formulates these set of constraints:

$$\sum_{v:(F_v \cap V_c^q) \neq \varnothing} x_v \leq c_q, \quad \forall q \in Q, \qquad (2)$$

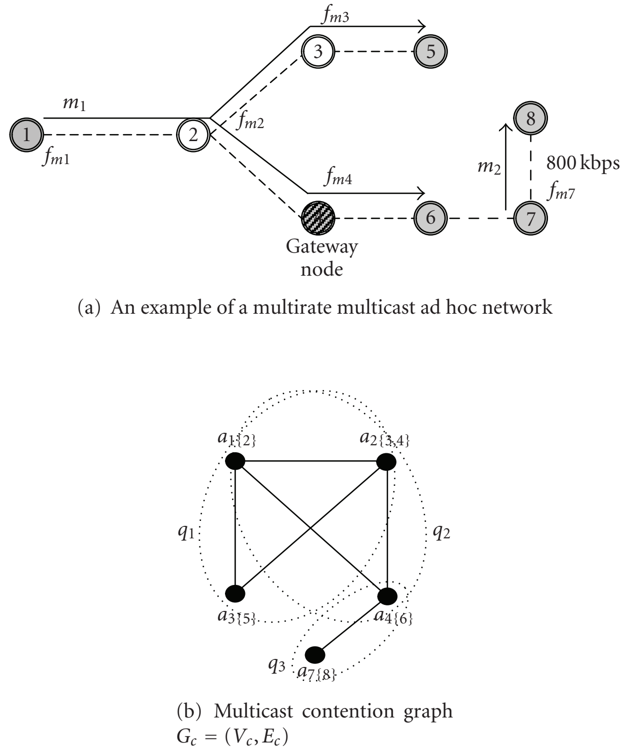
where q is a maximal clique in the set of all maximal cliques Q, c_q is the achieved channel capacity for clique q based on the scheduling discipline used in the MAC layer, and $V_c^q \subseteq A$ are the set of vertices in G_c that belong to clique q.

Next we present an example to illustrate the above notation and highlight the main difference between unirate and multirate multicast with respect to allocating rates in an ad hoc network. Figure 2 shows an example of an ad hoc network where there are 8 nodes connected through wireless links. The network contains 2 sessions m_1, m_2 where m_2 has a traffic with fixed rate 800 kbps. Session m_1 uses node 1 as the group source, and the receiving nodes are 5, 6, whereas m_2 uses node 7 as the group source, and the receiving node is 8. The aggregated subflows are represented by one node in the contention graph as shown in Figure 2(b). Assume that the channel capacity is 1 Mbps, which means that the aggregate rate for each maximal clique cannot exceed 1 Mbps. In this case, the rate on subflow f_{m4} cannot exceed 200 Kbps because the traffic on that subflow contends with f_{m7} and hence they both exist in the same maximal clique.

Table 1: Notations used by heterogeneous multicast model.

V_m	Set of ad hoc nodes on a multicast group m
E_m	Set of virtual wireless links used by a multicast group m
s_m	Source node for a multicast group m
R_m	Set of receivers on a multicast group m
v_{mi}	Gateway on node i controlling downstream nodes on multicast group m
f_{mi}	Subflow starting on node i on multicast group m
a_{iJ}	Set of active wireless links branching from node i to set of nodes J
$T_{v_{mi}}$	Multicast subtree starting at gateway node v_{mi}
x_v	Rate used by gateway node v
$x_{\pi_m(v)}$	Rate used by parent gateway of gateway node v
q	Maximal clique in the set Q
C_q	Estimated channel capacity for clique q
$\Lambda_m(v)$	Set of all children gateways of gateway node v along multicast group m
p_q	Price for utilizing resources on clique q
p'_v	Price due to forwarding traffic by gateway node v
λ_{vi}	Total price incurred by subflow f_{vi} on all cliques
$\lambda_v(i)$	Accumulated price for subtree T_v at node i
$\pi_v(i)$	The parent node of node i along subtree T_v
B	Configurable time window for rate and price calculations

Using unirate multicast, we cannot assign a rate to group m_1 higher than 200 Kbps because one of the receivers in this multicast session is congested. This means that using unirate multicast we allocate the rate based on the most congested receiver. On the other hand, multirate multicast using node 4 as a gateway node can make the rate allocation more efficient because, in this case, the rate used by source node 1 is allowed to exceed 200 Kbps provided that gateway node 4 will adjust this rate to 200 Kbps before forwarding the traffic to the downstream nodes. It can be shown that the rate used by source node 1 can be increased to 333 Kbps in this case.



(a) An example of a multirate multicast ad hoc network

(b) Multicast contention graph $G_c = (V_c, E_c)$

Figure 2: Example for resource allocation in unirate and multirate multicast.

2.2. Mathematical Formulation. First, we assign a utility function $U_v(x_v)$ for each gateway node on every multicast group $m \in M$ to measure the degree of service satisfaction based on assigning a specific rate x_v to that gateway node. An example of a logarithmic utility function to achieve intersession proportional fairness is given in Section 5. The utility function also serves as a network-wide efficient tool for achieving certain fair allocation behavior (e.g., proportional fairness, max-min fairness) as shown in [4]. The optimization problem is to find the set of rates assigned to all gateway nodes for all multicast groups such that the aggregated utility function of all gateway nodes is maximized. This can be formulated with the following modified set of constraints:

$$
\begin{aligned}
\text{(P): maximize } & \sum_{v \in \Upsilon} U_v(x_v) \\
\text{subject to } & \sum_{v:(F_v \cap V_c^q) \neq \varnothing} x_v \Gamma_{qv} \leq c_q, \quad \forall q \in Q, \\
& x_v \leq x_{\pi_m(v)}, \quad \forall v \in \Upsilon_m : \pi_m(v) \neq \varnothing \quad \forall m \in M, \\
& x_v \in I_v, \quad \forall v \in \Upsilon,
\end{aligned} \tag{3}
$$

where Γ_{qv} represents the number of multicast subflows which belong to both clique q and the subtree T_v. Throughout the rest of the paper, we will make the following assumptions to facilitate the solution for the *primal* problem (P).

Assumption 1. There exists at least one vector $\tilde{x} \in I$ such that $\sum_{v\in(F\cap V_c^q)}\tilde{x}_v \leq c_q \ \forall q \in Q$ and $\tilde{x}_v \in I_v$ (i.e., which means $\sum_{v\in(F\cap V_c^q)}\tilde{w}_v \leq c_q \ \forall q \in Q$).

Assumption 2. On the interval I_m, all the utility functions $U_v \ \forall v \in \Upsilon$ are increasing, strictly concave, and twice continuously differentiable.

Note that if we restrict each multicast group to have only one gateway (source) node, then the constraints in (1) will be eliminated and problem (P) will reduce to unirate multicast.

3. Solution Approach

Solving the resource allocation problem (P) with a centralized approach requires the knowledge of the utility functions and the knowledge of all contention domains and multicast groups, which is impractical. Instead we propose a decentralized scheme that minimizes the coordination between networks nodes and adapts robustly to network changes. The key to our solution is the use of the duality theory [13] which suggests solving the dual problem by introducing additional dual variables called *prices* using the same notation as in [4, 6, 8].

The first step is to define the Lagrangian function $L(x, p, p')$ for the optimization problem (P) as follows:

$$\begin{aligned} L(x,p,p') &= \sum_{v\in\Upsilon} U_v(x_v) + \sum_{q\in Q} p_q(c_q - x_v\Gamma_{qv}) \\ &\quad + \sum_{v\in\Upsilon} p'_v(x_{\pi_m(v)} - x_v) \\ &= \sum_{v\in\Upsilon}\left[U_v(x_v) - x_v\left(\sum_{q\in Q} p_q\Gamma_{qv} + p'_v - \sum_{v'\in\Lambda_m(v)} p'_{v'}\right)\right] \\ &\quad + \sum_{q\in Q} p_q c_q, \end{aligned} \tag{4}$$

where $\Lambda_m(v)$ is the set of all children gateway nodes of node v (if any) along multicast session m. Vectors $p = (p_q \ \forall q \in Q)$ and $p' = (p'_v \ \forall v \in \Upsilon)$ are two vectors of Lagrange multipliers. Γ_{qv} represents the number of multicast subflows that belong to subtree T_v and clique q simultaneously. Again we notice that the first term of (4) is separable in x_v, and this entails

$$\begin{aligned} &\max_{x_{v\in I_v}} \sum_{v\in\Upsilon}\left[U_v(x_v) - x_v\left(\sum_{q\in Q} p_q\Gamma_{qv} + p'_v - \sum_{v'\in\Lambda_m(v)} p'_{v'}\right)\right] \\ &\quad = \sum_{v\in\Upsilon}\max_{x_{v\in I_v}}\left[U_v(x_v) - x_v\left(\sum_{q\in Q} p_q\Gamma_{qv} + p'_v - \sum_{v'\in\Lambda_m(v)} p'_{v'}\right)\right]. \end{aligned} \tag{5}$$

Which means that this objective function can be divided into $|\Upsilon|$ separate subproblems. Each subproblem for subtree T_v can be solved locally if the values of clique prices $p_q \ \forall q : (F_v \cap V_c^q) \neq \varnothing$, gateway forwarding price p'_v, and all children gateway forwarding prices $p'_{v'} \ \forall v' \in \Lambda_m(v)$ are known. The objective function of the dual problem then becomes

$$\begin{aligned} &D(p,p') \\ &\quad = \max_{x_v\in I_v} L(x,p,p') \\ &\quad = \sum_{q\in Q} p_q c_q \\ &\quad + \sum_{v\in\Upsilon}\max_{x_{v\in I_v}}\left[U_v(x_v) - x_v\left(\sum_{q\in Q} p_q\Gamma_{qv} + p'_v - \sum_{v'\in\Lambda_m(v)} p'_{v'}\right)\right], \end{aligned} \tag{6}$$

and the dual problem (D) for the primal problem (P) as explained in [13] can then be defined as follows:

$$(\mathrm{D}): \min_{\substack{p\geq 0\\ p'\geq 0}} D(p). \tag{7}$$

Equation (7) suggests that to find the optimal rates in a decentralized fashion, we need to find the optimal prices p and p' by solving the constraint-less problem (D). In the following, we will see that p' can be calculated locally at each gateway node and p can also be calculated locally for each contention domain, hence decentralized solution for end-to-end optimal rates is possible as will be discussed later.

3.1. Interpretation of Prices. Consider $P_v(T_v)$ as the profit of the subtree T_v which can be defined as follows:

$$P_v(T_v) = U_v(x_v) - x_v\left(\sum_{q\in Q} p_q\Gamma_{qv} + p'_v - \sum_{v'\in\Lambda_m(v)} p'_{v'}\right). \tag{8}$$

This profit represents the difference between the utility that subtree T_v gains by having rate x_v (i.e., $U_v(x_v)$) minus the summation of prices (denoted by $\tilde{U}(x_v)$) that this subtree has to pay for gaining such transmission rate, which is defined as

$$\tilde{U}(x_v) = \sum_{q\in Q} p_q\Gamma_{qv}x_v + p'_v x_v - \sum_{v'\in\Lambda_m(v)} p'_{v'}x_v. \tag{9}$$

This summation of prices is divided into three components:

(i) $\sum_{q\in Q} p_q\Gamma_{qv}x_v$ which can be interpreted as the total price for utilizing resources on all cliques $\forall q \in Q$ such that $F_v \cap V_c^q \neq \varnothing$. In this case, p_q can be interpreted as the price per unit bandwidth consumed at clique q.

(ii) $p'_v x_v$ is the price that subtree T_v must pay to the parent subtree of the same group in order to have traffic with rate x_v forwarded to it. In this case, p'_v is the price per unit bandwidth for forwarding traffic to subtree T_v.

(iii) $\sum_{v'\in\Lambda_m(v)} p'_{v'}x_v$ is the total revenue that subtree T_v gains by forwarding traffic with rate x_v to all children subtrees with each term $p'_{v'}x_v$ indicating the revenue for forwarding traffic to subtree $T_{v'}$ such that $v' \in \Lambda_m(v)$.

Note that at optimality, $p'_v = 0$ if $x_v < x_{\pi_m(v)}$ since p'_v indicates the price when the constraints (1) are violated or the maximum possible rate is used (i.e., $x_v = x_{\pi_m(v)}$). This means that a subtree $T_{v'}$ is not charged for using rate x_v if this rate is *less than* the rate at parent gateway node $\pi_m(v)$.

For p_q we can, similarly, define the price for one subflow $f_{vi} \in T_v$ as the total price for consuming bandwidth on all maximal cliques $q \in Q$ as follows:

$$\lambda_{vi} = \sum_{q:(f_{vi} \in V_c^q) \neq \varnothing} p_q. \tag{10}$$

Moreover, we can also define the aggregated price for subtree T_v as a result of consuming bandwidth on all maximal cliques $q \in Q$ as follows:

$$\lambda_v = \sum_{q:(F_v \cap V_c^q) \neq \varnothing} p_q \Gamma_{qv}. \tag{11}$$

A crucial aspect of our solution is how to calculate the individual subtree clique prices $\lambda_v \ \forall v \in \Upsilon$ in a decentralized way given the prices of the individual maximal cliques $p_q \ \forall q \in Q$. To facilitate presentation, we introduce the following new terms:

(1) $\pi_v(i)$: the parent node of node i along subtree T_v;

(2) $\lambda_v(i)$: the accumulated price for subtree T_v at node i.

Note that, along subtree T_v, there is no parent node for the gateway node v, that is, $\pi_v(v) = \varnothing$, and the accumulated price at the $v\lambda_v(v) = 0$. We can then define the accumulated subtree price recursively as follows:

$$\lambda_v(i) = \frac{\lambda_v(\pi_v(i)) + \lambda_{v\pi_v(i)}}{K_{v\pi_v(i)}} \quad \forall i \in T_v, \tag{12}$$

where $K_{v\pi_v(i)}$ is the cardinality of subflow $f_{v\pi_v(i)}$.

Theorem 1. *If $\mathfrak{I}_v$ defines the set of terminal nodes for subtree T_v, then the subtree clique price can be calculated as follows:*

$$\lambda_v = \sum_{i \in \mathfrak{I}_v} \lambda_v(i). \tag{13}$$

Proof is given in Appendix A.

3.2. Aggregated Subtree Price Calculation. In Section 4 we will explain the iterative method for calculating both clique price p_q (hence subflow price from (10)) and the forwarding price for each gateway node p'_v. In order to calculate the total price defined by (9) at any gateway node v, we need to calculate the accumulated price on each branch recursively using (12) until we hit either a terminal node or another gateway node $v' \in \Lambda_m(v)$. Each gateway node $v' \in \Lambda_m(v)$ subtracts the forwarding price $p'_{v'}$ from the accumulated price to get the net price for the branch leading to that gateway node. Children gateway nodes and terminal nodes which are part of T_v then send the net price value back to node v to calculate the subtree aggregate price per unit bandwidth $\lambda(T_v)$ by simply aggregating all net branch prices and the forwarding price p'_v as follows:

$$\lambda(T_v) = \lambda_v + p'_v - \sum_{v' \in \Lambda_m(v)} p'_{v'}. \tag{14}$$

4. Optimal Resource Allocation for Heterogeneous Wireless Multicast (ORAHWM)

We present a distributed iterative algorithm that solves the primal problem (P) by applying the gradient projection method [13] to the dual problem (D). This implies that the clique prices $p_q(t+1) \ \forall q \in Q$ and forwarding prices $p'_v \ \forall v \in \Upsilon$ are calculated iteratively as follows:

$$\begin{aligned} p_q(t+1) &= \left[p_q(t) - \alpha \frac{\partial D(p(t))}{\partial p_q} \right]^+, \\ p'_v(t+1) &= \left[p'_v(t) - \alpha \frac{\partial D(p'(t))}{\partial p_v} \right]^+, \end{aligned} \tag{15}$$

where $\alpha > 0$ is the gradient step-size. Since $U_v \ \forall v \in \Upsilon$ are concave functions, $D(p, p')$ is continuously differentiable and the gradients for $D(p, p')$ with respect to p and p' are defined as follows:

$$\frac{\partial D(p, p')}{\partial p_q} = c_q - \sum_{v:(F_v \cap V_c^q) \neq \varnothing} x_v(t) \Gamma_{qv}, \quad q \in Q, \tag{16}$$

$$\frac{\partial D(p, p')}{\partial p'_v} = x_{\pi_m(v)}(t) - x_v(t)v, \quad \pi_m(v) \in V_m. \tag{17}$$

Substituting in (15) we get the supply and demand equations for calculating p and p' as follows:

$$p_q(t+1) = \left[p_q(t) - \alpha \left(c_q - \sum_{v:(F_v \cap V_c^q) \neq \varnothing} x_v(t) \Gamma_{qv} \right) \right]^+, \tag{18}$$

$$p'_v(t+1) = \left[p'_v(t) - \alpha \left(x_{\pi_m(v)}(t) - x_v(t) \right) \right]^+. \tag{19}$$

We calculate the subtree aggregate price $\lambda(T_v, t+1)$ defined by (14) at time $(t+1)$ using the clique, and forwarding price values from (18) and (19) as explained in Section 3.2. Finally, the transmission rate used by gateway node v at time $(t+1)$ is calculated as follows:

$$x_v(t+1) = \left[U'_v(\lambda(T_v, t+1)) \right]_{w_v}^{W_v}. \tag{20}$$

In order to prove the convergence of the algorithm described by (15)–(20), we define the following new terms. Let $Y_v = \sum_{q \in Q} \Gamma_{qv}$ indicate the number of subflows in subtree T_v, and $\overline{Y} = \max_{v \in \Upsilon} Y_v + |\Lambda_m(v)| - 1_v$ indicate the maximum summation of subflows in T_v plus number of children gateway nodes in $\Lambda_m(v) \ \forall v \in \Upsilon$, where $1_v = 1$ if $\pi_m(v) \neq \varnothing$, and zero otherwise. Let $\overline{Z} = \max_{q \in Q} \sum_{v \in \Upsilon} \Gamma_{qv}$ be the number of subflows in the most congested clique $q \in Q$, and $\overline{\gamma} = \max_{v \in \Upsilon} \gamma_v$ indicate the upper bound on all $-U''_v(x_v) \ \forall v \in \Upsilon$. We can obtain the following convergence result, the proof of which is in Appendix B.

Theorem 2. *For step-size values of α that satisfy the inequality $0 < \alpha < 2\overline{\gamma}/\overline{YZ}$, starting from any initial rate $x(0)$ ($x_v \in I_v \ \forall v \in \Upsilon$), clique prices $p(0) \geq 0 \ \forall q \in Q$, and forwarding prices $p'(0) \geq 0 \ \forall v \in \Upsilon$, the sequence of vectors $x(t) = (x_v(t), v \in \Upsilon)$ converges to the unique optimal solution of problem (P).*

4.1. Synchronous Versus Asynchronous Computations. Equations (15)–(20) assume that the price and rate iterations are performed at time $t = 1, 2, 3, \ldots$, which implies that the price and rate calculations happen at the same time using a synchronous computation scheme. Such synchronization is however difficult to attain in a distributed network environment where nodes do not have any global synchronization clock. Practically speaking, asynchronism inevitably happens for price and rate calculations at any node because the node may not have the exact current value of the rate, the clique, or the forwarding price. Instead, it receives a sequence of recent values at different time instances. Therefore, the node will use a weighted average of these values in estimating the price or the rate at any given time. For example, for node i to calculate $p_q(t+1)$ from (18), it needs all the rates $x_v(t)\ \forall v : (F_v \cap V_c^q) \neq \varnothing$, at exactly time t. However, because node i may not have the rates at time t, it keeps the rate values at times $(t - B) \leq t' \leq t$, where B is a configurable time window for rate and price calculations. Then, it estimates the rates at time t using the following weighted average:

$$\widehat{x}_v(t) = \sum_{t'=t-B}^{t} b_i^q(t', t)\, x_v(t') \quad \text{with} \sum_{t'=t-B}^{t} b_i^q(t', t) = 1. \tag{21}$$

This asynchronous mechanism is general and allows for deploying any estimation policy for the rates or prices. The simulations in Section 5 show that our algorithm attains convergence using some popular update policies such as

(i) *latest instant estimation*: only the last received value for $x_v(\tau)$ for some $\tau \in [t - B, \ldots, t]$, is used to estimate $\widehat{x}_v(t)$, that is, $b_i^q(t', t) = 1$ if $t' = \tau$ and 0 otherwise;

(ii) *latest average estimation*: only the average over the latest k received values is used for estimation, that is, $b_i^q(t', t) > 0$ for $t' = \tau - k + 1, \ldots, \tau$ and 0 otherwise.

The details for the asynchronous algorithm for heterogeneous wireless multicast are shown in Algorithm 1. In order to understand the association of this algorithm with the network architecture, we assume that each node i in the network has zero or more multicast subflows $f_{vi}\ \forall v \in \Upsilon$ depending on the traffic passing by this node. Even though the algorithm suggests that the *clique procedure* at clique q can be performed by a designated node from that clique (i.e., clique master), in our simulations we perform the clique procedure at each node i separately for all cliques that have $f_{vi} \cap V_c^q \neq \varnothing$. The *subflow procedure* is performed by each node i that has one or more multicast subflows $f_{vi}\ \forall v \in \Upsilon$ by simply calculating the accumulated prices at the branches of f_{vi} based on the accumulated price at i. Finally, *active gateway nodes* (i.e., gateway nodes that have traffic from one or more multicast groups passing by them) perform the *subtree procedure* by calculating the optimal rate $x_v(t+1)$ based on the aggregated prices for this subtree.

The estimation of the price and rate values at time t (i.e., $\widehat{x}_v(t), \widehat{p}_q(t), \widehat{\lambda}_v(i, t)$ and $\widehat{p}_i'(t)$) from the received values at time instances in the range $(t - B) \leq t' \leq t$ may follow any policy such as the *latest instant estimation* or the *latest average estimation* as explained before. The support for these different update policies demonstrates the versatility of our asynchronous algorithm. The following theorem illustrates the convergence of this model (detailed proof is given in [14]).

Theorem 3. *For step-size values of α that are sufficiently small, starting from any initial rate $x(0)$ ($x_m \in I_m\ \forall m \in M$) and clique prices $p(0) \geq 0\ \forall q \in Q$, every accumulation point $(x^*; p^*)$ of the sequence $(x(t); p(t))$ generated by the asynchronous Algorithm 1 (ORAWHM) is primal-dual optimal.*

4.2. Time-Varying Network Environment. So far, we assume that the cliques achieved capacity and the set of group utility functions are not functions of time (i.e., they do not change with time). However, due to online calculation and subproblem decomposition, it can be shown that our solution will work in the case when these quantities change with time.

For example, the clique capacity may be time-varying depending on the scheduling discipline used at the MAC layer. In this case, (16) will be the same except the current clique capacity $c_p(t)$ is used in place of the constant capacity c_p. For deploying our algorithm in a real network, we account for the time-varying channel capacity by using a bandwidth management mechanism for estimating the channel capacity based on [15]. In general, if the change in the environment parameters is relatively slow, our solution can track the changes in the optimal rates based on changing these quantities with time. This is shown in our experimental evaluation discussed in Section 5.

4.3. Cross-Layer Architecture for ORAHWM. Figure 3 depicts the cross-layer architecture of ORAHWM showing the main procedures and the interaction of ORAHWM with the different layers including MAC, routing, and transport layers. In this architecture, we use the common IEEE 802.11 DCF as the MAC protocol with multicast extensions as presented in [11]. Multicast ad hoc on demand distance vector (MAODV) [16] is used to provide a distributed routing scheme for the multicast sessions. We also use UDP with rate control in the transport layer to minimize the communication overhead through avoiding the excessive feedback packets used by other transport protocols (e.g., TCP).

We use the channel capacity estimator to measure the channel capacity in real time in the MAC layer. For this purpose, we devise a cross-layer mechanism which combines the multicast aware MAC protocol (MMP) [11] with a bandwidth management mechanism for measuring the channel capacity based on [15]. For details about this mechanism, please refer to [17]. We also use the HELLO packets for conveying the clique price information in the routing layer. The information from channel capacity estimator and HELLO packets jointly establish the requirements to calculate the dual gradient for clique prices described by (16). To construct the contention domains, we allow the price information to be broadcasted as part of the HELLO

```
**Clique procedure** (by clique $q$): At times $t \in T_q$
(1) Receive rates $x_v(t')$ from all subtrees $T_v$ where $F_v \cap V_c^q \neq \varnothing$
(2) Update clique price as follows
        $$p_q(t+1) = \left[ p_q(t) - \alpha\left( c_q - \sum_{v:(F_v \cap V_c^q) \neq \varnothing} \hat{x}_v(t)\Gamma_{qv} \right) \right]^+, \quad \forall q \in Q$$
(3) Send $p_q(t+1)$ to all nodes of group $m$ such that $F_m \cap V_c^q \neq \varnothing$

**Subflow procedure** (by subflow $f_{vi}$): At times $t \in T_m^i$
(1) Receive prices $p_q(t')$ from maximal cliques $q$ where $f_{vi} \cap V_c^q \neq \varnothing$
(2) Calculate the subflow price (per hop price) $\lambda_{vi}$ as follows
        $$\lambda_{vi}(t+1) = \sum_{q:(f_{vi} \cap V_c^q) \neq \varnothing} \hat{p}_q(t)$$
(3) Calculate the accumulated price on each branch $b_{vij} \in f_{vi}$
        $$\lambda_v(j, t+1) = \frac{\lambda_v(i,t) + \lambda_{vi}(t+1)}{K_{vi}}$$
(4) Forward $\lambda_v(j, t+1)$ to all children subflows of $f_{vi}$, if no children, send $\lambda_v(j, t+1)$ to $v$.

**Subtree procedure** (by gateway $v$): At times $t \in T_v$
(1) Receive the net prices $\lambda_v(i,t') - p_i'(t')$ from all terminal nodes of $T_v$ (i.e., $\forall i \in \mathfrak{I}_v$),
      and all children gateway nodes $\forall i \in \Lambda_m(v)$   // (note: $p_i' = 0 \ \forall i \in \mathfrak{I}_v$).
(2) If $\pi_m(v) \neq \varnothing$   // (i.e., $v \neq s_m$)
    (i) Receive rate $x_{\pi_m(v)}(t'')$ from parent subtree of $T_v$
    (ii) Calculate the next forwarding price $p_v'(t+1)$ as follows:
        $$p_v'(t+1) = \left[ p_v'(t) - \alpha\left( \hat{x}_{\pi_m(v)}(t) - x_v(t) \right) \right]^+$$
    Else $p_v'(t+1) = 0$
(3) Calculate the next subtree aggregate price $\lambda(T_v, t+1)$ as follows:
        $$\lambda(T_v, t+1) = p_v'(t+1) + \sum_{i \in (\mathfrak{I}_m \cup \Lambda_m(v))} \left( \hat{\lambda}_v(i,t) - \hat{p}_i'(t) \right)$$
(4) Calculate the next subtree rate as follows:
        $$x_v(t+1) = U_v'^{-1}(\lambda(T_v, t+1))$$
(5) Send $x_v(t+1)$ to all cliques $q$ where $F_v \cap V_c^q \neq \varnothing$
```

Algorithm 1: ORAHWM: asynchronous distributed algorithm.

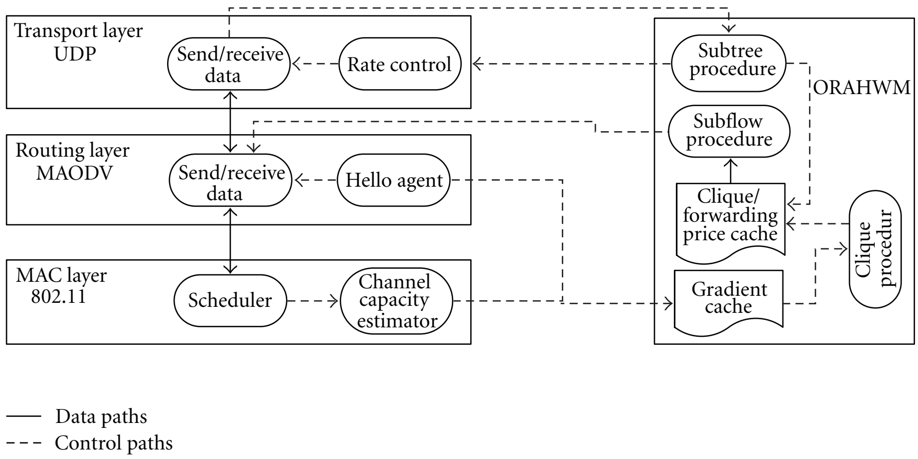


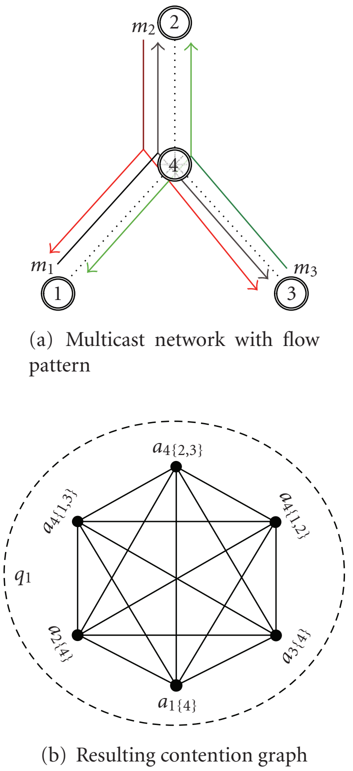
Figure 3: Cross-layer architecture for ORAHWM.

packets to all neighboring nodes within 3 hops away. Such multihop protocol-based scheme in calculating the maximal cliques is proven to work when the interference range is greater than or equal to the transmission range, given that nodes within the same interference range are reachable to each other through multihop communication, as shown in [6]. Feedback packets from the terminal nodes and gateway nodes can be used to convey the rate and accumulated price information which are used by the *subtree procedure* to calculate the dual gradient for forwarding prices described by (17) and the aggregated subtree price in (14) and hence the next subtree rate described by (20).

5. Simulation Results

In all our experiments, we use the utility functions $U_v(x_v) = g_v \ln(x_v)\ x_v > 0$ for imposing proportional fairness [4] amongst the multicast groups, where g_v is the differentiation gain for gateway node v, that is, $x_v(t) = g_v/\lambda_v(t)$. The default transmission and interference ranges are 250 m and 550 m, respectively. We implemented all the simulations using nanosecond-2 simulator. Unless otherwise stated, we use the *latest instant estimation* for asynchronous calculations.

5.1. Effect of Time-Varying Wireless Channel. In this experiment, we study the effect of time-varying wireless channel on the speed of convergence for our algorithm ORAHWM. we deployed our algorithm in a real network that uses multicast aware IEEE 802.11 DCF MAC scheduler [11] with a bandwidth management mechanism for measuring the channel capacity (i.e., channel capacity estimator in Figure 3) based on [15] and (MAODV) [16] for routing. We take as an example the network in Figure 4 with 3 multicast sessions as shown in Figure 4(a) and the corresponding contention graph as shown in Figure 4(b) . We use equal differentiation gains, that is, $g_v = 1\ \forall v \in M$. However, we start each of these sessions in a different time to test the ability of our algorithm to track network changes. The start times of sessions m_1, m_2, m_3 are 20, 40, 60 seconds, respectively, and the initial rates $x_v(0)\ \forall v \in M$ are selected from a uniform distribution in the range 50–250 kbps. We have fixed all the other parameters including the step-size, and we measured the rate of each multicast session against time. Figure 5 shows the result using a time-varying channel capacity realized by the MAC scheduler. From this figure, we observe that although the MAC channel capacity (i.e., the basic rate of sending data in IEEE 802.11 DCF) is set to 1 Mbps, the achieved channel capacity changes with time and does not go above 800 Kbps. Nevertheless, our algorithm continuously tracks the change in channel capacity and provides proportional fairness amongst all the multicast sessions based on the current available channel capacity. We also notice that the algorithm spends less than 2 seconds to achieve the optimal rates every time a new multicast session is added, which is deemed reasonable. However, intuitively, this convergence speed is affected by the number of hops that each multicast session spans and hence it may decrease in larger networks as we will see in the following experiments.



(a) Multicast network with flow pattern

(b) Resulting contention graph

Figure 4: Effect of time-varying wireless channel.

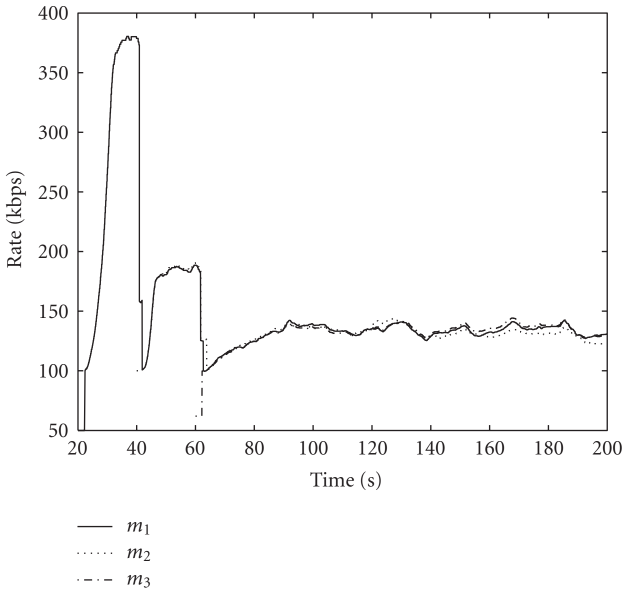


Figure 5: Convergence for time-varying channel capacity using IEEE 802.11 DCF scheduling.

5.2. Convergence in Random Network for Unirate Multicast. In this experiment, we study the convergence behavior of our algorithm ORAHWM with respect to both calculated rate and throughput in a randomly generated wireless network as shown in Figure 6. This network consists of 30 nodes deployed randomly over the 1000 × 1000 m^2 wireless space.

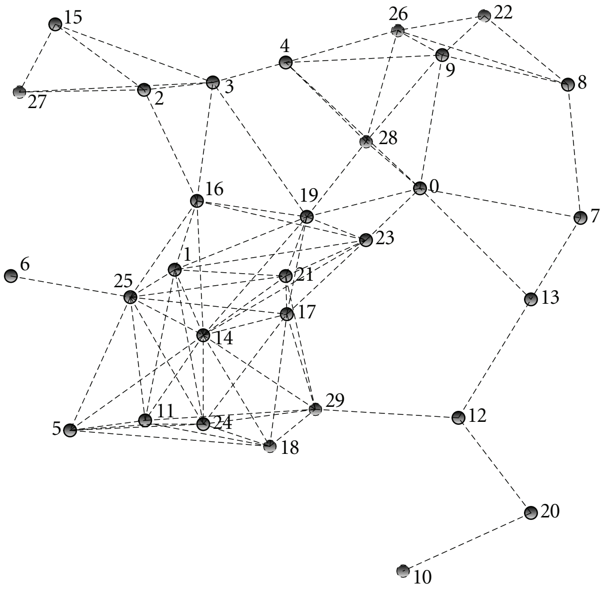
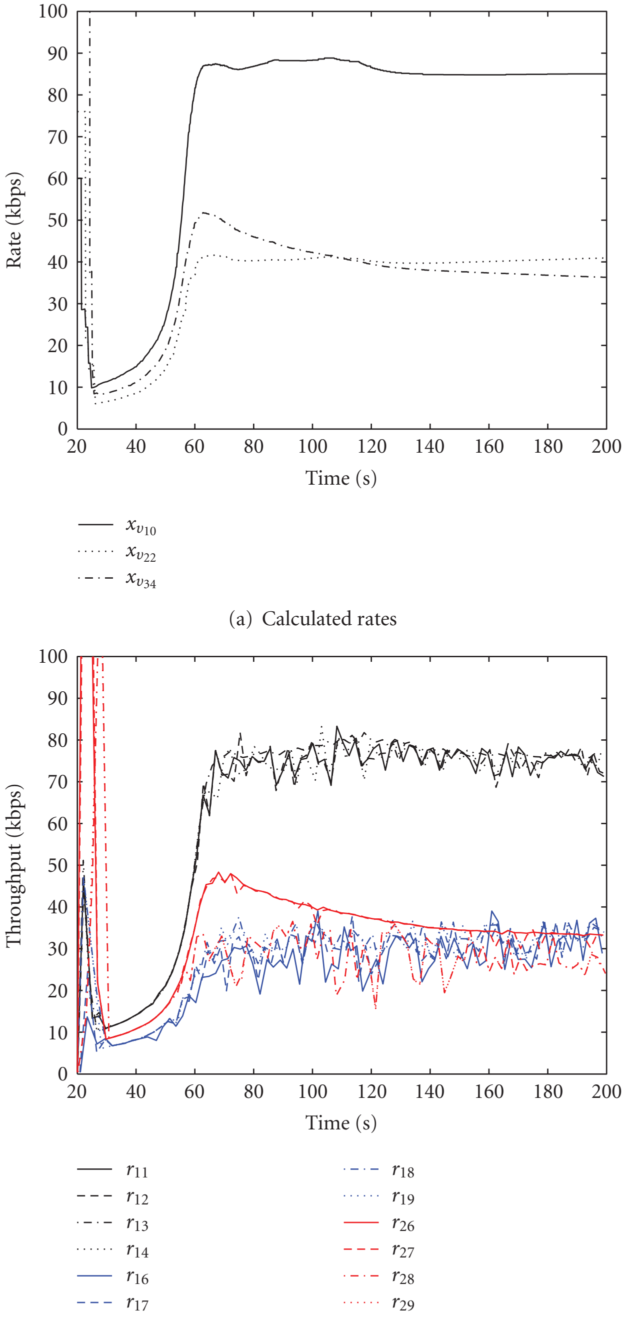


Figure 6: Random wireless network with 30 nodes.

Table 2: Multicast traffic pattern.

Session	Source/gateway	Receivers
m_1	v_{10}	$r_{11}, r_{12}, r_{13}, r_{14}$
m_2	v_{22}	$r_{16}, r_{17}, r_{18}, r_{19}$
m_3	v_{34}	$r_{26}, r_{27}, r_{28}, r_{29}$



(a) Calculated rates

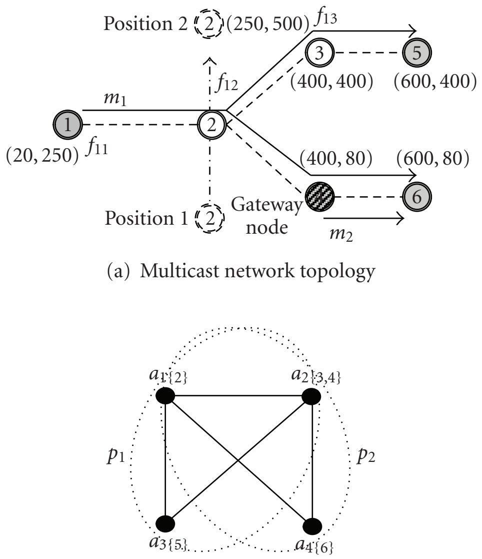
(b) Receiver throughput

Figure 7: Convergence of ORAHWM in large random networks.

We started 3 multicast sessions m_1, m_2, and m_3 at time 20 seconds, each with one source and gateway node v_{mi} and four receivers as shown in Table 2 using $\alpha = 10^{-8}$. The differentiation gain for all the three sessions is $g_v = 1 \ \forall v \in M$).

Figure 7 shows the calculated rates and receiver throughput of each multicast session with time. From these results, we observe that our algorithm attains convergence with satisfactory speed even in relatively large-scale networks. We also observe that the throughput achieved by each receiver on all sessions follows the calculated rates fairly well, which confirms the correctness of the calculated rates. Note that the optimal calculated rates are different for each session depending on the size of the multicast tree and how much resources each session consumes from the total network resources. If this discrimination based on tree topology is undesirable, it can be compensated using different differentiation gains (g_v) on each session, which will be discussed in the next experiment. We also observe that the time spent by the algorithm to achieve full optimality is almost 35 seconds in such large fully distributed network. However, after 5 seconds only, the rates start to approach optimal point gradually. This indicates that, although the algorithm may not have enough time to achieve full optimality especially in large-distributed environments, it will always attempt to approach optimal point and follow network changes concurrently and satisfactorily.

5.3. Effect of Changing Differentiation Gains on the Calculated Rates and Aggregate Utility. In these experiments, we study the effect of changing the differentiation gains on the calculated rates for unirate and multirate multicast sessions. We consider the small topology shown in Figure 8. Two sessions m_1 and m_2 are sharing this network with source and receiver nodes as shown in Figure 8. we consider 3 cases where we change the differentiation gain and show the effect on the calculated rates in each case. Case 1 is the unirate multicast where we use one gateway/source node for each



(a) Multicast network topology

(b) Contention graph

Figure 8: Effect of changing differentiation gains on the calculated rates and aggregate utility.

multicast group, and we use equal differentiation gains for both sessions (i.e., $g_{v_{11}} = g_{v_{24}} = 5$). For both cases 2 and 3, m_1 gateway node 4 for rate control and the differentiation gain $g_{v_{24}}$ is set to. Case 2 uses differentiation gains $g_{v_{11}} = 3$, $g_{v_{14}} = 2$ whereas case 3 uses $g_{v_{11}} = 4$, $g_{v_{14}} = 1$. In all cases we start both multicast groups at $t = 20$ seconds, we fix all the other parameters including $\alpha = 3 \times 10^{-7}$, and we set the channel capacity for all maximal cliques to 1 Mbps.

Figure 9 shows the calculated rates and the aggregated utility for the 3 cases. We notice that for case 1 (unirate), as expected, our algorithm ORAHWM will discriminate against session m_1 because it uses more wireless links and hence utilizes more network resources. This happens because for unirate, ORAHWM deals with each session as one entity regardless of how large this session is and how many links it uses. Multirate with additional gateway nodes can reduce this effect by providing more flexibility to assign more priority to some parts of the tree which in turn affects the aggregate utility of the entire system. This is depicted by the results in Figures 9(b) and 9(c). We notice that by increasing the differentiation gain for $T_{v_{11}}$, we can increase the aggregate utility (shown in Figure 9(d)). Therefore, assigning differentiation gains to different parts of the multicast trees is a crucial aspect of this algorithm and may call for a mechanism to assign these differentiation gains in real time in order to maximize the overall aggregate utility. For example, multicast subtrees which serve large number of uncongested receivers will be assigned higher differentiation gains, while multicast subtrees with fewer congested receivers will be assigned lower differentiation gains.

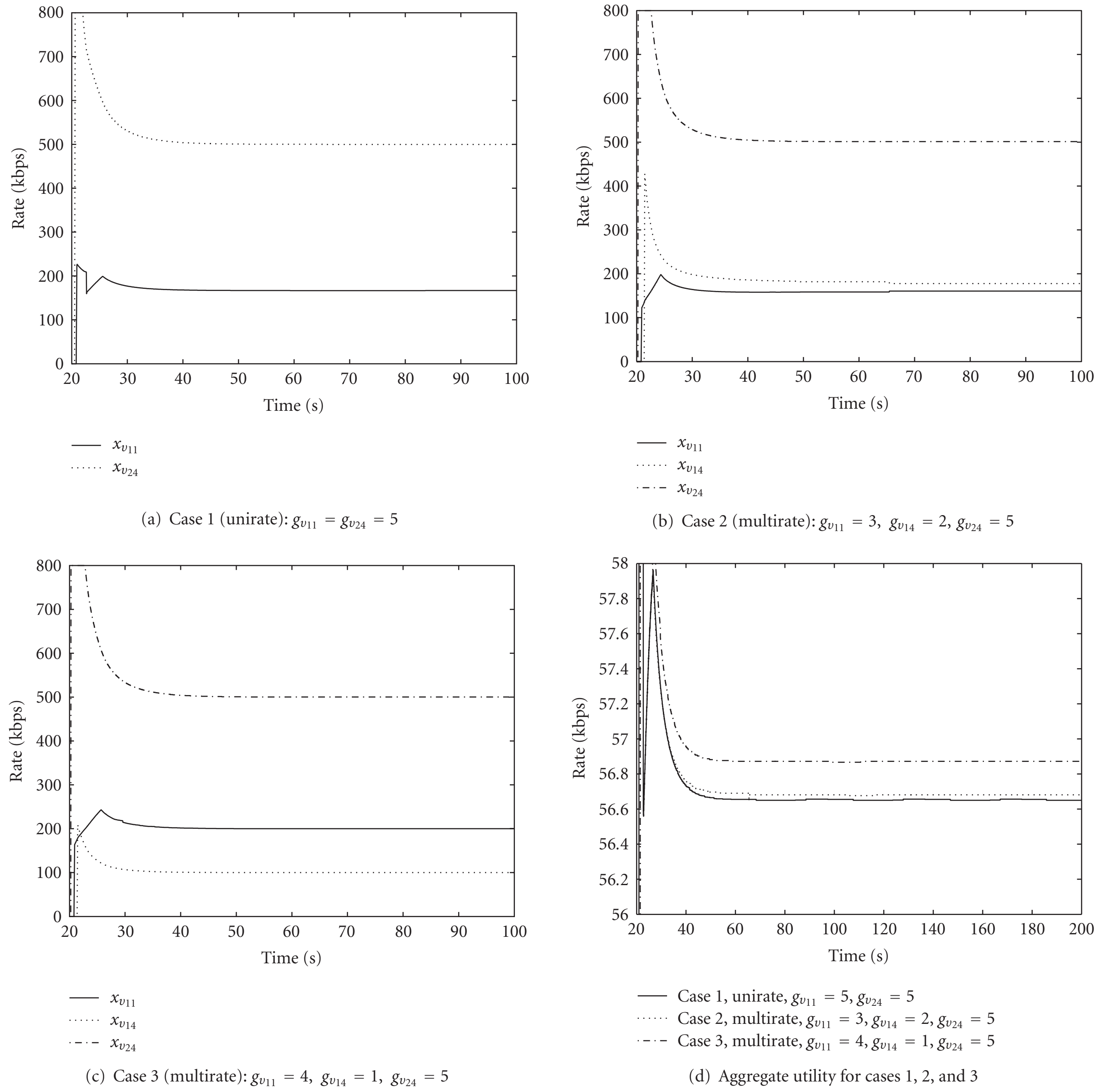
5.4. Effect of Time-Varying Channel and Mobility on the Convergence of ORAHWM. In these experiments, we study the effect of changing network conditions including changing capacity and node mobility on the convergence of our algorithm ORAHWM. We consider the same topology and multicast sessions shown in Figure 8, and we use the $g_{v_{11}} = 4$, $g_{v_{14}} = 1$.

First we study the effect of measuring the real capacity on each clique using the MAC layer channel capacity estimator as explained in Section 4.3. Figure 10 shows the result of using a time-varying channel capacity realized by the MAC scheduler IEEE 802.11 DCF with channel data rate 1 Mbps. From this figure, we observe that although the achieved channel capacity changes with time, our algorithm continuously tracks the change in channel capacity fairly well and provides proportional fairness amongst all the multicast sessions based on the current available channel capacity.

We also study the impact of mobility and route changes on the convergence of our algorithm by generating a mobility pattern where node 2 moves from *position* 1 to *position* 2 as shown in Figure 8(a) with average speed of 3 m/s and pause time 20 seconds. Figure 11 shows the rates calculated by our algorithm with time. The figure shows 3 different regions depending on the change of routes resulting from the node mobility. In region 1, only node 6 is receiving traffic for both multicast sessions. As expected in this case, our algorithm converges to the same rates of case 3 in the previous experiment. As node 2 moves to region 2, the routes for which are shown by Figure 8, both receivers at nodes 5, and 6 become active for session m_1 and the optimal rates converge to the same values, after some transient period, despite the route changes. When node 2 moves to region 3, both the receivers at node 6 and node 4 become inactive for session m_1 and session m_2 can now use the whole channel for its traffic. Therefore, the optimal rate for m_2 in this case is 1 Mbps whereas the capacity is divided amongst the 3 subflows f_{11}, f_{12}, and f_{13} for m_1.

5.5. Effect of Using Multirate on the Total Throughput for Multicast Flows. In this experiment we study the effect of using gateway nodes for rate control as part of a multicast group. Consider Figure 12 which shows two multicast groups m_1 and m_2 sharing an ad hoc network on 11 nodes as shown in Figure 12. m_1 uses gateway/source node 1 (i.e., v_{11}), and has 3 receivers, namely, r_7, r_8, and r_9 while m_2 uses gateway/source node 6 (i.e., v_{26}) and has two receivers, namely, r_{10} and r_{11}. Here, to study the impact of using multirate multicast we consider two cases. Case 1 is the unirate multicast with equal differentiation gains for both multicast groups (i.e., $g_{v_{11}} = g_{v_{26}} = 3$). For case 2, m_1 uses an additional gateway node at 4 (i.e., v_{14}) for rate control. In this case, we set $g_{v_{11}} = 2$, $g_{v_{14}} = 1$ so the total differentiation gain is similar to case 1, and we set $g_{v_{26}} = 3$.

Figures 13 and 14 show the calculated rates and receiver throughput for cases 1 and 2, respectively. We notice that in each case convergence is attained, and the throughput achieved by all receivers on each group tracks the calculated rates appropriately. Comparing the two figures, we notice the effect of using gateway node v_{14} for m_1 which lowers the optimal rate on the subtree $T_{v_{14}}$ (i.e., $x_{v_{14}}$) allowing the



(a) Case 1 (unirate): $g_{v_{11}} = g_{v_{24}} = 5$

(b) Case 2 (multirate): $g_{v_{11}} = 3,\ g_{v_{14}} = 2,\ g_{v_{24}} = 5$

(c) Case 3 (multirate): $g_{v_{11}} = 4,\ g_{v_{14}} = 1,\ g_{v_{24}} = 5$

(d) Aggregate utility for cases 1, 2, and 3

FIGURE 9: Effect of changing differentiation gains on the calculated rates and aggregate utility.

other rates $x_{v_{11}}$ and $x_{v_{26}}$ to increase drastically. This happens because we set the differentiation gain $g_{v_{14}} = 1$ giving this subtree lower priority based on our knowledge that this subtree has only one receiver (r_9), and the surrounding area has traffic load more than for example, $T_{v_{11}}$ and we used v_{14} to give us the flexibility of setting $x_{v_{14}}$ accordingly. Such knowledge can either be communicated between the receivers and gateway nodes or tuned manually by an administrator.

To study the effect of this heterogeneity within m_1 we measure the aggregate utility and the total throughput achieved by each group for cases 1 and 2. Figures 15 and 16 show the results for these measurements. We see from Figure 15 that the aggregate utility achieved for case 2 is better as a result of using gateway node v_{14} because both rates $x_{v_{11}}$ and $x_{v_{26}}$ increased significantly by reducing $x_{v_{14}}$. This increase in rates caused the overall throughput achieved by both multicast groups to increase drastically (i.e., ≈30%) as shown in Figure 16.

6. Concluding Remarks

In this paper we have presented the resource optimization algorithm for the case of multirate multicast (ORAHWM)

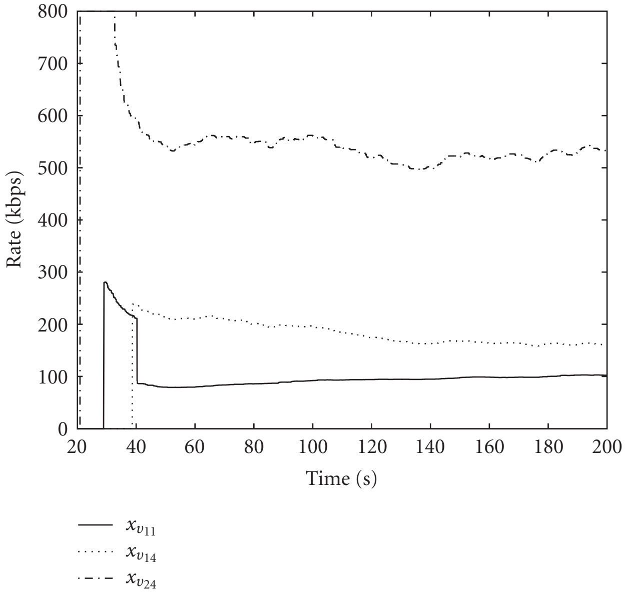


FIGURE 10: Calculated rates with dynamic capacity: $g_{v_{11}} = 4$, $g_{v_{14}} = 1$, $g_{v_{24}} = 5$.

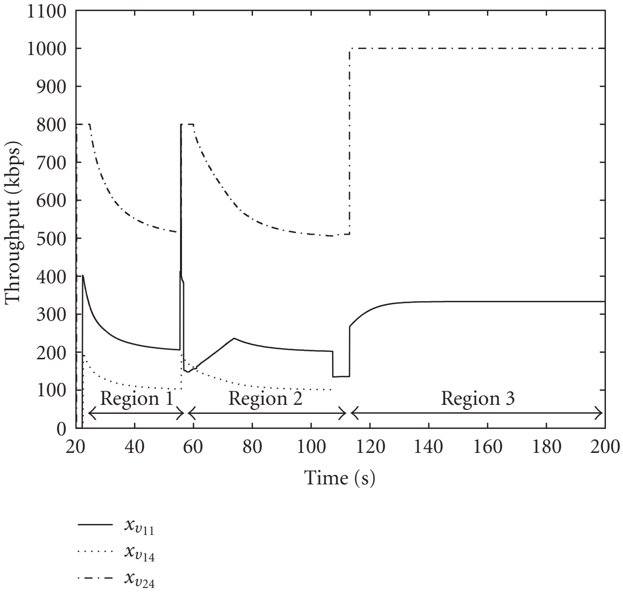


FIGURE 11: Calculated rates with mobility: $g_{v_{11}} = 4$, $g_{v_{14}} = 1$, $g_{v_{24}} = 5$.

over multihop ad hoc networks. We have introduced the notion of gateway nodes used to control the rates for multirate multicast groups and provided the optimization model that realizes the optimal rates used by each gateway node in order to maximize the overall aggregate utility for the entire system. We also discussed the cross-layer architecture that can be used for deploying this algorithm in real networks. We proposed a mechanism for calculating the subtree price based on the branch accumulated price which allows the calculation to occur in a totally distributed and

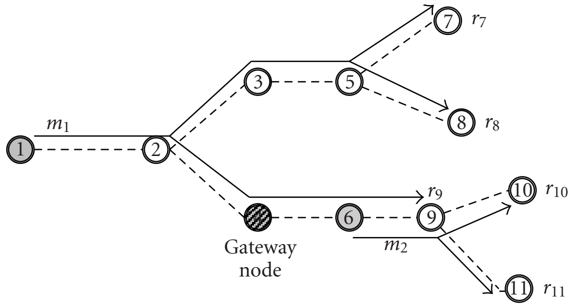
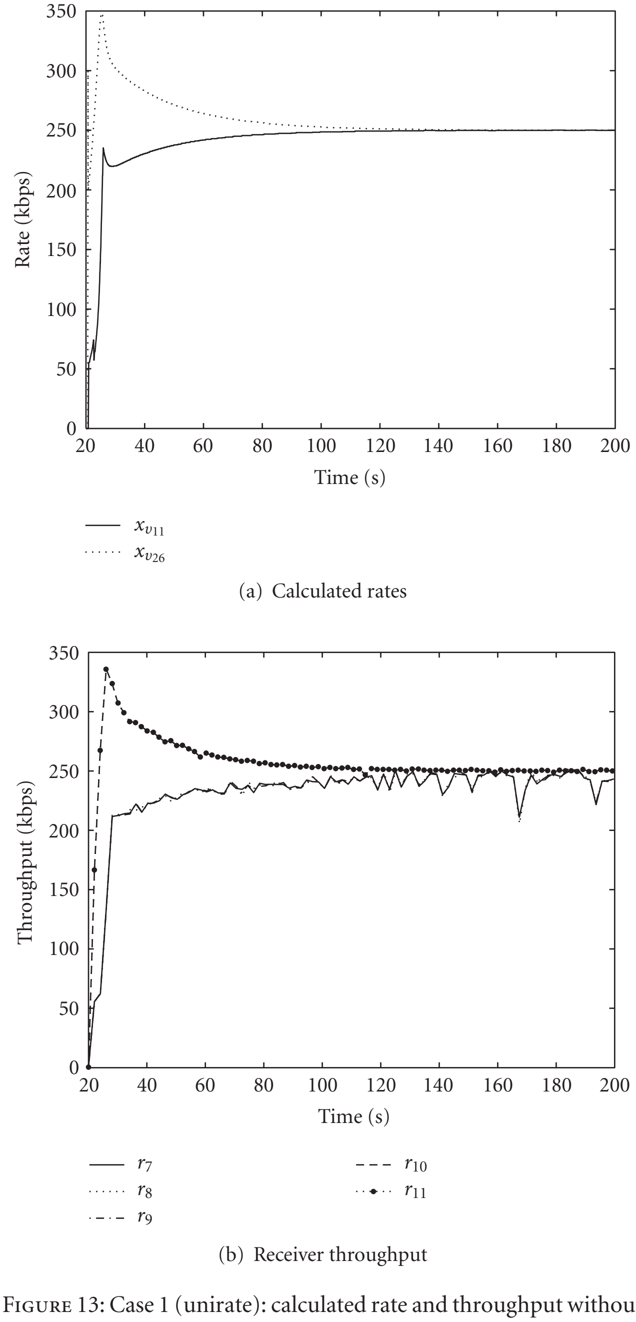


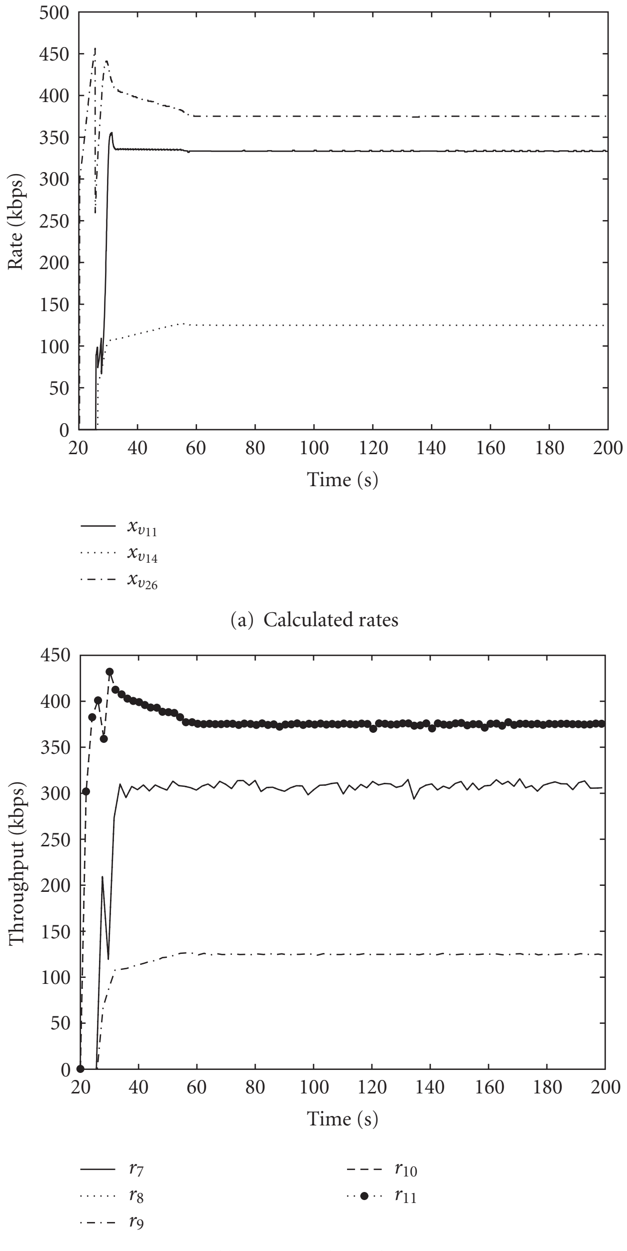
FIGURE 12: Multirate multicast network topology.



(a) Calculated rates

(b) Receiver throughput

FIGURE 13: Case 1 (unirate): calculated rate and throughput without using gateway node v_{14}.



(a) Calculated rates

(b) Receiver throughput

Figure 14: Case 2 (multirate): calculated rate and throughput using gateway node v_{14} for rate control on m_1.

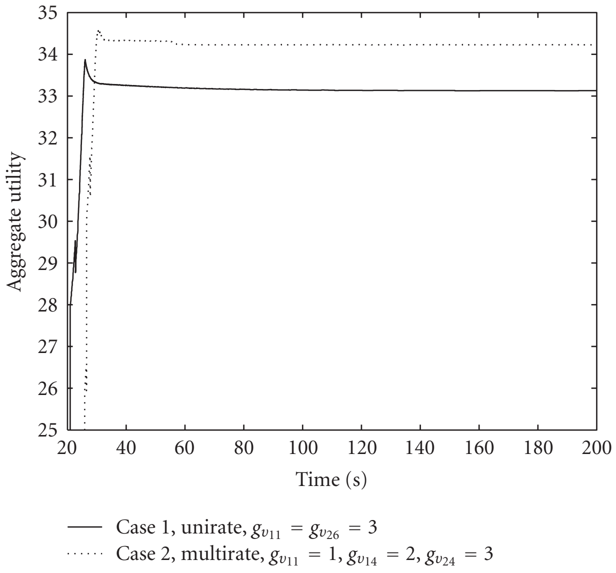


Figure 15: Aggregate utilities for cases 1 (unirate) and 2 (multirate).

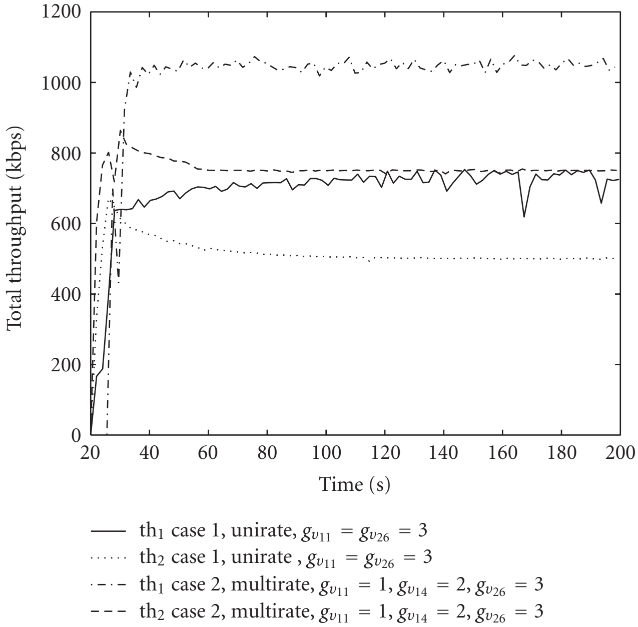


Figure 16: Total throughput for each multicast group for cases 1 and 2: th_1 is total throughput for m_1, th_2 is total throughput for m_2.

asynchronous way. Utilizing the flexibility of using gateway nodes across the multicast trees, ORAHWM is expected to increase the aggregate utility of the system and boost the overall throughput achieved by each multicast group by as high as 30% provided that the differentiation gains are set appropriately.

Appendices

A. Proof of Theorem 1

Proof. Assume that $\Im_v(h)$ is the set of all nodes $i \in T_v$ such that the depth of i is h, and H is the maximum depth of the subtree T_v. Now it is easy to recognize that

$$\lambda_v = \lambda_{vv} + \sum_{i\in\Im_v(1)} \lambda_{vi} + \cdots + \sum_{i\in\Im_v(H-1)} \lambda_{vi}, \tag{A.1}$$

where λ_{vv} is the clique price for subflow $f_{vv} \in F_v$ branching from the gateway node v. Next, we proceed by induction based on H as follows.

(i) For $H = 1$,

$$\sum_{i\in\Im_v} \lambda_v(i) = \sum_{i\in\Im_v(1)} \lambda_v(i) = \frac{K_{vv} \times \lambda_{vv}}{K_{vv}} = \lambda_v. \tag{A.2}$$

(ii) For $H = 2$,

$$\begin{aligned}\sum_{i\in\mathfrak{I}_v(2)} \lambda_v(i) &= \sum_{i\in\mathfrak{I}_v(2)} \frac{\lambda_{mv}/K_{mv} + \lambda_{m\pi_v(i)}}{K_{m\pi_v(i)}} \\ &= \sum_{i\in\mathfrak{I}_v(1)} \frac{\lambda_{vv}/K_{vv} + \lambda_{vi}}{K_{vi}} \times K_{vi} \\ &= \lambda_{vv} + \sum_{i\in\mathfrak{I}_v(1)} \lambda_{vi} = \lambda_v.\end{aligned} \tag{A.3}$$

Notice that if $f_{vi} \notin F_v$, then $\lambda_{vi} = 0$.

(iii) Assume that for $H = n - 1$,

$$\sum_{i\in\mathfrak{I}_v(n-1)} \lambda_v(i) = \lambda_v = \lambda_{vv} + \cdots + \sum_{i\in\mathfrak{I}_v(n-2)} \lambda_{vi}, \tag{A.4}$$

(iv) Hence, for $H = n$,

$$\begin{aligned}\sum_{i\in\mathfrak{I}_v(n)} \lambda_v(i) &= \sum_{i\in\mathfrak{I}_v(n-1)} \frac{\lambda_v(i) + \lambda_{vi}}{K_{vi}} \times K_{vi} \\ &= \lambda_{vv} + \cdots + \sum_{i\in\mathfrak{I}_v(n-1)} \lambda_{vi} = \lambda_v,\end{aligned} \tag{A.5}$$

therefore the result follows. □

B. Proof of Theorem 2

The proof follows the same way as [4, Theorem 1]. We define $\widetilde{\Lambda}$ to be the set of gateway nodes that have $\Lambda_m(v) \neq \varnothing$. Then the vector of forwarding prices p' is defined as $p' = (p'_v, \forall v \in \widetilde{\Lambda})$. First we prove the following lemma.

Lemma 1. *If $\hat{u}, \overline{u} = (p, p')$ are any two $(|Q| + |\widetilde{\Lambda}|) \times 1$ feasible system price vectors, that is, $\hat{u}, \overline{u} \geq 0$, then based on Assumptions 1 and 2, ∇D satisfies the Lipscitch condition*

$$\|\nabla D(\hat{u}) - \nabla D(\overline{u})\|_2 \leq \frac{\overline{YZ}}{\overline{\gamma}} \|\hat{u} - \overline{u}\|_2. \tag{B.1}$$

Proof. From (15), we have $\nabla D = \hat{C} - \hat{\Gamma} x$, where $\hat{C}$ is the $(|Q| + |\widetilde{\Lambda}|) \times 1$ capacity vector with $\hat{c}_i = 0 \ \forall i \in \widetilde{\Lambda}$, and $\hat{\Gamma}$ is the $(|Q| + |\widetilde{\Lambda}|) \times |\Upsilon|$ constraints matrix.

Let $(\partial x/\partial u)(u)$ denote the $|\Upsilon| \times (|Q| + |\widetilde{\Lambda}|)$ matrix whose (i, j) element $(\partial x_i/\partial u_j)(u)$ is

$$\frac{\partial x_i}{\partial u_j}(u) = \begin{cases} \dfrac{\hat{\Gamma}_{ji}}{U_i''(x_i(u))}, & \text{if } U_i'(W_i) \leq u_j \leq U_i'(w_i), \\ 0, & \text{o.w.} \end{cases} \tag{B.2}$$

If we define $\beta_i(u)$ as follows:

$$\beta_i(u) = \begin{cases} -\dfrac{1}{U_i''(x_i(u))}, & \text{if } U_i'(W_i) \leq u_j \leq U_i'(w_i), \\ 0, & \text{o.w,} \end{cases} \tag{B.3}$$

then $(\partial x_i/\partial u_j)(u)$ in matrix form can be written as

$$\left[\frac{\partial x_i}{\partial u_j}(u)\right] = -B(u)\,\hat{\Gamma}^T, \tag{B.4}$$

where $B(u) = \mathrm{Diag}(\beta_i(u);\ i \in \Upsilon)$ is the diagonal matrix with diagonal elements $\beta_i(u)$. Hence,

$$\nabla^2 D = -\hat{\Gamma}\left[\frac{\partial x_i}{\partial u_j}(u)\right] = \hat{\Gamma}\, B(u)\, \hat{\Gamma}^T. \tag{B.5}$$

Now from [18, Proposition A.25(e)] and knowing that $\nabla^2 D = \hat{\Gamma}\, B(p)\, \hat{\Gamma}^T$ is symmetric (i.e., $\|\hat{\Gamma}\, B(p)\, \hat{\Gamma}^T\|_1 = \|\hat{\Gamma}\, B(p)\, \hat{\Gamma}^T\|_\infty$), then we have

$$\begin{aligned}\|\hat{\Gamma}\, B(u)\, \hat{\Gamma}^T\|_2 &\leq \|\hat{\Gamma}\, B(u)\, \hat{\Gamma}^T\|_\infty \\ &= \max_j \sum_{j'} [\hat{\Gamma} B(u) \hat{\Gamma}^T]_{jj'} \\ &= \max_j \sum_{j'} \sum_i \beta_i(u) \hat{\Gamma}_{ji} \hat{\Gamma}_{j'i} \\ &= \max_j \sum_i \beta_i(u) \hat{\Gamma}_{ji} \sum_{j'} \hat{\Gamma}_{j'i},\end{aligned} \tag{B.6}$$

where $\sum_{j'} \hat{\Gamma}_{j'i}$ represents the sum of subflows in each maximal clique $j' \ \forall j' : (F_i \cap V_c^{j'} \neq \varnothing)$ plus the number of children gateway nodes for each subtree T_i, which is by definition $\leq \overline{Y}$. Then we have

$$\|\hat{\Gamma}\, B(u)\, \hat{\Gamma}^T\|_2 \leq \frac{\overline{YZ}}{\overline{\gamma}}. \tag{B.7}$$

From [19, Theorem 9.19] we have for (B.7)

$$\|\nabla D(\hat{u}) - \nabla D(\overline{u})\|_2 \leq \frac{\overline{YZ}}{\overline{\gamma}} \|\hat{u} - \overline{u}\|_2, \tag{B.8}$$

hence the result follows. □

Proof. (Theorem 2) from Lemma 1, the dual objective function D is lower bounded and ∇D is Lipschitz. Then, limit point u^* of the sequence $\{u(t)\}$ generated by the gradient projection algorithm for the dual problem is dual optimal provided that $0 < \alpha < 2\overline{\gamma}/\overline{YZ}$ (see [18, Proposition 3.4]).

Let $\{u(t)\}$ be a subsequence converging to u^*. Since $U_i'(x_i)$ is defined on a compact interval $[w_i, W_i]$, it is continuous and one-to-one (because of the strict concavity of $U_i(x_i)$). Thus, its inverse is continuous (see [19, Theorem 4.17]) and hence from (20), $x(u)$ is continuous. Therefore, $\lim_{t\to\infty} x(t) = x(u^*)$ and that proves the result of Theorem 2. □

Acknowledgment

This research was supported by a collaborative R&D grant from the Qatar National Research Fund Project no. NPRP 26-6-7-10.

References

[1] X. Li, S. Paul, and M. Ammar, “Layered video multicast with retransmissions (LVMR): evaluation of hierarchical rate control,” in *Proceedings of the 17th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM ’98)*, vol. 3, pp. 1062–1072, San Francisco, Calif, USA, March-April 1998.

[2] Y. Cui, Y. Xue, and K. Nahrstedt, "Optimal resource allocation in overlay multicast," in *Proceedings of the 11th International Conference on Network Protocols (ICNP '03)*, pp. 71–81, Atlanta, Ga, USA, November 2003.

[3] E. Amir, S. McCanne, and R. Katz, "An active service framework and its application to real-time multimedia transcoding," *Computer Communication Review*, vol. 28, no. 4, pp. 178–189, 1998.

[4] S. H. Low and D. E. Lapsley, "Optimization flow control—part I: basic algorithm and convergence," *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, pp. 861–874, 1999.

[5] R. J. La and V. Anantharam, "Utility-based rate control in the internet for elastic traffic," *IEEE/ACM Transactions on Networking*, vol. 10, no. 2, pp. 272–286, 2002.

[6] Y. Xue, B. Li, and K. Nahrstedt, "Optimal resource allocation in wireless ad hoc networks: a price-based approach," Tech. Rep. UIUCDCS-R-2004-2505, University of Illinois at Urbana-Champaign, Urbana, Ill, USA, June 2004.

[7] K. Kar, S. Sarkar, and L. Tassiulas, "Optimization based rate control for multirate multicast sessions," in *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '01)*, vol. 1, pp. 123–132, Anchorage, Alaska, USA, April 2001.

[8] K. Kar, S. Sarkar, and L. Tassiulas, "A scalable low-overhead rate control algorithm for multirate multicast sessions," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 8, pp. 1541–1557, 2002.

[9] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Springer, Berlin, Germany, 1985.

[10] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, 2000.

[11] H. Gossain, N. Nandiraju, K. Anand, and D. P. Agrawal, "Supporting MAC layer multicast in IEEE 802.11 based MANETs: issues and solutions," in *Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks (LCN '04)*, pp. 172–179, Tampa, Fla, USA, November 2004.

[12] H. Luo, S. Lu, and V. Bharghavan, "A new model for packet scheduling in multihop wireless networks," in *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking (MOBICOM '00)*, pp. 76–86, Boston, Mass, USA, August 2000.

[13] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, Mass, USA, 1999.

[14] A. Mohamed and H. Alnuweiri, "Optimal resource allocation for homogeneous wireless multicast," Tech. Rep., University of British Columbia, Vancouver, Canada, 2006, http://www.ece.ubc.ca/~amrm/tr/orawm2006_TR.pdf.

[15] S. H. Shah, K. Chen, and K. Nahrstedt, "Dynamic bandwidth management in single-hop ad hoc wireless networks," *Mobile Networks and Applications*, vol. 10, no. 1, pp. 199–217, 2005.

[16] E. M. Royer and C. E. Perkins, "Multicast operation of the ad-hoc on-demand distance vector routing protocol," in *Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM '99)*, pp. 207–218, Seattle, Wash, USA, August 1999.

[17] A. Mohamed and H. Alnuweiri, "Cross-layer optimization framework for rate allocation in wireless multicast," in *Proceedings of IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS '06)*, pp. 1–10, Vancouver, Canada, October 2006.

[18] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[19] W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, New York, NY, USA, 1976.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 379059, 13 pages
doi:10.1155/2009/379059

Research Article

A Novel Approach to Fair Routing in Wireless Mesh Networks

Juho Määttä and Timo Bräysy

Centre for Wireless Communications, University of Oulu, P.O. Box 4500, 90014 Oulu, Finland

Correspondence should be addressed to Juho Määttä, jumaatta@cc.hut.fi

Received 2 June 2008; Accepted 8 December 2008

Recommended by Mohamed Hossam Ahmed

Multiradio wireless mesh network (WMN) is a feasible choice for several applications, as routers with multiple network interface cards have become cheaper. Routing in any network has a great impact on the overall network performance, thus a routing protocol or algorithm for WMN should be carefully designed taking into account the specific characteristics of the network. In addition, in wireless networks, serious unfairness can occur between users if the issue is not addressed in the network protocols or algorithms. In this paper, we are proposing a novel centralized routing algorithm, called Subscriber Aware Fair Routing in WMN (SAFARI), for multiradio WMN that assures fairness, leads to a feasible scheduling, and does not collapse the aggregate network throughput with a strict fairness criterion. We show that our protocol is feasible and practical, and exhaustive simulations show that the performance is improved compared to traditional routing algorithms.

1. Introduction

Wireless mesh network (WMN) [1] has recently appeared as a promising technology, which can increase coverage area and capacity of existing wireless networks. With the help of of-the-shelf wireless mesh routers, large, previously possibly unreachable, areas can have wireless access to, for example, the Internet. As these routers are becoming less expensive, the introduction of multiple radios to each router is becoming economically possible. multiradio concept with multiple noninterfering channels can significantly improve the overall network capacity, thus current WMN research has been concentrated to multiradio WMN.

In wireless networks, users or subscribers can experience unfairness depending on their location in the network. Users with multiple hops to destination are given less bandwidth than those with fewer hops. The unfairness stems from the shared wireless medium and unfair network protocols that are designed to maximize network capacity, that is, the aggregate throughput or do not take into account the fairness at all. Maximizing capacity and ensuring fairness are contradictory requirements and usually maximizing capacity has been preferred [2]. Unfairness is also present in multihop multichannel WMN. Users with multiple hops can be completely starved, while capacity, in terms of throughput, is maximized. This is naturally not fair, especially if the users pay the same amount for the service.

Usually routing in WMN has been seen from the point of view of the mesh routers (e.g., in [3]). As they are, mesh routers do not generate traffic, they only forward traffic of users and other routers. Thus, routing should be seen from the point of view of the users, who are also the paying customers. In addition, subscribers can be unevenly distributed in the network; the number of subscribers registered to a mesh router can vary significantly. This is neglected in most of capacity and routing studies, where one user per router is assumed (e.g., in [4]). Therefore, as the number of subscribers per router increase, so should its share to the limited network capacity. As discussed above, there is a need for a new or improved routing protocol or algorithm, which takes into account the special characteristics and applications of WMN as well as the distinct needs of users.

The rest of this paper is organized as follows. In Section 2 some related studies are discussed briefly. Section 3 presents needed concepts and definitions. Section 4 presents the SAFARI algorithm and shows simulation results. Section 5 interprets the simulation results and draws conclusions.

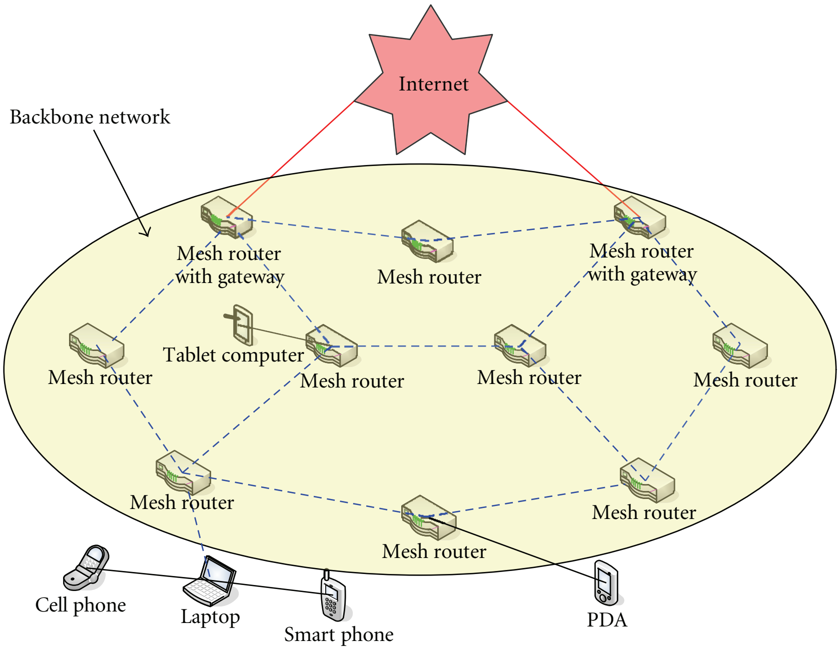


Figure 1: Hybrid wireless mesh network architecture.

2. Related Work

Fairness in medium access control (MAC), scheduling and network layer has been studied to some extent (e.g., [5–7]). These papers observe fairness in the different layers of the protocol stack and propose their solutions. However, fairness is a cross-layer problem, and thus MAC-layer solutions are useless if higher layer protocols are unfair. This is not true vice versa; an ideal transport protocol can enforce fairness even if the underlying MAC protocol is unfair [7].

Several papers have appeared that have taken linear programming (LP) approach to routing and fairness. One of the benchmark paper in LP-based routing, with several linear constraints, is presented in [8]. The paper addresses two interrelated questions: *what is the maximum throughput capacity of an arbitrary (ad hoc) network with given source-destination pairs can this maximum throughput capacity be achieved by jointly routing packets and scheduling transmissions?*

The authors devise an LP formulation that maximizes aggregate rates and incorporates any requirements that can be modeled as linear constraints. The paper provides a proof that using their LP formulation, all needed packet transmissions can be feasibly scheduled and that their solution to the maximum concurrent flow problem is a constant factor away from the optimal. The problem in their proposed scheme is that the authors use an infinitesimally divisible flow model for data transmission. This means that data packet can be divided into pieces and transmitted along *all possible paths* between source and destination, which lead to very complex receiver structures and possibly to a long delay between the arrival of the first and the last data segment. In addition, storing and updating of all possible routing paths leads to large routing tables and network overhead.

In [3], optimized routing in WMN is considered with fairness constraints. The paper points out that past work can be categorized into two different strategies: heuristic and optimization problem. Heuristic methods lack the theoretical foundation to analyze how well the method is working, while optimization problems can be far too complex in practise or make too much simplified assumptions. The paper inspects and analyzes optimal routing with uncertain traffic demand and fairness constraints, thus the authors end up with a stochastic maximum concurrent flow optimization problem. Unfortunately, their LP-formulation seeks to maximize scaling factor δ, which defines the fraction of traffic that can be transmitted for each flow, instead of guaranteeing fairness.

In [4], a topology control algorithm (TCA) and a new routing metric suitable for WMN, namely, collision domain (CD), are presented. The term topology control refers to any set of network operations that lead to a connected topology, for example, node placement, channel assignment, power control, and routing. It is shown that the proposed TCA performs better than conventionally used metrics, that is, hop count and interference, in the terms of minimum collision domain. On the other hand, the paper makes simplified assumptions such as one user per router, absolute fairness is said to be enforced and only one radio per router is assumed.

Our work is mainly based on the work by Malekesmaeili et al. [4] and Kumar et al. [8]. From [4], the topology control concept and collision domain routing metric are taken as baseline for routing with modifications. From [8], linear programming-based approach to rout and rate maximization are adopted with modifications to constraints and routing path selection. The essence of this work is to develop a fair subscriber-aware routing algorithm for WMN, in which the positions of subscribers are taken into account in order to ensure fairness without crippling the network performance. The algorithm is called Subscriber Aware Fair Routing in WMN (SAFARI).

3. Preliminaries

In this section, basic definitions and concepts are introduced and explained. We consider multiradio WMN modeled as a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is the set of nodes and $\mathbf{E}$ the set of wireless links (edges). Each link $e \in \mathbf{E}$ has a certain amount of data to send, $x(e)$, and each e has a set of interfering links $\mathbf{I}(e)$, which is based on the transmitter-receiver (Tx-Rx) model [8].

3.1. Network Model. We consider WMN comprising of mesh users, mesh routers, and mesh gateway routers. Mesh users can be mobile and nomadic with stringent power constraints, mesh routers are considered to be stationary without power constraints, and mesh gateway routers are similar to mesh routers except that they have gateway properties, that is, they can connect to an external network. Our network model is illustrated in Figure 1.

3.2. Fairness. In the context of wireless networks, fairness means that every user receives a fair share of the network resources (e.g., time and frequency), taking into account user's service requirements. Different services can have very different requirements, for example, voice calls have strict delay requirements and relatively low data rates, while file downloading has high bandwidth and low delay requirements. These different requirements should be taken into account, when designing a fair network protocol.

It is important to notice that assuring fairness is a cross-layer problem, since unfairness occurs in MAC (e.g., channel access and scheduling) and transport layers (e.g., congestion control). Current network protocols (e.g., IEEE 802.11) ensure user fairness only on one-hop communication or seek to maximize aggregate throughput of the network [4].

Three popular definitions of fairness are absolute, max-min, and proportional fairness. Absolute fairness is defined as equal rates among all users, max-min fairness is enforced if no user can increase its rate without decreasing some other users' smaller rate at the same time, and a set of allocated rates is proportionally fair if any other feasible rate allocation results in zero- or negative-aggregate change.

In this work, we use a simple fairness index $\lambda \in [0, 1]$:

$$\lambda = \frac{\min(\mathbf{R})}{\max(\mathbf{R})}, \tag{1}$$

where $\mathbf{R}$ is the set of user rates, $\mathbf{R} = \{r_1, r_2, \ldots, r_{|N_u|}\}$, where N_u is the number of users. When $\lambda = 0$, some user's rate are allowed to starve and when $\lambda = 1$, absolute fairness is enforced. Together with linear programming-based rate allocation, our fairness index enforces proportional fairness when $\lambda > 0$ and also satisfies quality of service (QoS) requirements if minimum allowable rate is set to QoS threshold.

3.3. Collision Domain. In the work in [9], WMN capacity has been addressed in form of a bottleneck collision domain (BCD). In order to get a formal definition of BCD, we need to first define CD and the corresponding CD load. CD of link $e(i, j)$, C_e, is the set of wireless links $\mathbf{E}$, which need to be silent due to the shared nature of the wireless medium, when link $e(i, j) \in \mathbf{E}$ is active. The link e itself is also included in C_e, since it also contends over the medium. Indices i and j are the transmitting and receiving nodes, respectively. More formally, C_e is defined as

$$C_e = e + \mathbf{I}(e), \tag{2}$$

where $\mathbf{I}$(e) is the set of edges interfering with edge e, for all $e \in \mathbf{E}$.

Each link $e(i, j)$ for all i, $j \in \mathbf{V}$ has a certain amount of data to send, $x(e)$, and all the data is accumulated in the collision domain. Thus, CD load of link e is defined as

$$C_l(e) = x(e) + \sum_{f \in \mathbf{I}(e)} x(f), \tag{3}$$

where $x(f)$ is the amount of data on link $f \in \mathbf{E}$.

BCD is the collision domain that has the most data to forward in an arbitrary topology, thus limiting the capacity of the network. More formally, BCD of a network C_b is defined as

$$C_b = \max(\mathbf{C}_l), \tag{4}$$

where $\mathbf{C}_l$ is the set of CD loads, that is, $\mathbf{C}_l = \{C_l(1), C_l(2), \ldots, C_l(n)\}$, $n = |\mathbf{E}| \cdot |\mathbf{E}|$ is the number of edges in the network.

In Figure 2, collision domain of a link 4 → 3 is illustrated with the two shaded circles. In other words, collision domain of a link 4 → 3 is the set of links included or intersecting the two shaded circles. The consideration of collision domain models the performance degradation of multihop communication in contrast to single-hop, thus it captures essential properties of MAC protocol without actually making assumptions of the used MAC-layer protocol.

The technology-dependent link capacity (theoretical maximum throughput (TMT)) is calculated in [10], and it was used in [9] to assess link capacity, which is also limiting the network capacity since the accumulated traffic of a link cannot exceed the link capacity. In Figure 2, the total load of the collision domain $C_{4\rightarrow 3}$ is 20 U since it is the accumulated traffic of links 7 → 6, 9 → 5, 5 → 4, 6 → 4, 4 → 3, 3 → 2 and 2 → 1, where U is the amount of data that mesh user transmits and it is same to all users (Figure 2) for simplicity of notation. Thus, the throughput per node G_m is bounded by $G_m \leq \mathrm{TMT}/20$ [9]. Note that this is not necessarily the BCD of the network. The above calculation needs to be done to every link in order to find the BCD.

C_l can be used as a cumulative routing metric, combined with Tx-Rx model, it reflects wireless interference, it takes

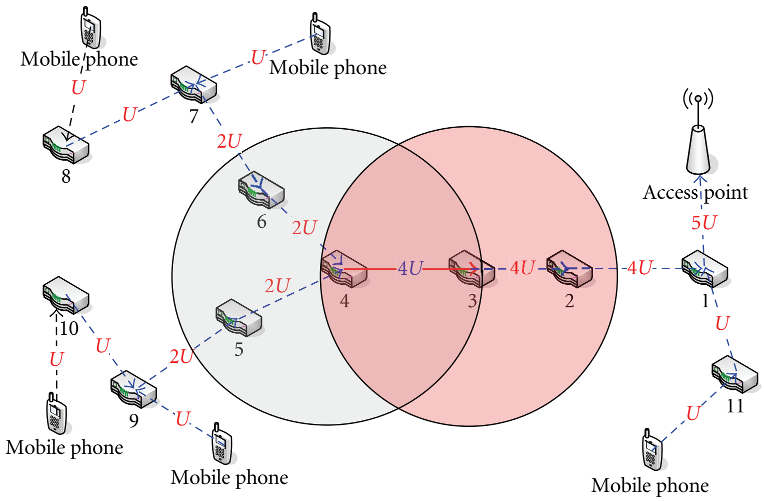


Figure 2: Illustration of collision domain of link 4 → 3.

into account network congestion in a certain area and models MAC-layer collisions since interfering nodes are not allowed to transmit simultaneously. The BCD can be used to estimate the maximum number of users in a network with a fixed data rate since if each user transmits at rate r, then $C_b = m\cdot r$, where $m \in \mathbb{Z}^+$, and the link bandwidth is L, the throughput per node G_m will be bounded by $G_m \leq L/m$ [9]. Thus, the number of users the network can support is

$$N_u = \frac{G_m}{R_n} \leq \frac{L}{mR_n}, \tag{5}$$

where the required data rate for each user is R_n.

3.4. Routing Metrics. A good routing metric for WMN is aware of network topology, takes into account network characteristics, and is isotonic [11]. Isotonicity means that the order of path lengths of two paths is preserved if they are appended or prefixed by a common third path. An isotonic metric assures loop-free routing, simple implementation, and minimum weight paths using Dijkstra's algorithm.

Proposed routing metrics for WMN are hop count, distance, weighted cumulative expected transmission time (WCETT), and CD. Hop count is used in AODV [12], but it fails to address WMN characteristics and network congestion. Distance-based metric is usually used with modified Dijkstra's algorithm and it suffers from same things as hop count-based metric. WCETT was proposed by Draves et al. [13] and it is a combination of loss rate with a priori-known packet loss probability, bandwidth, and interference of a link. Unfortunately, WCETT is not isotonic as shown in [11]. CD was proposed as a routing metric by [4], which is an excellent choice since it models wireless interference, MAC layer collisions, and is isotonic. Based on the above discussion, CD is used in this work as a routing metric.

3.5. Linear Programming. LP is a mathematical optimization method that seeks to optimize (i.e., minimize or maximize) a linear objective function subject to equality and inequality constraints. In our work, we are using LP to maximize the user rates with capacity and fairness constraints with a selected path. Our LP is modified from [8] and is formulated as follows:

$$\max \sum_{i \in \mathbf{V}} r_i \quad \text{subject to}, \tag{6}$$

$$x(e) + \sum_{f \in \mathbf{I}(e)} x(f) \leq \text{TMT} \quad \forall e \in \mathbf{E}, \tag{7}$$

$$r_i \geq \lambda r_j \quad \forall i, j \in \mathbf{V},\ i \neq j, \tag{8}$$

$$R_n \leq r_i \leq R_m, \tag{9}$$

where r_i is the rate of user i and TMT is the theoretical maximum throughput (i.e., physical data rate a link can transmit [10]), R_n is the minimum required rate, and R_m is the maximum feasible rate. However, (7) is the capacity constraint, (8) is the fairness constraint, and (9) is constraining the rates. Solving this optimization problem leads to a rate allocation $\mathbf{R} = \{r_1, r_2, \ldots, r_{|N_u|}\}$ that can be feasibly scheduled, as shown later on.

The obtained rate allocation is dependent on the random positions of the users. Thus, the obtained aggregate throughput varies significantly with different user positions and there is a need for statistical processing. For this reason, standard deviation, σ, is introduced as

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=1}^{N} (z_k - \overline{z})^2}, \tag{10}$$

where N is the number of random drop of users to the network area, the kth rate allocation, z_k is the sum of user rates on kth random drop:

$$z_k = \sum_{l=1}^{|\mathbf{R}_k|} \mathbf{R}_k(l), \tag{11}$$

where $\mathbf{R}_k$ is the rate allocation of kth random drop of users, and $\overline{z}$ is the average of all rate allocations with certain number of users:

$$\overline{z} = \frac{1}{N}\sum_{l=1}^{N} z_l. \qquad (12)$$

3.6. Channel Assignment. The main purpose of any channel assignment (CA) algorithm is to minimize interference, maximize aggregate throughput, as well as capacity or fairness. The assignment of radios and channels to mesh nodes is far from trivial. In [14], it is proved that simply assigning first channel to the first node and second channel to second node, and so forth, is far from optimal.

In [15], a taxonomy of CA schemes is presented and a new CA algorithm, called mesh-based traffic and interference aware channel assignment (MeshTiC), is introduced and evaluated. The MeshTiC assigns channels to links in decreasing order based on a link's rank

$$\text{Rank}(i) = \frac{A(i)}{H(i)\cdot N_r(i)}, \qquad (13)$$

where $A(i)$ is the aggregate traffic that traverses through a certain node i, $H(i)$ is the minimum number of hops from node i that needs to be done in order to reach a gateway, and $N_r(i)$ is the number of radios in node i. MeshTiC has been chosen here since it takes into account the traffic load on links, can be modified to incorporate interference, and has low complexity.

4. Proposed Algorithm: SAFARI

Next, the centralized SAFARI algorithm is explained in detail, pseudocode and simulation results are presented. The SAFARI algorithm uses CD as a cumulative routing metric, assigns channels to links using a modified version of the MeshTiC algorithm, and uses a linear programming framework to assign rates to users taking into account capacity, fairness, and rate constraints, see (6), (7), (8), (9).

MeshTiC algorithm is modified such that in (13), $A(i)$ is estimated by using CD of link i based on the initial geographical positions of users and $H(i)$ is estimated as distance to the nearest gateway. This way CA is fixed until user positions change dramatically, and channels can be assigned before routing and rate allocation.

Next, a high-level pseudocode of the SAFARI algorithm is presented in Algorithm 1. In Table 1, the used abbreviations in the pseudocode are explained. In Algorithm 1, on line 1, necessary network information is collected, that is, router, gateway router, and user positioning **V**, and the set of wireless links **E**. This serves as a basis for the whole algorithm to work. The positions of routers and gateways are easy-to-obtain since they are stationary, and they are handled by a centralized entity. The positions of users can be obtained by multilateration or simply modeling the position by the routers the user can reach.

```
(1) Collect network information: V and E.
(2) Compute initial estimate of G.
(3) Assign channels to E, update G accordingly.
(4) Solve best known paths using G and FW.
(5) for $k = 1$ to $|\mathbf{S}|$ do
(6)     For user $k$, choose the router from which the
        best known path
(7)     to any gateway is shortest.
(8)     Connect to this router.
(9) end for
(10) Sort users such that users in low CD regions are
     routed first.
(11) Store the order in $\mathbf{S}_{\text{new}}$.
(12) for $k = 1$ to $|\mathbf{S}_{\text{new}}|$ do
(13)     Calculate paths from $\mathbf{S}_{\text{new}}(k)$ to all gateway
         routers using FW.
(14)     Choose optimal gateway router and select the
         corresponding path.
(15)     Update G.
(16) end for
(17) Solve the LP-problem in order to find optimized
     rates.
```

Algorithm 1: SAFARI.

Table 1: Abbreviations used in the pseudocode.

Abbreviation	Explanation
CD	Collision domain
E	Set of edges
FW	Floyd-Warshall's algorithm
$\mathbf{G}(i, j)$	A graph representing C_l on each link for all $i, j \in \mathbf{V}$
S	Set of mesh users (i.e., sources)
$\mathbf{S}_{\text{new}}$	New routing order based on to which router each user is connected to
V	Set of all nodes in the network

On line 2, link weight matrix **G** is calculated based on the initial positions of users, CD and Tx-Rx models. The **G** ultimately determines the routing path selection and it is modified several times in SAFARI so that it always reflects the current network condition. The first calculation of **G** does not take into account CA, since the used MeshTiC CA algorithm needs an estimate of the traffic demand and it is estimated using CD based on the initial positions of users.

On line 3, channels are assigned using modified MeshTiC and **G** is updated to match CA. Channels can be now assigned to **E**, since we have an estimate of traffic in the network. **G** needs now to be updated to match CA. In other words, the Tx-Rx model takes into account the CA, that is, links interfere only if links are within the interference range and use the same channel.

On line 4, the *best known paths* are solved using **G** and FW's algorithm. In this context, the *best-known paths* are the *"shortest"* paths to gateways and they are used in the determination of the best router for each user to attach to (lines 6–8). Modified FW is used, since it can be made to incorporate CD metric and performs necessary routing with relatively short time, that is, running time of FW is $O(|\mathbf{V}|^3)$. The determination of the router each user attaches to is decided now by simply selecting the router from which the path to any gateway is *"shortest"*. With this kind of router selection, the randomness of user positions is diminished and overall network throughput is increased, since in most cases, the router selection procedure leads to smaller number of hops for user to reach destination.

The routing order is decided on line 10. Low CD areas, that is, links and corresponding nodes with low C_l are routed first since these areas are usually at the border of a network, thus their routing is essential. This comes from the fact that when the users are far away from gateways, the number of hops increases. Now, if far away users are routed last, their number of hops increases even more. Keep in mind that as the number of hops increase, the capacity constraints become stricter and the throughput decreases while delay increases. Thus, the aggregate number of hops in the network should be minimized, and routing far-away-users first is one way to do it.

On lines 13–15, every user is routed in the decided order to the best gateway and **G** is updated to reflect current network condition. Each user is routed individually using FW's algorithm and the best gateway is selected according to cumulative CD metric. The main reason for using FW is that even a large number of gateways does not increase running time of the algorithm. This is the final routing path selection. After every user's routing, **G** is updated according to and along the chosen path.

Line 17 executes LP-problem, which allocates the highest possible rates subject to capacity and fairness constraints. Solving the LP-problem (6), (7), (8), (9), optimal rate allocation with chosen paths is performed.

The original contributions of SAFARI are as follows.

(1) Positions of users are taken into account in

  (i) CA by traffic load estimation with the help of CD,

  (ii) determination of which router each user attaches to.

(2) Determination of routing order.

The positions of users are taken into account in CA so that in the rank calculation (13), the traffic load is estimated with CD. In addition, the positions of users help to determine the router each user attaches to. This is determined by finding the *best-known paths* to the best gateways using FW's algorithm with CD estimated by user positions. With our router selection scheme, the number of users attached to each router is not random, as in cases where simply the closest router is chosen, but determined by considering transmit powers, available gateways, and other users' positions.

4.1. Feasibility of the Algorithm. When comparing the SAFARI algorithm to any wireless network routing algorithm, several similarities and differences arise. Every routing algorithm needs to collect network information, at least **V** and **E**, in order to be able to route data from source to destination. Also, every routing algorithm should have at least an estimate of link weights, that is, hop count, distance, interference, bandwidth, or CD, in order to compute **G**. Finally, every routing algorithm needs a path selection algorithm (e.g., Dijkstra or FW). These properties are also implemented in SAFARI, and thus there is no extra complexity in that regard.

There are a few factors that increase SAFARI's complexity compared to, for example, a simple distance-based routing algorithm. The calculation of *best-known paths* and the following router selection for each user increases complexity compared to algorithms where simply the closest or the farthest router is selected. Sorting users so that low CD regions are routed first increases complexity only slightly since all the necessary information is already calculated and stored in **G**. The biggest factor increasing complexity is the recursive path selection with FW and updating **G**. This recursion is done because it allows the routing algorithm to adapt to changing traffic conditions. In addition, rate allocation by LP-problem solving increases the complexity and running time especially with a large number of users. Based on the above discussion, it can be stated that the performance gain of SAFARI, as shown later, comes with the cost of increased complexity. Nevertheless, this increase in complexity is not too great to make SAFARI infeasible for practical implementation since FW's algorithm is the most complex with a running time proportional to $O(|\mathbf{V}|^3)$. Thus, SAFARI can be solved in polynomial time.

Since SAFARI's rate allocation is based on the LP-formulation by Kumar et al. [8], it can be shown that this rate allocation leads to feasible scheduling. The feasible scheduling of SAFARI is formalized in Theorem 1.

Theorem 1. *The LP formulation* (6), (7), (8), (9) *(i.e., rate allocation) used in the SAFARI algorithm results in a stable schedule, that is, flows are given enough transmission opportunities in a finite period of time. In addition, the rate allocation is a constant factor away from the optimal solution to the corresponding flow problem.*

The proof of feasible scheduling in a TDMA-based system is based on [8, Lemma 1]. The intuition behind the proof is that link flows can only be scheduled in finite time if there are enough transmission opportunities for each flow, that is, there is enough bandwidth on the link. The detailed proof is available in [8].

Lemma 1. *A sufficient condition for link flow stability,*

$$\forall\, e \in \mathbf{E}, \quad x(e) + \sum_{f \in \mathbf{I}_{\geq}(e)} x(f) \leq \mathrm{TMT}, \tag{14}$$

where $\mathbf{I} \geq (e)$*, is a subset of edges in* $\mathbf{I}(e)$ *which are greater than or equal to e in length.*

4.2. Simulations. The simulations are performed using MATLAB software version R2007b. In Table 2, the most important simulation parameters are presented. The communication

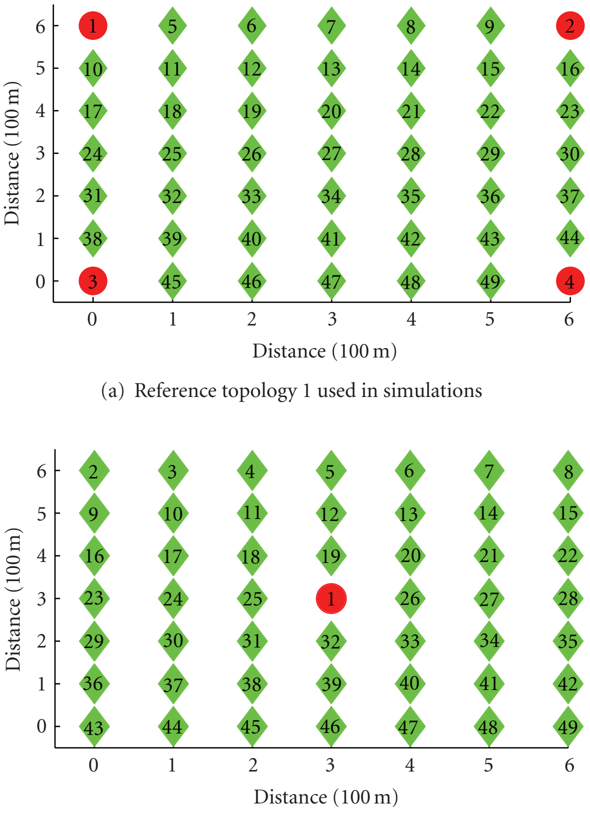
Table 2: Simulation parameters.

Parameter	Value
Communications range	$20\sqrt{50}$ m
Interference range	$40\sqrt{50}$ m
Link capacity (TMT)	43 Mbps
Number of radios	1, 3 and 12
Number of users	1–20
N	750
R_n	0 Mbps
R_m	3 Mbps
Step size λ	0.1
Step size users	1
Topology	Figure 3

range is fixed and set to $20\sqrt{50}$ m, since routers are 100 m and $20\sqrt{50}$ m apart as shown in Figures 3(a) and 3(b). Thus, most of the routers can reach eight other routers. The interference range is fixed and set to $40\sqrt{50}$ m, which is twice the communication range. The number of users in simulations varies between 1–20. This choice allows the observation of the effect of number of users on throughput and fairness, and keeps simulation times bearable. Number of random drop of users, N, into the simulation area is set to 750, since the achieved throughput varies significantly with different user positions. Users are dropped following a uniform distribution into the simulation area. The step size of λ is set to 0.1 and it defines the incremental value λ is given in simulations. This choice allows for observing the tradeoff between throughput and fairness. The step size of users is set to the minimum (i.e., one) in order to observe the effect of users on throughput. Link capacity is set to 43 Mbps, as one of the options for link capacity defined and calculated in [10], and is assumed to be constant. The lower and upper bounds for user rates, R_n and R_m, respectively, are set such that total starvation of users is enabled and that user rates have a realistic upper bound enabled by the physical layer data rate.

In the simulations, it is assumed that each user has the same QoS requirement and the corresponding data rate is tried to achieve with limitations from the LP constraints. The number of routers is kept relatively low since when there is too many routers leads to very long simulation times. On the other hand, using only a few routers is not practical, since then the routing algorithm is tied to only a couple of possible paths. In the simulations, defined numbers of users are dropped uniformly into an area covered by a certain predefined topology (e.g., Figures 3(a) and 3(b)), are allowed to exceed this area by 100 m, and routed to destinations using the algorithm in question. This is done several hundred times since the distribution of users has a significant effect on performance.

There are three algorithms that are used throughout the following simulations. The first one implements SAFARI algorithm and is referred to as SAFARI in the following. The second one implements the TCA proposed in [4] and is referred to as CD *metric* in what follows. The third one is a simple distance-based algorithm that uses Dijkstra's algorithm, and is referred to as *distance metric* in what follows.



(a) Reference topology 1 used in simulations

(b) Reference topology 2 used in simulations

Figure 3: Two reference topologies used in simulations.

The following simulation results are obtained in a 7 × 7 grid topology with four gateways and 45 routers as shown in Figure 3(a) (see [4]), and a 7 × 7 grid topology with one gateway and 48 routers as shown in Figure 3(b), where red circles are gateways and green diamonds are routers.

4.2.1. Comparison. In Figures 4–6 and Table 3, comparison of the three used routing schemes by illustrating routing paths, corresponding throughputs, and BCD are presented with 30 users, 12 channels, and $\lambda = 1$. The red circles are gateways, green diamonds routers, blue dots users, red lines router-to-router routing paths, and blue lines user-to-router hops. Figure 4 shows how each user, using SAFARI, selects the best router for itself and how two paths, for two different users, are separated at a node in order to avoid congestion on that link. In other words, the two users are guided now along noninterfering paths. It is obvious that using SAFARI leads to higher transmit powers on users, see the blue lines, while the number of hops is diminished. The increased transmit power can be unwanted in some scenarios but if power consumption is not a crucial issue, higher data rates are achieved.

Figure 5 shows the corresponding routing paths with the same user positions. Now, users are connected to the

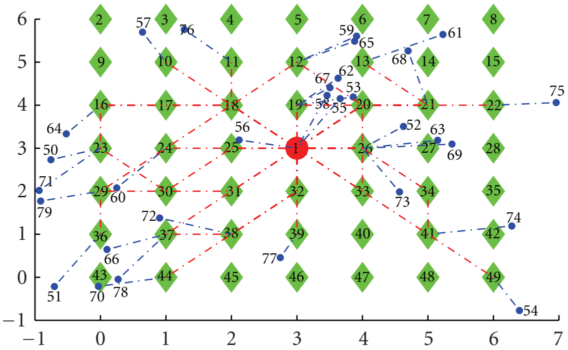


Figure 4: Routing done with SAFARI.

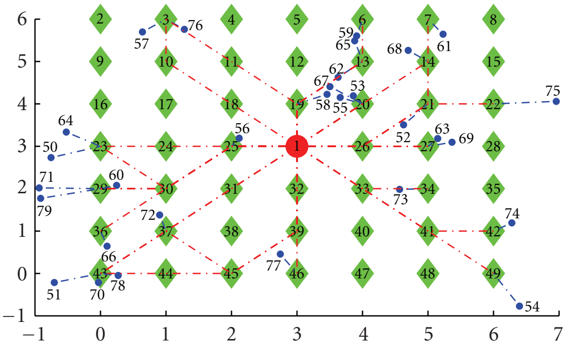


Figure 5: Routing done with CD *metric*.

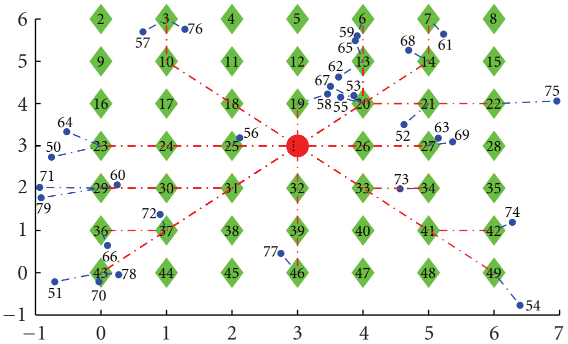


Figure 6: Routing done with *distance metric*.

Table 3: Corresponding-achieved throughput and BCD from Figures 4–6.

Algorithm	Throughput (Mbps)	BCD (Mbps)
SAFARI	5.584	2.45
CD metric	2.972	3.7
Distance metric	2.687	5.5

nearest router and are then routed using CD as a metric. It is obvious that using CD *metric* leads to lower power consumption while the number of hops is increased. Similar to SAFARI, CD *metric* also guides users to noninterfering paths. The main difference between these two schemes is that with CD *metric*, the number of hops is greater, thus finding noninterfering paths is harder. This is shown so that a fewer number of noninterfering paths are selected.

Figure 6 illustrates the path selection with *distance metric* scheme. It can be seen that also this scheme selects the nearest router for each user to attach to. Then, the paths are selected blindly without considering CA and link congestion. This leads to shorter paths than using CD *metric* scheme but some links are heavily congested, and thus limiting the network capacity. The *distance metric* scheme is the simplest scheme while worst on the performance, as seen later on.

Table 3 shows the achieved throughput and BCD using SAFARI, CD *metric*, and *distance metric* schemes with the shown user positions in Figures 4–6. SAFARI achieves almost twice as much throughput than the two others and has significantly lower-average BCD. CD *metric* performs slightly better than *distance metric*. As mentioned before, this performance gain comes with the cost of increased transmit power and algorithmic complexity.

Figures 7–9 show the average traffic distribution on each gateway and router using the three models with topology shown in Figure 3(a), 12 channels, $\lambda = 1$, 20 users, and 1000 random drop or users. The traffic distribution is obtained so that the number of users attached or passing a router/gateway is summed up in each random drop, and after 1000 drops, it is divided by the number of users and number of random drops. Horizontal axis shows the router/gateway indices as presented in Figure 3(a). It can be seen that using SAFARI in Figure 7, some routers are used rarely, especially the ones far away from gateways. This is due to the fact that SAFARI selects routers to users so that routers close to gateways are preferred. This location-dependent router starvation should be taken into account in the deployment of routers, that is, sometimes network deployment cost can be reduced by deploying less routers. Another remark is that besides the starved routers, SAFARI performs load balancing to some extent, that is, traffic is divided evenly among routers that are at equivalent network positions (e.g., routers next to gateways).

In Figure 8, the same traffic profile is presented with CD *metric*. This scheme performs load balancing, which is shown especially in gateways, indices 1–4, as the number of users per gateway is equal in the long run. With CD *metric*, routers are not starved in any location and the traffic is divided smoothly among routers. This is another advantage of CD as a routing metric, it inherently performs load balancing. The difference to SAFARI, which also uses CD as a metric, is the router-selection procedure and routing order, which disables full-load balancing among routers.

In Figure 9, the traffic profile using *distance metric* model is shown. It is obvious that this model fails to achieve load balancing, which is shown in uneven gateway utilization and heavy congestion in some routers. This illustrates the effect of using blind distance-based routing and not taking into account network condition.

These results show that SAFARI is superior to the two other schemes with this topology, number of users, λ, and number of channels. Next, the performance of SAFARI is shown in scenarios where several parameters are changed.

Table 4: Average number of hops using the three simulation models.

Algorithm	1 gateway, 1 channel	1 gateway, 12 channels	4 gateways, 1 channel	4 gateways, 12 channels
SAFARI	2.08	2.25	1.43	1.43
CD *metric*	2.48	2.65	1.85	1.91
Distance metric	2.48	2.48	1.86	1.84

Table 5: Percentage of starved users.

Algorithm	1 gateway, 1 channel	1 gateway, 12 channels	4 gateways, 1 channel	4 gateways, 12 channels
SAFARI	0.820	0.464	0.401	0.148
CD *metric*	0.910	0.554	0.631	0.290
Distance metric	0.913	0.628	0.640	0.337

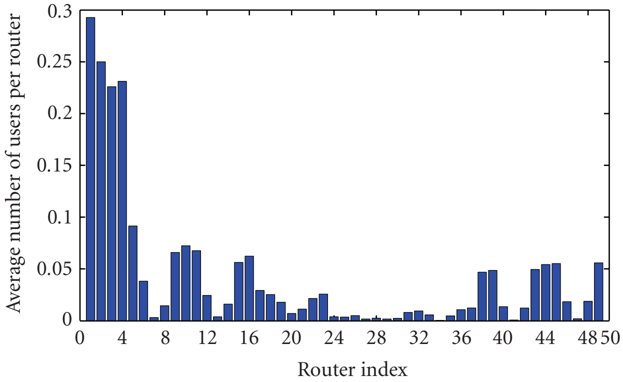


Figure 7: Traffic distribution among routers with SAFARI scheme.

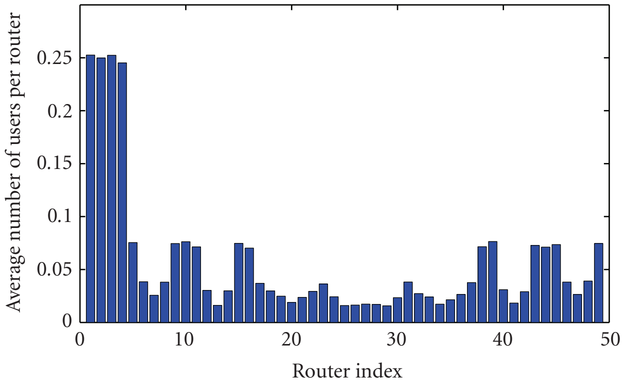


Figure 8: Traffic distribution among routers with CD *metric* scheme.

4.2.2. Number of Hops and Starved Flows. First, the average number of hops users need to make in order to reach a gateway, and the percentage and positions of starved users are observed in topologies presented in Figures 3 with 1 and 12 channels, $\lambda = 0$, 20 users, and 1000 random drops. In Table 4, the average number of hops users need to make in order to reach a gateway are presented. It can be seen that SAFARI has the lowest number of hops in all the four simulated cases, on the average approximately 0.5 hops less than the other simulation cases. With one channel, the CD *metric* and *distance metric* have the same number of hops, while with 12 channels, the CD *metric* has more hops. This stems from the fact that CD *metric* avoids congested areas, which inevitably leads to more hops. In addition, the number of hops is lower with 4-gateway case, since now there is a gateway closer to more users than in 1-gateway case (see Figures 3(a) and 3(b)).

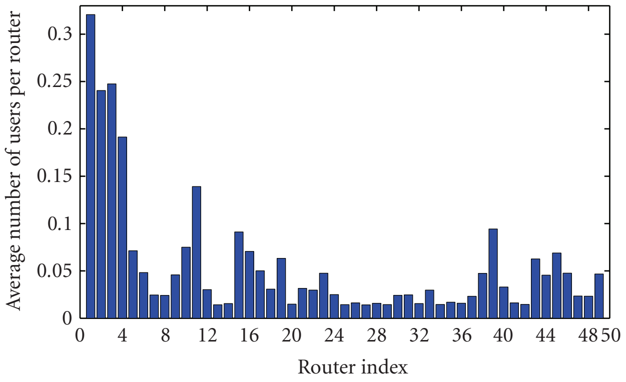


Figure 9: Traffic distribution among routers with *distance metric* scheme.

In Table 5, the percentage of starved users (i.e., when $r_i = 0$ bps) are presented with the same parameters as in Table 4. It is obvious that severe unfairness occurs especially with low number of channels and gateways. Blocking 90% of users in order to maximize the aggregate throughput is very unfair and noneconomical to service providers as users will not tolerate such blocking percentages. The SAFARI has the lowest percentage in all the cases, even though it only guarantees reasonable performance with 4 gateways and 12 channels with 14.8 % blocking rate. CD *metric* is better than *distance metric* and they both have a poor performance in all the four cases. The results in Table 5 points out why $\lambda = 0$ is not a good choice even though it maximizes the aggregate throughput.

Table 6: Average distance in meters to gateway of starved users.

Algorithm	1 gateway, 1 channel	1 gateway, 12 channels	4 gateways, 1 channel	4 gateways, 12 channels
SAFARI	307	307	199	205
CD *metric*	321	327	237	236
Distance metric	322	318	240	246

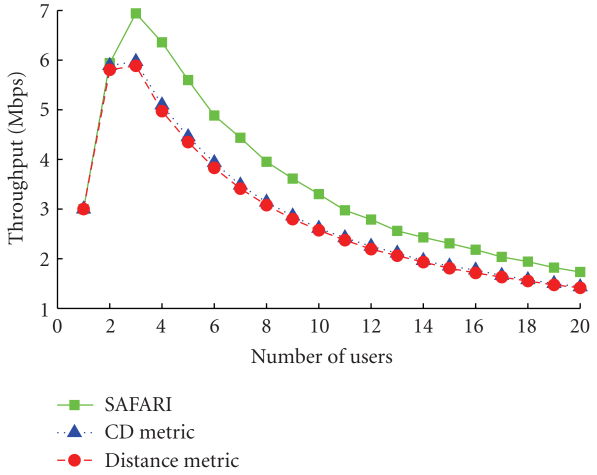


Figure 10: Throughput versus number of users with 1 channel and $\lambda = 1$.

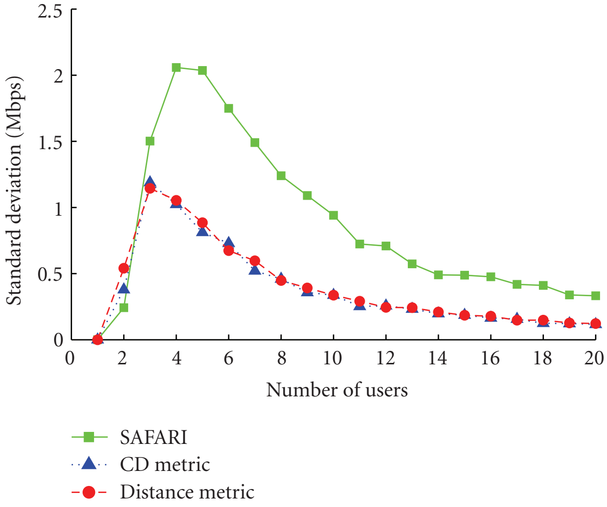


Figure 11: Standard deviation versus number of users with 1 channel and $\lambda = 1$.

In Table 6, the average distance to nearest gateway of starved users are presented. One should know that the maximum distance to a gateway is $400\sqrt{2} \approx 566$ m since users are allowed to be dropped 100 m outside the routers. It is obvious that the starvation distance does not depend on the number of channels rather it depends on the availability of gateways nearby. SAFARI has the lowest starvation distance value in all the cases, which means that users that are far away are not that easily starved. CD *metric* and *distance metric* have starvation distances of same magnitude.

4.2.3. Effect of Number of Channels. Next, the number of channels is limited to one in order to see how the three considered models perform in a single-channel environment. Naturally, there is no need for a CA in this case and all links that are within each other's interference range interfere with each other.

Figure 10 shows how throughput behaves as a function of number of users in a network defined by Figure 3(b), with one channel and $\lambda = 1$. It can be seen that when there is only a few users in the network, all of the users can transmit at their peak rate. After 2-3 users, the network becomes crowded and all the users rates need to be constrained, which results in a steady decrease in the overall throughput. As the number of users grow, the throughput starts to saturate. SAFARI achieves the best performance when the number of users is greater than two, and CD *metric* is slightly better than *distance metric*.

In Figure 11, the standard deviation (see (10)) of the three models is plotted in the same case as in Figure 10. In these simulations, the standard deviation measures the variation in the aggregate throughput between each random drop of users. It is apparent that all the three models have a large standard deviation when compared to the corresponding throughput. This reflects the fact that user positions have a significant effect on the throughput, thus taking into account the user positions can lead to performance gain. The standard deviation of CD *metric* and *distance metric* are almost identical conforming the superiority of CD to distance as a path metric. SAFARI has the highest standard deviation, which can be explained by considering the following two cases.

(1) Users are positioned so that it can be exploited, for example, near gateways or far away from each other, thus using SAFARI leads to high throughput.

(2) Users are poorly positioned, for example, forming clusters, and taking into account their positions, does not lead to a significant performance gain.

Figure 12 plots throughput as a function of the fairness index (1) with 1 channel, 20 users, and in the topology shown in Figure 3(b). This simulation result points out the tradeoff between throughput and user fairness. When $\lambda = 0$, some users are allowed to completely starve and other users, who are usually near gateways, are given the whole bandwidth. This leads to high throughput but is very

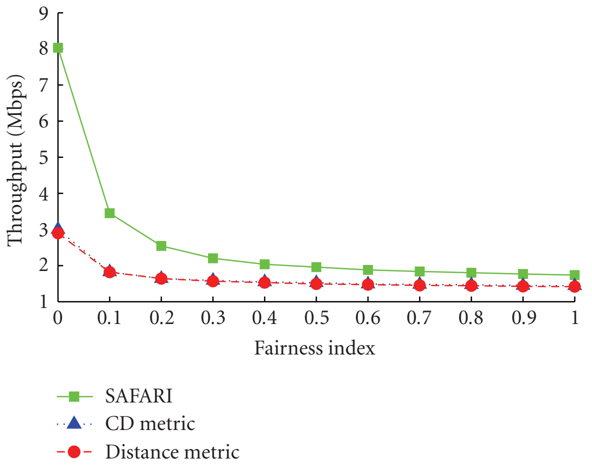


Figure 12: Throughput versus farness index with 1 channel and 20 users.

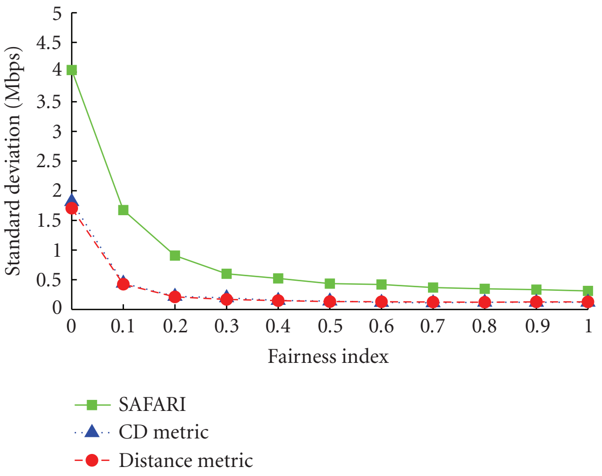


Figure 13: Standard deviation versus farness index with 1 channel and 20 users.

unfair and undesirable. When the fairness index increases, the throughput decreases, which stems from the fact that now user rates are restricted to or near the smallest user rate. In addition, in this simulation scenario, SAFARI achieves the highest throughput, it can provide consistently twice as much throughputs than the other two models, which achieve throughput of same magnitude. It should be pointed out that since throughput saturates quickly as fairness index increases, even a relatively low fairness criterion is able to lower the overall throughput; but as pointed out earlier, cases with low λ are unfair and most of the attention should be focused near $\lambda = 1$.

Standard deviation, in the same simulation case as in Figure 12, is presented in Figure 13. As expected, the highest standard deviation occurs with $\lambda = 0$, since with this fairness index value, the aggregate throughput is also highest, and CD *metric* and *distance metric* cases have a very similar standard deviation curves. SAFARI has again the highest standard deviation which stems from the above enumerated reasons.

Figure 14 plots the average BCD with respect to the number of users with 1 channel, $\lambda = 1$, and in a topology illustrated in Figure 3(b). Now, rate allocation is not used since user rates are fixed beforehand. Using (5), it is obvious that the average BCD should be as low as possible in order to have maximum number of users in a network. Considering this fact, the SAFARI is once again the best one and CD *metric* the second best. SAFARI's dominance starts show with 4 users, and CD *metric* starts to outperform *distance metric* after ten users. Average BCD versus fairness index is not plotted here since with fixed equal user rates (i.e., $\lambda = 1$), the BCD versus fairness index plot would be meaningless.

In the following, the results presented in Figures 10–14 are referred to as *baseline simulation set*. Next, the number of channels is increased to 12 and the corresponding results as in the *baseline simulation set* are presented in Figures 15–19. Once again, SAFARI achieves the best performance measured in throughput and average BCD. Now, the performance gain compared to *distance metric* is almost 100% and compared to CD *metric* it is approximately 40% (see Figures 15-16). One should notice that with increasing number of channels, CD *metric* starts to outperform *distance metric*. This stems from the fact that with many channels, CD metric can choose noninterfering paths for different flows, which leads to smaller CD loads on links, thus throughput increases.

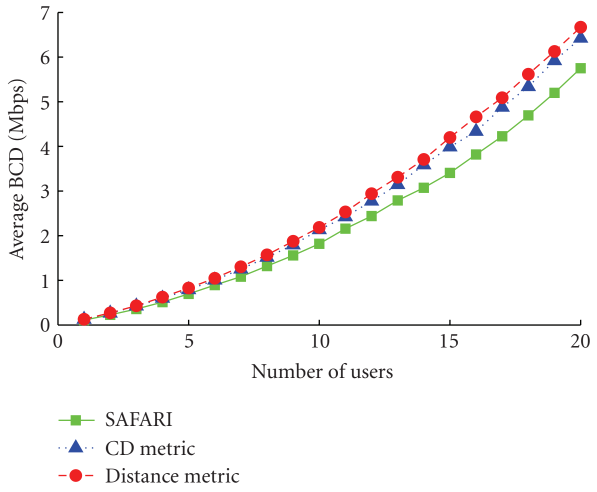


Figure 14: BCD versus number of users with 1 channel and $\lambda = 1$.

It can bee seen by comparing Figures 15-16 to Figures 10, 12, that using 12 channels instead of one results in 500% throughput increase when $\lambda = 1$. This shows the benefit of multiradio concept (i.e., with increasing cost comes increased performance). One should notice that even though the number of orthogonal channels is increased from one to 12, the throughput is not increased with the same ratio. This stems from the fact that 12 channels does not result in empty $\mathbf{I}(e)$, that is, some links still interfere with each other which leads to strict capacity constraints and lower throughput enhancement.

Figures 17 and 18 point out that the higher throughputs of SAFARI and CD *metric*, in Figures 15 and 16, compared to *distance metric* come with the cost of increased σ. One might notice that the standard deviations of SAFARI and CD *metric* fluctuate somewhat, while *distance metric* results in smooth

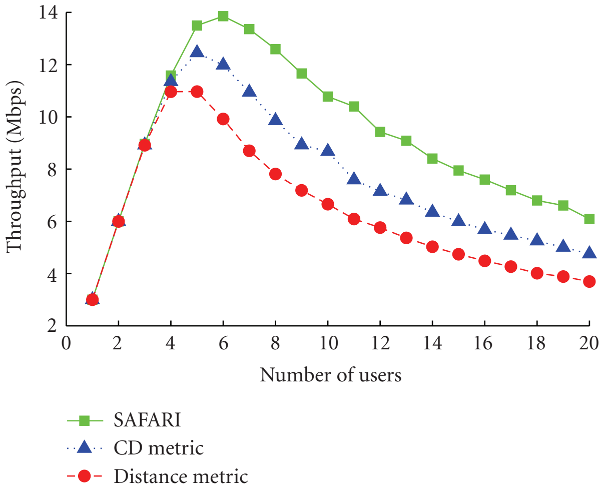


Figure 15: Throughput versus number of users with 12 channels and $\lambda = 1$.

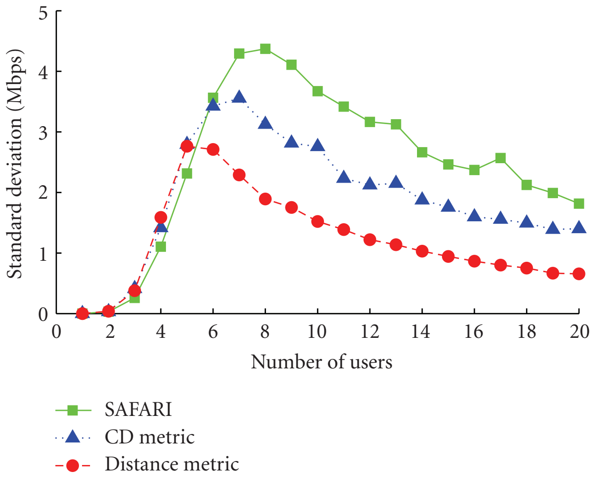


Figure 17: Standard deviation versus number of users with 12 channels and $\lambda = 1$.

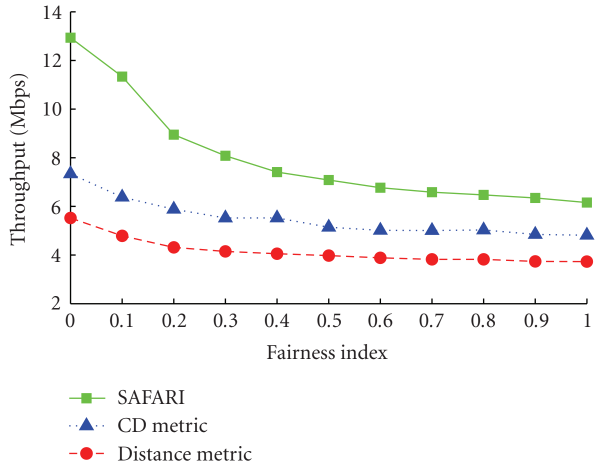


Figure 16: Throughput versus fairness index with 12 channels and 20 users.

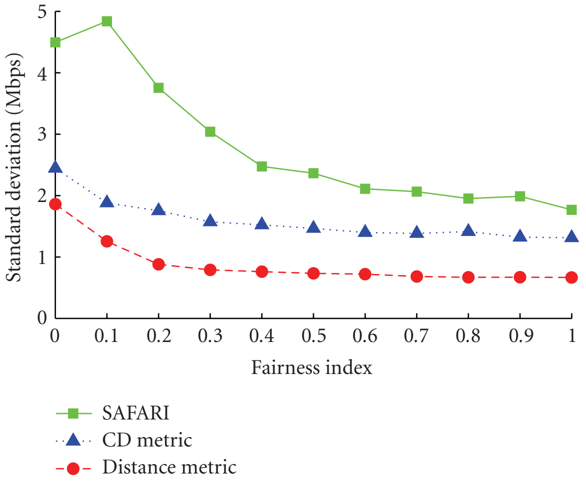


Figure 18: Standard deviation versus fairness index with 12 channels and 20 users.

curves. The fluctuation illustrates the significant effect on user positions within the network, even averaging over 750 random drop of users (Table 2) it cannot completely average the achieved throughput.

Figure 19 shows that SAFARI has the lowest average BCD, and CD *metric* increases gap to *distance metric*. This result reassures the benefit of CD as a routing metric compared to simple distance-based metric. Comparing Figures 14 and 19, it is clear that increasing the number of channels from one to 12 decreases the average BCD to one third with SAFARI and *CD metric*. *Distance metric* case does not decrease its average BCD as much as the others.

5. Conclusions and Summary

The simulation results show that the proposed routing algorithm SAFARI outperforms CD and distance-based routing algorithms in terms of the increased network throughput and the number of admitted users. The performance gain comes mainly from the fact that users positions are taken into account instead of neglecting them, as in the CD and distance-based routing. The information of user position is exploited in the CA and in the selection of the best router to each user to attach to. The second factor that contributes to the performance gain is the routing order. By first routing the users in low CD regions (i.e., usually users far away from the gateways), shorter paths are obtained and which leads to less strict capacity constraint and fairness is easier to achieve. The CD metric is shown to be a suitable metric for WMN and its inherent capability to avoid congested areas in the network is a very useful quality. In addition, SAFARI's LP-based rate allocation leads to user rates that can be scheduled.

The performance gain comes with the cost of increased complexity, transmit power, and statistical variation of the achieved throughput. The increase in complexity can be remarkable when compared to a simple hop count-based routing with fixed rates. Factors effecting the complexity are the router selection procedure, LP-based rate allocation, and

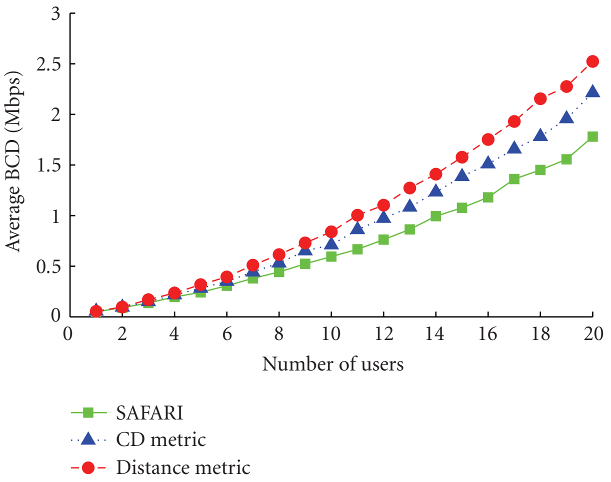


Figure 19: BCD versus number of users with 12 channels and $\lambda = 1$.

the recursive calling of FW's algorithm and CD estimate update in the final routing phase. The increased complexity and the need for more transmit power can be too much for some systems or users. Nevertheless, SAFARI can be solved in polynomial time. If the increased complexity of SAFARI is too much for a system, CD metric-based routing can be used with reasonable performance. An estimate of the CD of each link can be obtained by a centralized entity or by spectrum sensing at each node.

The scientific contribution of this work is the developed SAFARI algorithm. The novelty of SAFARI comes from the usage of the information of user positions in CA, router selection, and routing. Another new feature is the routing order selection that is based on the network congestion so that users in low-congestion areas are routed first. This routing order selection leads to higher throughput, mainly since users in low-congested areas are usually at the edge of a network and thus routing them first leads to shorter routing paths on the average.

Since our rate allocation is based on the one proposed in [8], the assigned rates can be feasibly scheduled. On a more widespread scope for future research, the overall feasibility and practicality of SAFARI needs to be investigated in more detail.

Acknowledgments

This work was supported by the Finnish Funding Agency for Technology and Innovation (Tekes), Nokia, Nokia Siemens Networks, Elektrobit, and the Centre for Wireless Communications.

References

[1] I. F. Akyildiz, X. Wang, and W. Wang, "Wireless mesh networks: a survey," *Computer Networks*, vol. 47, no. 4, pp. 445–487, 2005.

[2] B. Radunović and J.-Y. L. Boudec, "Rate performance objectives of multihop wireless networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 4, pp. 334–349, 2004.

[3] L. Dai, Y. Xue, B. Chang, and Y. Cui, "Throughput optimization routing under uncertain demand for wireless mesh networks," in *Proceedings of IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS '07)*, pp. 1–11, Pisa, Italy, October 2007.

[4] M. Malekesmaeili, M. Shiva, and M. Soltan, "Topology optimization for backbone wireless mesh networks," in *Proceedings of the 5th Annual Conference on Communication Networks and Services Research (CNSR '07)*, pp. 221–230, Fredericton, Canada, May 2007.

[5] K. Duffy, D. J. Leith, T. Li, and D. Malone, "Improving fairness in multi-hop mesh networks using 802.11e," in *Proceedings of 4th IEEE International Symposium on Modelling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt '06)*, pp. 1–8, Boston, Mass, USA, April 2006.

[6] E.-C. Park, D.-Y. Kim, C.-H. Choi, and J. So, "Improving quality of service and assuring fairness in WLAN access networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 4, pp. 337–350, 2007.

[7] A. Raniwala, P. De, S. Sharma, R. Krishnan, and T.-C. Chiueh, "End-to-end flow fairness over IEEE 802.11-based wireless mesh networks," in *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM '07)*, pp. 2361–2365, Anchorage, Alaska, USA, May 2007.

[8] V. S. A. Kumar, M. V. Marathe, S. Parthasarathy, and A. Srinivasan, "Algorithmic aspects of capacity in wireless networks," in *Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '05)*, pp. 133–144, Banff, Canada, June 2005.

[9] J. Jun and M. L. Sichitiu, "The nominal capacity of wireless mesh networks," *IEEE Wireless Communications*, vol. 10, no. 5, pp. 8–14, 2003.

[10] J. Jun, P. Paddabachagari, and M. Sichitiu, "Theoretical maximum throughput of IEEE 802.11 and its applications," in *Proceedings of the 2nd International Symposium on Network Computing and Applications (NCA '03)*, pp. 249–256, Cambridge, Mass, USA, April 2003.

[11] Y. Yang, J. Wang, and R. Kraves, "Designing routing metrics for mesh networks," in *Proceedings of the 1st IEEE Workshop on Wireless Mesh Networks (WiMesh '05)*, Santa Clara, Calif, USA, September 2005.

[12] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc on-demand distance vector (AODV) routing," Tech. Rep. RFC 3561, Internet Engineering Task Force, Fremont, Calif, USA, 2003.

[13] R. Draves, J. Padhye, and B. Zill, "Routing in multi-radio, multi-hop wireless mesh networks," in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking (MOBICOM '04)*, pp. 114–128, Philadelphia, Pa, USA, September-October 2004.

[14] A. Raniwala, K. Gopalan, and T. C. Chiueh, "Centralized channel assignment and routing algorithms for multi-channel wireless mesh networks," *ACM Mobile Computing and Communications Review*, vol. 8, no. 2, pp. 50–65, 2004.

[15] H. Skalli, S. Ghosh, S. K. Das, L. Lenzini, and M. Conti, "Channel assignment strategies for multiradio wireless mesh networks: issues and solutions," *IEEE Communications Magazine*, vol. 45, no. 11, pp. 86–93, 2007.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 102064, 17 pages
doi:10.1155/2009/102064

Research Article

On Throughput-Fairness Tradeoff in Virtual MIMO Systems with Limited Feedback

Alexis A. Dowhuszko,[1] Graciela Corral-Briones,[1] Jyri Hämäläinen,[2] and Risto Wichman[3]

[1] *Digital Communications Research Laboratory, National University of Cordoba (UNC) - CONICET, Avenida Velez Sarsfield 1611, X5016GCA Cordoba, Argentina*
[2] *Department of Communications and Networking, Helsinki University of Technology (TKK), P.O. Box 3000, FIN-02015 TKK, Finland*
[3] *Department of Signal Processing and Acoustics, Helsinki University of Technology (TKK) - Smart and Novel Radios (SMARAD) Centre of Excellence, P.O. Box 3000, FIN-02015 TKK, Finland*

Correspondence should be addressed to Alexis A. Dowhuszko, adowhuszko@efn.unc.edu.ar

Received 1 July 2008; Revised 10 December 2008; Accepted 24 January 2009

Recommended by Zhu Han

We investigate the performance of channel-aware scheduling algorithms designed for the downlink of a wireless communication system. Our study focuses on a two-transmit antenna cellular system, where the base station can only rely on quantized versions of channel state information to carry out scheduling decisions. The motivation is to study the interaction between throughput and fairness of practical spatial multiplexing schemes when implemented using existing physical layer signaling, such as the one that exists in current wideband code division multiple access downlink. Virtual MIMO system selects at each time instant a pair of users that report orthogonal (quantized) channels. Closed-form expressions for the achievable sum-rate of three different channel-aware scheduling rules are presented using an analytical framework that is derived in this work. Our analysis reveals that simple scheduling procedures allow to reap a large fraction (in the order of 80%) of the sum-rate performance that greedy scheduling provides. This overall throughput performance is obtained without affecting considerably the optimal short-term fairness behavior that the end users would perceive.

1. Introduction

The deployment of multiple *transmit* (Tx) antennas at the *base station* (BS) has emerged as an effective way for improving the overall throughput in a wireless communication system. This is because multiuser *multiple-input multiple-output* (MIMO) downlink systems offer multiple channel directions to send independent information streams to multiple users simultaneously within the same resource block, capitalizing the so-called spatial multiplexing gain [1]. However, resource allocation in multiuser MIMO systems is not a trivial task because users should be selected taking into account not only their spatial compatibility, but also their individual channel strengths [2]. The construction of optimal schedulers in terms of throughput makes imperative the investigation of the sum-rate upper bound that can be achieved in this situation. However, such a myopic approach is not enough for real-life wireless applications if the scheduler does not share common channel resources fairly among all the participating users as well. Based on this, intensive research has been carried out in the past few years to study the interaction between these two conflicting goals and design fair channel-aware scheduling rules for delay-constrained data connections. In this context, this work provides an analytical framework for quantifying the throughput gain of different multiuser scheduling strategies in MIMO downlink channel with different types of partial *channel state information* (CSI) in the transmitter. The use of a well-known fairness performance metric, Jain's fairness index [3], is also suggested as a simple way to evaluate the short-term fairness that is traded off at the expense of additional throughput gain.

Recent theoretical results show that the optimal transmission scheme in an MIMO downlink channel is *dirty paper coding* (DPC) [4], but it faces serious implementation issues in practical systems due to its high complexity, especially when the number of participating users is large. *Linear beamforming* (LBF) is a suboptimal strategy in which each user stream is multiplied independently by a beamforming weighting vector for transmission through multiple antennas. Despite its reduced complexity, LBF achieves a large portion of DPC capacity and exhibits the best tradeoff between complexity and performance [5]. In particular, a simpler strategy based on *zero-forcing beamforming* (ZFBF) has been shown to be optimal in terms of sum capacity in the limit of a large number of users [6]. All these capacity results rely on the assumption that perfect CSI is available at the transmitter. However, this condition is hard to satisfy in practical systems, particularly when *frequency-division duplex* (FDD) is implemented because in practice mobiles report their channel estimates to the BS via a rate-constrained reverse channel.

One of the simplest approaches to reduce feedback overhead involves each user quantizing his instantaneous vector channel according to a finite collection of vectors (beamformer codebook) that is maintained at both extremes of the link [7]. After selecting the optimal quantization vector, receiver feeds back the corresponding codeword index through a B-bit (per user) reverse channel at the beginning of each transmission block. This feedback is used to capture the *channel direction information* (CDI), and was first considered for point-to-point MIMO channels in [8, 9]. System sum-rate capacity with only CDI is bounded as the number of users increases because *channel quality information* (CQI) is not available in transmission to exploit multiuser diversity and obtain the double-logarithmic growth in system throughput with the number of users [10]. Based on this, both CDI and CQI feedbacks are necessary if we want to achieve both multiplexing and multiuser diversity gains at the same time. As expected, we later show that CQI should be the channel magnitude in low Tx power regime, while it should be proportional to the *signal-to-interference* power *ratio* (SIR) when Tx power is high.

Limited feedback techniques have already been considered in 3G cellular standards, where two antenna schemes have been emphasized so far due to implementation constraints. In *3G Partnership Project* (3GPP), *closed-loop* (CL) *transmit-diversity* (TD) techniques come in two classes: quantized phase information (mode 1) and direct channel quantization (mode 2) [11]. The quantized phase algorithm uses a fixed number of bits to quantize phase angles to perform equal gain beamforming at the transmitter. The direct channel quantization allocates a fixed number of bits for the gain and phase of each channel entry separately, as opposed to more sophisticated vector quantization techniques that quantize gain and phase jointly. Our motivation is to study the performance when combining channel-aware scheduling rules with ZFBF prefiltering in case of practical (commercial) beamformer codebook designs. Note that this principle is equivalent to virtual MIMO concept for the uplink of a *time-division multiple-access* (TDMA)-based cellular system, where many users with only one Tx antenna transmit independently to the BS on the same resource block. Our analysis reveals that in the presence of 3GPP physical layer signaling, the additional multiuser diversity gain that is obtained at the cost of relegating fairness considerations over short time scales is quite important. However, it was also observed that the implementation of simpler scheduling procedures, such as the one presented in [12], offers a good balance between implementation complexity, short-term fairness, and system sum-rate performance. Although we concentrate on the two CL techniques in the FDD mode of the *wideband code division multiple access* (W-CDMA) downlink, a similar procedure can be used to extend the analysis to other FDD MIMO systems with limited feedback.

The rest of the paper is organized as follows: Section 2 introduces the system model, presents the feedback model for CDI and CQI, and describes the scheduling strategies and spatial prefiltering technique that will be analyzed. Section 3 studies the statistics of desired signal energy and mutual interference, proposes a probability distribution approximation for them, and derives an accurate closed-form expression for the achievable rate per user when BS simultaneously transmits to a pair of semiorthogonal users without exploiting multiuser diversity. Section 4 extends the analysis when channel norm CQI or SIR CQI is available in transmission to perform user selection. Section 5 introduces the criterion that is used to carry out the fairness study of the different schemes over short-time scales. Section 6 analyzes the performance of the different scheduling strategies, quantifying the different tradeoffs between throughput and fairness that they provide. Finally, conclusions are drawn in Section 7.

2. System Model

The system consists of a single BS with $M_t = 2$ Tx antennas and K active *user equipments* (UEs) with single-element antennas. In case of flat fading and rich scattering, the channel gain from a Tx antenna t to a UE k is described by a zero-mean circularly symmetric complex Gaussian *random variable* (RV) $h_{k,t}$, for $t = 1, \ldots, M_t$ and $k = 1, \ldots, K$. We assume that all UEs are homogeneous and experience independent fading, and that they have a low-rate, reliable, and delay-free feedback channel to the BS.

A block fading channel model is employed, that is, channels remain constant during each block of transmitted symbols, and channels between temporally separate transmission blocks are independent. Transmitted codewords of fixed rate span multiple independent fading blocks; therefore, when the number of blocks is large, the system is able to achieve nonzero ergodic capacity. Note that instead of fixed rate codes, *high-speed downlink packet access* (HSDPA) [13] exploits variable rate coding, where the BS selects modulation and coding scheme according to CQI reports. However, it has been shown in [14] that both fixed rate and variable rate coding strategies achieve the same capacity when channel variation satisfies a compatibility assumption meaning that the input distribution that maximizes mutual information is the same regardless of the channel state. We note that block fading channels with constant Tx power satisfy this compatibility assumption.

In our system model, the signal received by a user k is

$$r_k = \mathbf{h}_k \mathbf{x} + n_k, \quad k = 1, \ldots, K, \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^{M_t \times 1}$ is the transmitted vector signal from the BS antennas containing information symbols of selected users, $\mathbf{h}_k \in \mathbb{C}^{1 \times M_t}$ is the channel gain vector, and n_k is zero-mean complex additive white Gaussian noise with power N_0. In order to facilitate the analysis, the channel and noise entries are normalized to have unit variance. The average power constraint of the input signal implies that $\mathbb{E}\{\mathbf{x}^\dagger \mathbf{x}\} \leq P$, where P is the total Tx energy per channel use, $(\cdot)^\dagger$ denotes Hermitian transpose, and $\mathbb{E}\{\cdot\}$ denotes expectation. As with HSDPA, we do not consider the possibility of employing fast power control mechanisms at the BS; thus, P remains fixed. Since the noise has unitary variance, P takes on the meaning of total Tx *signal-to-noise* power *ratio* (SNR).

As the number of participating users grows, the introduction of user selection mechanisms enables the BS to choose up to M_t out of K mobiles to use the channel. In this context, $\mathcal{S}$ is the set that contains the indices of selected UEs at any given time. Transmit vector $\mathbf{x}$ is related to information symbols $\{s_i : i \in \mathcal{S}\}$ via linear beamforming; that is, $\mathbf{x} = \sum_{i \in \mathcal{S}} \mathbf{w}_i s_i$, where Tx weights $\{\mathbf{w}_i : i \in \mathcal{S}\}$ are appropriately selected according to BS spatial prefiltering technique and quantized versions of channel states $\{\widehat{\mathbf{h}}_i : i \in \mathcal{S}\}$ available in transmission. Based on this, rewriting (1) in a more convenient way, it is possible to observe that received signal

$$r_k = \underbrace{(\mathbf{h}_k \mathbf{w}_k) s_k}_{d_k\text{: Desired Signal}} + \underbrace{\sum_{l \in \mathcal{S}, l \neq k} (\mathbf{h}_k \mathbf{w}_l) s_l}_{q_k\text{: Mutual Interference}} + \underbrace{n_k}_{\text{Noise}}, \quad k \in \mathcal{S} \tag{2}$$

is actually composed by three different parts. Active user set $\mathcal{S}$ is chosen according to the implemented scheduling policy and will ideally try to provide a reasonable tradeoff between throughput and fairness according to quality-of-service requirements of the supported application.

2.1. Feedback Model for Channel Direction Information and Channel Quality Information. The feedback scheme assumes that each UE has perfect CSI in reception, and each of them quantizes the normalized channel vector $\widetilde{\mathbf{h}}_k = \mathbf{h}_k / \|\mathbf{h}_k\|$ to a unit norm M_t-dimensional vector $\widehat{\mathbf{h}}_k$, which is selected from a common quantization codebook $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_{2^B}\}$, where B refers to the number of reported CDI bits per mobile user. Each UE quantizes its channel vector to the codeword that forms the minimum angle to it, or equivalently

$$\widehat{\mathbf{h}}_k = \arg\max_{\mathbf{c}_i \in \mathcal{C}} \cos^2 \angle(\widetilde{\mathbf{h}}_k, \mathbf{c}_i) = \arg\max_{\mathbf{c}_i \in \mathcal{C}} |\widetilde{\mathbf{h}}_k \mathbf{c}_i^\dagger|^2. \tag{3}$$

Note that only the index i needs to be reported because quantization codebook $\mathcal{C}$ is known to both transmitter and mobile users a priori.

Narula et al. noticed in [7] that CL beamforming is invariant to the channel being multiplied by $e^{j\vartheta}$ for any phase angle ϑ. Therefore, it can be assumed that the first coefficient $\widehat{h}_{k,1}$ of channel vector $\widehat{\mathbf{h}}_k$ is real, and without loss of generality, CDI feedback solution can be fully characterized by $(M_t - 1)$ complex coefficients. More precisely, when focusing on 2 Tx antennas, CDI feedback is composed by a single complex coefficient $\widehat{h}_{k,2} = \widehat{\alpha}_k e^{-j\widehat{\phi}_k}$, where $\widehat{\alpha}_k$ and $\widehat{\phi}_k$ are quantized magnitude and phase of the weight applied in the second Tx antenna. Specifically, in W-CDMA mode 1 CL TD solution, only phase information of the feedback weight is quantized with 2 bits (i.e., the magnitude remains constant), while in mode 2 both magnitude and phase are independently quantized with 1 bit and 3 bits, respectively [11]. In both cases, uniform quantization is applied for phase information. In mode 2, the stronger channel receives 6 dB more power than the weaker one. Even though CL TD mode 2 was later removed from the specification with the motivation of simplifying the 3GPP standard, we also consider this feedback mode in order to quantify performance gain when the amount of reported CDI grows.

In addition to the CDI, each user feeds back a CQI that is used at BS for scheduling purposes. In this work, we consider two different definitions for CQI:

$$\begin{aligned} Q(\mathbf{h}_k) &= \|\mathbf{h}_k\|^2, \\ \breve{Q}(\mathbf{h}_k) &= \frac{\mathbb{E}\{|d_k|^2\}}{\mathbb{E}\{|q_k|^2\}} \\ &= \frac{|\mathbf{h}_k \mathbf{w}_k|^2}{\sum_{l \in \mathcal{S}, l \neq k} |\mathbf{h}_k \mathbf{w}_l|^2}, \end{aligned} \tag{4}$$

that are proportional to the selected users channel norm and SIR, respectively. Channel norm CQI is suitable for noise-limited communication systems, such as those that employ TDMA schemes or spatial multiplexing strategies with imperfect CSI at the transmitter in low-SNR regions. Note that this is the CQI definition that W-CDMA specification contains. On the other hand, SIR CQI does a better job in the presence of interference-limited systems, such as those that implement spatial multiplexing policies with imperfect CSI at the transmitter in high Tx power regimes. We assume that the CQI is reported to the BS without quantization; however, previous works have already observed that the number of bits for CQI quantization can be kept relatively low with an appropriate CQI feedback design [15, 16].

2.2. Scheduling Strategies. Seeking a reasonable balance between throughput and fairness, a scheduler achieving *proportional fairness* (PF) criterion was first proposed in [17]. This PF scheduler selects at each time slot the user with the highest transmission rate relative to its current average throughput. In the classical version, the average rate is tracked by an exponential window with time constant t_c. The proper selection of parameter t_c allows to control the maximum starvation period (i.e., the maximum time between two successive service offerings) for the packet scheduling scheme. The combination of PF scheduler along with ZFBF precoding has been proposed in [6] as a natural alternative to provide an equal share of common resources among users in a *space-division multiple-access* (SDMA) system with multiuser diversity. Even though important results on tradeoffs between throughput and fairness have been reported, no closed-form formula for sum-rate performance has been provided since PF algorithm is hard to analyze. Keeping this in mind, we will study the behavior of simpler schemes that will allow us to derive closed-form expressions for the performance of ZFBF-PF when t_c is tuned to maximize system throughput or fairness, respectively.

Optimal user set solution in terms of throughput demands an exhaustive search over all possible groups with up to M_t out of K members at a time, taking into account their spatial compatibility and CQI. In order to avoid this search in the presence of many users, suboptimal *semiorthogonal user selection* (SUS) procedure was used instead as an accurate approximation for ZFBF-PF upper bound, see [6, 10]. On the other hand, we use pure *orthogonal round-robin* (ORR) scheme proposed in [12] as a simple lower bound for ZFBF-PF when CSI is not taken into account to perform scheduling. We note that the idea behind pure ORR scheme is simple. Select both, primary and secondary users in round-robin (RR) taking into account their spatial compatibility. Since primary user is selected according to its waiting time at the BS, all active users will have an explicit guarantee to be scheduled at least one time in a round.

2.3. Zero-Forcing Beamforming Scheme. Let $\widehat{\mathbf{H}}(\mathcal{S}) = [\widehat{\mathbf{h}}_{\pi(1)}^T, \ldots, \widehat{\mathbf{h}}_{\pi(|\mathcal{S}|)}^T]^T$ be the concatenated unit norm quantized vectors of selected users in set $\mathcal{S} = \{\pi(1), \ldots, \pi(|\mathcal{S}|)\}$, where $(\cdot)^T$ denotes vector transpose. The ZFBF matrix is given by the pseudoinverse of the channel as

$$\mathbf{W}(\mathcal{S}) = \widehat{\mathbf{H}}(\mathcal{S})^{+} = \widehat{\mathbf{H}}(\mathcal{S})^{\dagger}\left[\widehat{\mathbf{H}}(\mathcal{S})\widehat{\mathbf{H}}(\mathcal{S})^{\dagger}\right]^{-1}, \tag{5}$$

where Tx weight $\mathbf{w}_{\pi(i)} \in \mathbb{C}^{M_t \times 1}$, obtained by normalizing the ith column of $\mathbf{W}$, represents Tx weight for user $\pi(i)$. In ZFBF, Tx weights satisfy orthogonality criterion in transmission; that is, $\widehat{\mathbf{h}}_j \mathbf{w}_i = 0$ for $i, j \in \mathcal{S}$, $j \neq i$. Even though ZFBF is not the optimal choice among all possible LBF schemes, we focus on it because its analytical simplicity enables to obtain closed-form expressions for achievable sum-rate that are asymptotically optimal as the Tx power grows. Note that when the number of users K is large and the codebook contains orthogonal codewords (such as W-CDMA CL modes), $\mathbf{W}(\mathcal{S}) = \widehat{\mathbf{H}}(\mathcal{S})^{\dagger}$.

3. Achievable Rate per Beam without Exploiting Multiuser Diversity

This section derives a closed-form expression for the achievable rate per user (beam) when BS simultaneously transmits to a pair of spatial-compatible UEs (i.e., semiorthogonal in terms of their quantized CDI) without considering CQI reports to perform scheduling. The derived expression is used in Section 6 to quantify the actual system throughput of pure ORR and hybrid ORR proposals as well.

3.1. Probability Distributions of Desired Signal and Mutual Interference. Following the model (2) we construct two RVs,

$$X_k = \left|\mathbf{h}_k \mathbf{w}_k\right|^2, \qquad Y_k = \left|\mathbf{h}_k \mathbf{w}_l\right|^2. \tag{6}$$

Here, the first RV gives the desired signal energy while the second RV represents the contribution of mutual interference due to simultaneous transmission. In these equations, $\mathbf{w}_k = \widehat{\mathbf{h}}_k^{\dagger}$ is the Tx weight vector that maximizes received energy for user k (i.e., the best Tx weight), while $\mathbf{w}_l = \widehat{\mathbf{h}}_l^{\dagger}$ is the Tx weight vector that minimizes received energy of the same user (i.e., worst Tx weight). In the coming sections, we deduce usable formulae for achievable rates in different cases based on modeling the distributions of X_k and Y_k, denoted by $f_x(x)$ and $f_y(y)$, by chi-square (χ^2) distribution approximations. To justify this claim, we use Nakagami's distribution [18]

$$f(y) = \frac{1}{\Gamma(\mathcal{F})}\left(\frac{\mathcal{F}}{\mathcal{G}}\right)^{\mathcal{F}} y^{\mathcal{F}-1} e^{-(\mathcal{F}/\mathcal{G})y} \tag{7}$$

as an accurate approximation to model the signal energy behavior of our RVs, where $\Gamma(\cdot)$ denotes the Gamma function and

$$\mathcal{G} = \mathbb{E}\{y\}, \qquad \mathcal{F} = \frac{\mathcal{G}^2}{\mathbb{E}\{(y-\mathcal{G})^2\}} \tag{8}$$

represent the so-called SNR gain and fading figure, respectively. Note that the SNR gain provides information on the coherent combining gain, whereas the fading figure indicates the degree of signal variation. If $\mathcal{F} \in \mathbb{N}$, then $f(y)$ is the normalized χ^2-distribution with $r = 2\mathcal{F}$ degrees of freedom. If we select Tx weights randomly, then there is neither coherent combining power gain (i.e., $\mathcal{G} = 1$), nor Tx diversity gain (i.e., $\mathcal{F} = 1$). On the other hand, in the presence of unquantized Tx weights, full Tx beamforming gain is achieved (i.e., $\mathcal{G} = 2$ and $\mathcal{F} = 2$).

According to the analysis presented in Appendix A.2, the first-order corrected version when approximating $f_x(x)$ by an χ^2-distribution with 4 degrees of freedom is given by

$$\begin{gathered} f_x(x) \approx \left(\frac{2}{\mathcal{G}_x}\right)^2 x e^{-(2/\mathcal{G}_x)x}\left(b_2 x^2 + b_1 x + b_0\right), \\ b_0 = 2\left(\frac{\mathbb{E}\{X_k^2\}}{\mathcal{G}_x^2} - 1\right), \quad b_1 = \frac{2(1-b_0)}{\mathcal{G}_x}, \quad b_2 = -\frac{2(1-b_0)}{(3\mathcal{G}_x^2)}. \end{gathered} \tag{9}$$

Similarly, Appendix A.1 derives the first-order corrected version when $f_y(y)$ is approximated by an exponential distribution (i.e., χ^2-distribution with 2 degrees of freedom). In this case,

$$\begin{gathered} f_y(y) \approx \left(\frac{1}{\mathcal{G}_y}\right) e^{-(1/\mathcal{G}_y)y}\left(a_2 y^2 + a_1 y + a_0\right), \\ a_0 = \frac{\mathbb{E}\{Y_k^2\}}{(2\mathcal{G}_y^2)}, \quad a_1 = \frac{2(1-a_0)}{\mathcal{G}_y}, \quad a_2 = -\frac{(1-a_0)}{(2\mathcal{G}_y^2)}. \end{gathered} \tag{10}$$

3.2. Probability Distribution Approximations with Deterministic Codebook Design. Let us assume that CDI codebook is selected using a deterministic design with fixed number of bits to quantize the gain and phase portions of each channel independently, see Section 2.1. Then, each weight vector admits an orthogonal counterpart. Hence, for any weight $\mathbf{w}_k$, there exists a weight $\mathbf{w}_l$ such that $\mathbf{w}_k^\dagger \mathbf{w}_l = 0$. Note that while beams are orthogonal in transmission, the orthogonality is lost in the receiver because Tx weights are selected based on quantized versions of actual channel direction $\widetilde{\mathbf{h}}_k$. After working out (6), we arrive at

$$\begin{aligned} X_k &= |w_{1,k}|^2|h_{k,1}|^2 + |w_{2,k}|^2|h_{k,2}|^2 \\ &\quad + 2|w_{1,k}||w_{2,k}||h_{k,1}||h_{k,2}|\cos\varphi_k, \\ Y_k &= |w_{2,k}|^2|h_{k,1}|^2 + |w_{1,k}|^2|h_{k,2}|^2 \\ &\quad - 2|w_{1,k}||w_{2,k}||h_{k,1}||h_{k,2}|\cos\varphi_k, \end{aligned} \tag{11}$$

where $\varphi_k = \angle(h_{k,1}) - \angle(h_{k,2}) + \widehat{\phi}_k$ is the phase difference between both channel gains after applying the corresponding Tx weight vector. Let us denote by Z_k the sum of X_k and Y_k. Then, we find by (11) that

$$Z_k = X_k + Y_k = |h_{k,1}|^2 + |h_{k,2}|^2 \tag{12}$$

follows an χ^2-distribution with 4 degrees of freedom and mean $\mathbb{E}\{Z_k\} = 2$.

The SNR gains and fading figures for both W-CDMA CL TD modes are derived analytically in Appendix B, see Table 1. According to these results, fading figures $\mathcal{F}_x \approx 2$ and $\mathcal{F}_y \approx 1$ in the both CL TD methods. This indicates that the shapes of distributions $f_x(x)$ and $f_y(y)$ are close to χ^2-distribution with $r_x = 4$ and $r_y = 2$ degrees of freedom, respectively. (A similar procedure can be used to compute both SNR gains and fading figures when the BS is equipped with more than two Tx antennas, with the only difference that mutual interference would become the sum of $(M_t - 1)$ i.i.d. RVs in this situation.) It has already been observed in [12] that these approximations greatly simplify the computation of closed-form expressions for achievable sum-rate. However, in order to have a better distribution fitting, we propose to use the first-order correction for χ^2-distribution approximation, as detailed in Appendix A. Coefficients a_i and b_i for both CL TD feedback modes have been derived analytically based on the first two raw moments of RVs X_k and Y_k. These moments are computed in Appendix B, and coefficients are presented in Table 2.

3.3. Achievable Rate for Spatial Multiplexing with CDI and No CQI. When BS applies SDMA to simultaneously serve a pair of UEs that report orthogonal CDI codewords (no CQI), the achievable rate per user when Tx power is evenly divided between both users (i.e., $P/2$) is

$$C_k(P) = \mathbb{E}\left\{\log_2\left(1 + \frac{(1/2)PX_k}{(1/2)PY_k + N_0}\right)\right\} \tag{13}$$

$$= \log_2(e)\left[\mathbb{E}\{\log_e(Z_k + N_0')\} - \mathbb{E}\{\log_e(Y_k + N_0')\}\right], \tag{14}$$

Table 1: SNR gains and fading figures in case of CL TD mode 1 and 2.

(a) SNR gains

	Mode 1	Mode 2
$\mathcal{G}_x$	$1 + \sqrt{\frac{1}{2}}$	$1.3 + 1.6\sin\frac{\pi}{8}$
$\mathcal{G}_y$	$1 - \sqrt{\frac{1}{2}}$	$0.7 - 1.6\sin\frac{\pi}{8}$

(b) Fading figures

	Mode 1	Mode 2
$\mathcal{F}_x$	1.9104	1.9919
$\mathcal{F}_y$	0.7714	0.6816

Table 2: Coefficients for first-order correction χ^2-approximation with CL TD mode 1 and 2.

(a) $f_x(x)$ approximation

	Mode 1	Mode 2
b_0	1.0469	1.0041
b_1	−0.0549	−0.0043
b_2	0.0107	0.0007

(b) $f_y(y)$ approximation

	Mode 1	Mode 2
a_0	1.1481	1.2336
a_1	−1.0116	−5.3265
a_2	0.8634	15.1827

where $N_0' = 2N_0/P$. Based on the fact that Z_k is χ^2 distributed with 4 degrees of freedom ($\mathcal{G}_z = 2$),

$$\mathbb{E}\{\log_e(Z_k + N_0')\} = \int_0^\infty \log_e(z + N_0')ze^{-z}\,dz. \tag{15}$$

At this stage, we use the relation derived in Appendix C

$$\int_0^\infty \log_e(\gamma + c)\beta(\beta\gamma)^n e^{-\beta\gamma}\,d\gamma = n!\left[\log_e(c) + e^{\beta c}\sum_{i=0}^{n} E_{i+1}(\beta c)\right], \tag{16}$$

where $E_n(z)$ represents the exponential integral function of order n, see (5.1.4) of [19]. After combining (15) and (16), we obtain

$$\mathbb{E}\{\log_e(Z_k + N_0')\} = \log_e(N_0') + e^{N_0'}\left[E_1(N_0') + E_2(N_0')\right]. \tag{17}$$

To compute the latter expectation in (14), we use approximation (10) and formula (16), that is,

$$\begin{aligned}
&\mathbb{E}\{\log_e(Y_k + N_0')\}\\
&\approx \int_0^\infty \log_e(y + N_0')\left(\frac{1}{\mathcal{G}_y}\right)e^{-(1/\mathcal{G}_y)y}(a_2y^2 + a_1y + a_0)\,dy\\
&= (2a_2\mathcal{G}_y^2 + a_1\mathcal{G}_y + a_0)\left[\log_e(N_0') + e^{N_0'/\mathcal{G}_y}E_1\left(\frac{N_0'}{\mathcal{G}_y}\right)\right]\\
&\quad + (2a_2\mathcal{G}_y^2 + a_1\mathcal{G}_y)e^{N_0'/\mathcal{G}_y}E_2\left(\frac{N_0'}{\mathcal{G}_y}\right) + (2a_2\mathcal{G}_y^2)e^{N_0'/\mathcal{G}_y}E_3\left(\frac{N_0'}{\mathcal{G}_y}\right),
\end{aligned} \tag{18}$$

where coefficients a_i and SNR gain $\mathcal{G}_y$ depend on the number of bits assigned to report CDI to the transmitter (see Tables 2 and 1). Replacing (17) and (18) in (14), final approximation to estimate the achievable rate per beam with only CDI feedback is obtained.

3.3.1. Low-SNR Regime. Assume that Tx power is small. Then, after applying Taylor series expansion in (13), we find that

$$C_k(P) \approx \log_2(e)\frac{1}{2}\frac{P}{N_0}\mathbb{E}\{X_k\}, \quad P \ll N_0. \tag{19}$$

Hence, achievable rate of an individual user decays linearly with Tx power in a low-SNR regime.

3.3.2. High-SNR Regime. As expected, proposed SDMA scheme admits an interference-limited behavior in high Tx power regime since reported CDI is not perfect. A formula for this upper bound is obtained from expression of $C_k(P)$, given by (14) combined with (17) and (18), as follows. First, we write all exponential integral functions of order $n > 1$ in terms of $E_1(z)$ using recursive relationship (5.1.14) of [19], that is,

$$E_{n+1}(z) = \frac{1}{n}\left[e^{-z} - zE_n(z)\right], \quad n = 1, 2, \ldots. \tag{20}$$

Then, we let P grow and apply approximation for $E_1(z)$ that is valid for small z values, that is,

$$E_1(z) \approx -\epsilon_0 - \log_e(z) \quad z \longrightarrow 0. \tag{21}$$

Here, $\epsilon_0 = 0.5772\ldots$ is Eulers' constant. After these preparations, we find that all terms containing logarithm of P vanish, and we are able to compute the final limit when $P \to \infty$. It turns out that asymptotic formula admits expression in terms of SNR gain and fading figure as

$$\lim_{P\to\infty} C_k(P) \approx \log_2(e)\left[\frac{3\mathcal{F}_y + 1}{4\mathcal{F}_y} - \log_e(\mathcal{G}_y)\right]. \tag{22}$$

According to this formula, asymptotic upper bounds are equal to 3.3211 and 5.1223 bps/Hz for CL TD modes 1 and 2, respectively.

3.4. Achievable Rate for Single User Transmission with CDI and No CQI. For comparison purposes, we also introduce a single user approach (or TDMA scheme), where all Tx power is assigned to a single user in RR fashion. In this situation, achievable rate becomes

$$\begin{aligned}
C^{\mathrm{TDMA}}(P) &= \mathbb{E}\left\{\log_2\left(1 + \frac{PX_k}{N_0}\right)\right\}\\
&= \log_2(e)\left[\mathbb{E}\left\{\log_e\left(X_k + \frac{N_0}{P}\right)\right\} - \log_e\left(\frac{N_0}{P}\right)\right].
\end{aligned} \tag{23}$$

Here, we could use first-order corrected distribution $f_x(x)$ according to (9), but from Table 2 we find that $b_0 \approx 1$ and $b_1, b_2 \approx 0$ for both CL TD feedback modes. Hence,

$$\mathbb{E}\left\{\log_e\left(X_k + \frac{N_0}{P}\right)\right\} \approx \int_0^\infty \log_e\left(x + \frac{N_0}{P}\right)\left(\frac{2}{\mathcal{G}_x}\right)^2 xe^{-(2/\mathcal{G}_x)x}\,dx \tag{24}$$

and we can apply relation (16) to derive the final closed-form expression

$$C^{\mathrm{TDMA}}(P) \approx \log_2(e)e^{(2N_0)/(P\mathcal{G}_x)}\left[E_1\left(\frac{2N_0}{P\mathcal{G}_x}\right) + E_2\left(\frac{2N_0}{P\mathcal{G}_x}\right)\right], \tag{25}$$

where SNR gain $\mathcal{G}_x$ depends on the number of bits used for CDI quantization, see Table 1.

3.4.1. Low-SNR Regime. Applying Taylor series expansion in (23), we arrive to an approximation that resembles the one presented in (19). Based on this, it is possible to conclude that achievable rate admit linear dependence on Tx power when SNR is low.

3.4.2. High-SNR Regime. Rewriting $E_2(\cdot)$ in (25) using recursive formula (20) and considering approximation (21), we see that

$$\begin{aligned}
C^{\mathrm{TDMA}}(P) &\approx \log_2(e)\left[1 - \epsilon_0 - \log_e\left(\frac{2N_0}{P\mathcal{G}_x}\right)\right]\\
&\approx \log_2\left(\frac{P}{N_0}\right) + \log_2(\mathcal{G}_x), \quad P \gg N_0.
\end{aligned} \tag{26}$$

Thus, achievable rate increases logarithmically with the Tx power when SNR is high. As expected, the use of CL TD provides an additional logarithmic SNR gain in this case.

4. Achievable Rate per Beam When Exploiting Multiuser Diversity

When the number of active users is large, there exist with high probability more than one user reporting any given CDI codeword. In this situation, the best strategy in terms of throughput is to select the UE with the best CQI among all users with identical CDI. In this section, we extend the previous analysis to a scenario where the BS exploits users CQI to reap multiuser diversity gain. Derived expressions are used in Section 6 to quantify the actual system throughput of ZFBF-SUS and hybrid ORR proposals as well.

4.1. Alternative RVs to Study the Effect of CQI Feedback. According to the model introduced in Section 2.1, each UE feeds back a quantized version of its CDI selected from a common codebook. Thus, we construct the following two RVs:

$$\begin{aligned}\widetilde{X}_k &= \left|\widetilde{\mathbf{h}}_k \mathbf{w}_k\right|^2 = \frac{1}{\|\mathbf{h}_k\|^2}\left|\mathbf{h}_k \mathbf{w}_k\right|^2 = \frac{X_k}{Z_k},\\ \widetilde{Y}_k &= \left|\widetilde{\mathbf{h}}_k \mathbf{w}_l\right|^2 = \frac{1}{\|\mathbf{h}_k\|^2}\left|\mathbf{h}_k \mathbf{w}_l\right|^2 = \frac{Y_k}{Z_k}.\end{aligned} \tag{27}$$

In our model both, channel direction $\widetilde{\mathbf{h}}_k$ and channel magnitude $\|\mathbf{h}_k\|$ are independent. Therefore, Tx weight vector $\mathbf{w}_k$ (and $\mathbf{w}_l$) does not depend on the channel strength. Thus, it is possible to conclude that both $\widetilde{X}_k$ and $\widetilde{Y}_k$ are independent with respect to Z_k. This property will be useful when deriving performance behavior for the different SDMA schedulers that will be analyzed.

4.2. Achievable Rate for Spatial Multiplexing with CDI and Channel Norm CQI. In this part, we analyze the effect of exploiting multiuser diversity when CQI reports are proportional to the channel norm (i.e., $Q(\mathbf{h}_k) = \|\mathbf{h}_k\|^2 = Z_k$). The procedure consists of selecting the user with the largest channel norm among all the users that report a given CDI codeword. The analysis that we apply here is similar to the one already employed in Section 3.3. However, the main difference is found in the modeling of the desired signal and mutual interference, that become $Z_{(n)}\widetilde{X}_k$ and $Z_{(n)}\widetilde{Y}_k$, respectively, with $Z_{(n)} = \max_{i=1,\dots,n} Z_i$. Based on these considerations, the achievable rate per beam when there are n users reporting the same CDI codeword is

$$C_{(n)}^{\text{Norm}}(P) = \mathbb{E}\left\{\log_2\left(1 + \frac{(1/2)PZ_{(n)}\widetilde{X}_k}{(1/2)PZ_{(n)}\widetilde{Y}_k + N_0}\right)\right\} \tag{28}$$

$$= \log_2(e)\left[\mathbb{E}\left\{\log_e\left(Z_{(n)} + N_0'\right)\right\} - \mathbb{E}\left\{\log_e\left(Z_{(n)}\widetilde{Y}_k + N_0'\right)\right\}\right], \tag{29}$$

where $Z_{(n)}$ is the largest order statistic of n *independent* and *identically distributed* (i.i.d.) χ^2 RVs with 4 degrees of freedom ($\mathcal{G}_z = 2$). Based on this, the *probability distribution function* (PDF) of $Z_{(n)}$ becomes

$$f_{z_{(n)}}(z) = \frac{\partial}{\partial z}\left[F_{z_{(n)}}(z)\right] = n\left[F_{z_{(n-1)}}(z)\right] f_z(z), \tag{30}$$

where $F_{z_{(n)}}(z)$ is the corresponding highest *cumulative distribution function* (CDF) given by

$$F_{z_{(n)}}(z) = \sum_{k=0}^{n}(-1)^k\binom{n}{k}e^{-zk}(1+z)^k. \tag{31}$$

At this stage, combining PDF expression (30) along with relation (16), we are now able to compute

$$\begin{aligned}\mathbb{E}\left\{\log_e\left(Z_{(n)} + N_0'\right)\right\} &= \int_0^{\infty}\log_e\left(z + N_0'\right) f_{z_{(n)}}(z)\,dz\\ &= \sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k}\sum_{l=1}^{k} L_{k,l}\left(N_0'\right),\\ L_{k,l}\left(N_0'\right) &= \left[\frac{(k-1)!}{(k-l)!}\frac{l}{k^l}\right]\\ &\quad\times\left\{\left[\sum_{m=1}^{l+1} E_m\left(kN_0'\right)e^{kN_0'}\right] + \log_e\left(N_0'\right)\right\}.\end{aligned} \tag{32}$$

We now compute the approximation for the distribution of the resulting mutual interference $Z_{(n)}\widetilde{Y}_k$. Firstly, based on the fact that RVs Z_k and $\widetilde{Y}_k$ are independent, we have that

$$\begin{aligned}\mathbb{E}\left\{Z_{(n)}\widetilde{Y}_k\right\} &= \frac{\mathcal{G}_y}{\mathcal{G}_z}\mathbb{E}\left\{Z_{(n)}\right\} = \frac{\mathcal{G}_y}{2}\mathbb{E}\left\{Z_{(n)}\right\},\\ \mathbb{E}\left\{Z_{(n)}^2\widetilde{Y}_k^2\right\} &= \frac{\mathbb{E}\left\{Y_k^2\right\}}{\mathbb{E}\left\{Z_k^2\right\}}\mathbb{E}\left\{Z_{(n)}^2\right\} = \frac{\mathbb{E}\left\{Y_k^2\right\}}{6}\mathbb{E}\left\{Z_{(n)}^2\right\},\end{aligned} \tag{33}$$

where the statistics of Y_k are computed in Appendix B and first two raw moments of RV $Z_{(n)}$ can be derived analytically based on the PDF information presented in (30) (see [20]):

$$\begin{aligned}\mathbb{E}\left\{Z_{(n)}\right\} &= \sum_{k=1}^{n}\left[(-1)^{k+1}\binom{n}{k}\sum_{l=1}^{k}\binom{k}{l}\frac{l(l+1)!}{k^{l+2}}\right],\\ \mathbb{E}\left\{Z_{(n)}^2\right\} &= \sum_{k=1}^{n}\left[(-1)^{k+1}\binom{n}{k}\sum_{l=1}^{k}\binom{k}{l}\frac{l(l+2)!}{k^{l+3}}\right].\end{aligned} \tag{34}$$

When dealing with numbers of users that can be handled in realistic scenarios, it is possible to observe that $\mathcal{F}_{z_{(n)}\tilde{y}} \approx 1$. However, in order to show that this fading figure does not grow indefinitely with n, the following asymptotic upper bound for $\mathcal{F}_{z_{(n)}\tilde{y}}$ is derived:

$$\begin{aligned}\lim_{n\to\infty}\mathcal{F}_{z_{(n)}\tilde{y}} &= \lim_{n\to\infty}\frac{\mathbb{E}^2\left\{Z_{(n)}\widetilde{Y}_k\right\}}{\mathbb{E}\left\{Z_{(n)}^2\widetilde{Y}_k^2\right\} - \mathbb{E}^2\left\{Z_{(n)}\widetilde{Y}_k\right\}}\\ &= \frac{3\mathcal{G}_y^2}{2\mathbb{E}\left\{Y_k^2\right\} - 3\mathcal{G}_y^2},\end{aligned} \tag{35}$$

based on the fact that $\mathbb{E}^2\{Z_{(n)}\}/\mathrm{Var}\{Z_{(n)}\} \to 0$ as n grows. We note that according to this formula, asymptotic upper bounds for this fading figure are equal to 1.8838 and 1.5509 for CL TD modes 1 and 2, respectively.

At this stage, we use (10) to approximate $f_{z_{(n)}\tilde{y}}(u)$ in order to compute

$$
\begin{aligned}
&\mathbb{E}\{\log_e(Z_{(n)}\widetilde{Y}_k + N_0')\} \\
&\quad \approx \int_0^\infty \log_e(u+N_0')\left(\frac{1}{\mathcal{G}_{z_{(n)}\tilde{y}}}\right)e^{-(1/\mathcal{G}_{z_{(n)}\tilde{y}})u} \\
&\qquad \times [a_2^{(n)}u^2 + a_1^{(n)}u + a_0^{(n)}]du \\
&\quad \approx [2a_2^{(n)}\mathcal{G}^2_{z_{(n)}\tilde{y}} + a_1^{(n)}\mathcal{G}_{z_{(n)}\tilde{y}} + a_0^{(n)}] \\
&\qquad \times \left[\log_e(N_0') + e^{N_0'/\mathcal{G}_{z_{(n)}\tilde{y}}}E_1\left(\frac{N_0'}{\mathcal{G}_{z_{(n)}\tilde{y}}}\right)\right] \\
&\qquad + [2a_2^{(n)}\mathcal{G}^2_{z_{(n)}\tilde{y}} + a_1^{(n)}\mathcal{G}_{z_{(n)}\tilde{y}}]e^{N_0'/\mathcal{G}_{z_{(n)}\tilde{y}}}E_2\left(\frac{N_0'}{\mathcal{G}_{z_{(n)}\tilde{y}}}\right) \\
&\qquad + 2a_2^{(n)}\mathcal{G}^2_{z_{(n)}\tilde{y}}e^{N_0'/\mathcal{G}_{z_{(n)}\tilde{y}}}E_3\left(\frac{N_0'}{\mathcal{G}_{z_{(n)}\tilde{y}}}\right),
\end{aligned}
\tag{36}
$$

where $\mathcal{G}_{z_{(n)}\tilde{y}} = \mathbb{E}\{Z_{(n)}\widetilde{Y}_k\}$ and coefficients $a_i^{(n)}$ are derived analytically based on the first two raw moments of RV $Z_{(n)}\widetilde{Y}_k$ according to

$$
\begin{aligned}
a_2^{(n)} &= \frac{2\mathbb{E}\{Y_k^2\}\mathbb{E}\{Z_{(n)}^2\} - 6\mathcal{G}_y^2\mathbb{E}^2\{Z_{(n)}\}}{3\mathcal{G}_y^4\mathbb{E}^4\{Z_{(n)}\}}, \\
a_1^{(n)} &= \frac{-4\mathbb{E}\{Y_k^2\}\mathbb{E}\{Z_{(n)}^2\} + 12\mathcal{G}_y^2\mathbb{E}^2\{Z_{(n)}\}}{3\mathcal{G}_y^3\mathbb{E}^3\{Z_{(n)}\}}, \\
a_0^{(n)} &= \frac{1}{3}\frac{\mathbb{E}\{Y_k^2\}\mathbb{E}\{Z_{(n)}^2\}}{\mathcal{G}_y^2\mathbb{E}^2\{Z_{(n)}\}}.
\end{aligned}
\tag{37}
$$

Replacing (32), (36), and (37) in (29), the final closed-form approximation to estimate the achievable rate per beam with CDI and channel norm CQI feedback is obtained.

4.2.1. Low-SNR Regime. Applying Taylor series expansion in (28) when the Tx power is low,

$$
\begin{aligned}
C_{(n)}^{\text{Norm}}(P) &\approx \log_2(e)\frac{1}{2}\frac{P}{N_0}\mathbb{E}\{Z_{(n)}\}\mathbb{E}\{\widetilde{X}_k\} \\
&= \frac{\mathbb{E}\{Z_{(n)}\}}{\mathbb{E}\{Z_k\}}C_k(P) \quad P \ll N_0.
\end{aligned}
\tag{38}
$$

The asymptotic behavior of the largest order statistic of n i.i.d. χ^2 RVs with $2M_t$ degrees of freedom has been reported in [15] to be

$$
\mathbb{E}\{\gamma_{(n)}\} \doteq \log_e(n) + \log_e\left[\frac{n^{M_t-1}}{(M_t-1)!}\right] + \epsilon_0, \tag{39}
$$

where notation $c_n \doteq d_n$ denotes asymptotic equivalence, defined as $\lim_{n\to\infty}(c_n/d_n) = 1$. Based on this, multiuser diversity gain in case of channel norm CQI and $M_t = 2$ is given by

$$
\frac{C_{(n)}^{\text{Norm}}}{C_k} \approx \frac{\mathbb{E}\{Z_{(n)}\}}{\mathbb{E}\{Z_k\}} \doteq \log_e(n) + \frac{1}{2}\epsilon_0 \quad P \ll N_0. \tag{40}
$$

4.2.2. High-SNR Regime. A scheduler that relies on channel norm CQI to perform user selection has always the same asymptotic behavior, which does not depend on the number of active users. This is because SIR feedback is not considered for scheduling purposes; therefore, both the desired signal and mutual interference tend to grow with the same proportion as Tx power increases. So, we conclude that the upper bound for any smart scheduling scheme in this situation is identical to the one already obtained in Section 3.3.

4.3. Achievable Rate for Spatial Multiplexing with CDI and SIR CQI. The effect of exploiting multiuser diversity when users reports are proportional to the received SIR is analyzed in this part. The procedure consists of scheduling the user with the largest SIR CQI

$$
\check{Q}(\mathbf{h}_k) = \frac{X_k}{Y_k} = \frac{Z_k - Y_k}{Y_k} = \frac{Z_k(1-\widetilde{Y}_k)}{Z_k\widetilde{Y}_k} = \frac{1-\widetilde{Y}_k}{\widetilde{Y}_k}, \tag{41}
$$

which reduces to select the user that minimize $\widetilde{Y}_k$. Based on this, the achievable rate per beam when there are n users reporting the same CDI codeword and SIR CQI is given by

$$
\begin{aligned}
C_{(n)}^{\text{SIR}}(P) &= \mathbb{E}\left\{\log_2\left(1 + \frac{(1/2)PZ_k[1-\widetilde{Y}_{(1)}]}{(1/2)PZ_k\widetilde{Y}_{(1)} + N_0}\right)\right\} \\
&= \log_2(e)[\mathbb{E}\{\log_e(Z_k + N_0')\} \\
&\qquad - \mathbb{E}\{\log_e(Z_k\widetilde{Y}_{(1)} + N_0')\}],
\end{aligned}
\tag{42}
$$

where $\widetilde{Y}_{(1)} = \min_{i=1,\ldots,n}\widetilde{Y}_i$. We now need to find out an approximation for the distribution of the mutual interference $Z_k\widetilde{Y}_{(1)}$. Thus, we first study the behavior of RV $\widetilde{Y}_{(1)}$.

According to [9, 21], the CDF $F_{\tilde{y}}(y)$ for any well-designed codebook satisfies $F_{\tilde{y}^\star}(y) \geq F_{\tilde{y}}(y)$ for $0 \leq y \leq 1$, where

$$
F_{\tilde{y}^\star}(y) = \begin{cases} 2^B y^{M_t-1}, & 0 \leq y < 2^{-B/(M_t-1),} \\ 1, & y \geq 2^{-B/(M_t-1)} \end{cases} \tag{43}
$$

represents quantization error CDF when *quantization cell upper bound* (QUB) approach is employed as a performance upper bound for any CDI codebook design. Based on this, it is possible to observe that the CDF of RV $\widetilde{Y}_k = 1 - |\bar{\mathbf{h}}_k\mathbf{w}_k|^2$ in case of $M_t = 2$ (and both CL TD feedback modes) will be upper bounded (in all its range) by the CDF of a uniform RV in $[0, 2^{-B}]$. Since the kth-order statistic of n uniformly distributed RVs in $[0, 1]$ is Beta distributed according to

$$
f_{u_{(k)}}(u) = \frac{n!}{(k-1)!(n-k)!}u^{k-1}(1-u)^{n-k}, \quad 0 \leq u \leq 1, \tag{44}
$$

the following expressions for the first two-ordered raw moments result:

$$
\begin{aligned}
\mathbb{E}\{U_{(k)}\} &= \frac{k}{n+1}, \\
\mathbb{E}\{U_{(k)}^2\} &= \frac{k(k+1)}{(n+1)(n+2)}.
\end{aligned}
\tag{45}
$$

Taking into account that RVs Z_k and $\widetilde{Y}_k$ are independent, it is possible to lower bound the first two raw moments of mutual interference $Z_k\widetilde{Y}_{(1)}$ as

$$\begin{aligned}\mathbb{E}\{Z_k\widetilde{Y}_{(1)}\} &= \mathcal{G}_z\mathbb{E}\{\widetilde{Y}_{(1)}\} \gtrapprox 2\left(\frac{1}{2^B}\right)\left[\frac{1}{(n+1)}\right],\\ \mathbb{E}\{Z_k^2\widetilde{Y}_{(1)}^2\} &\gtrapprox 6\left(\frac{1}{4^B}\right)\left[\frac{2}{(n+1)(n+2)}\right],\end{aligned} \tag{46}$$

where these approximations are asymptotically tight as n grows. According to these results, we see that fading figure $\mathcal{F}_{z\widetilde{y}_{(1)}} \approx (n+2)/(2n+1)$, which is still close to 1 for both CL TD feedback modes when dealing with numbers of users that can be handled in realistic scenarios.

Using first-order corrected version presented in (10) to approximate $f_{z\widetilde{y}_{(1)}}(u)$ by an exponential distribution with parameter $\mathcal{G}_{z\widetilde{y}_{(1)}} = \mathbb{E}\{Z_k\widetilde{Y}_{(1)}\}$, it is possible to see that

$$\begin{aligned}&\mathbb{E}\{\log_e(Z_k\widetilde{Y}_{(1)} + N_0')\}\\ &\quad\approx \int_0^\infty \log_e(u+N_0')\left(\frac{1}{\mathcal{G}_{z\widetilde{y}_{(1)}}}\right)e^{-(1/\mathcal{G}_{z\widetilde{y}_{(1)}})u}\\ &\qquad\times[\breve{a}_2^{(n)}u^2 + \breve{a}_1^{(n)}u + \breve{a}_0^{(n)}]du\\ &\quad\approx \left[2\breve{a}_2^{(n)}\mathcal{G}_{z\widetilde{y}_{(1)}}^2 + \breve{a}_1^{(n)}\mathcal{G}_{z\widetilde{y}_{(1)}} + \breve{a}_0^{(n)}\right]\\ &\qquad\times\left[\log_e(N_0') + e^{N_0'/\mathcal{G}_{z\widetilde{y}_{(1)}}}E_1\left(\frac{N_0'}{\mathcal{G}_{z\widetilde{y}_{(1)}}}\right)\right]\\ &\qquad+\left[2\breve{a}_2^{(n)}\mathcal{G}_{z\widetilde{y}_{(1)}}^2 + \breve{a}_1^{(n)}\mathcal{G}_{z\widetilde{y}_{(1)}}\right]e^{N_0'/\mathcal{G}_{z\widetilde{y}_{(1)}}}E_2\left(\frac{N_0'}{\mathcal{G}_{z\widetilde{y}_{(1)}}}\right)\\ &\qquad+2\breve{a}_2^{(n)}\mathcal{G}_{z\widetilde{y}_{(1)}}^2e^{N_0'/\mathcal{G}_{z\widetilde{y}_{(1)}}}E_3\left(\frac{N_0'}{\mathcal{G}_{z\widetilde{y}_{(1)}}}\right),\end{aligned} \tag{47}$$

where coefficients $\breve{a}_i^{(n)}$ are derived analytically based on the first two raw moments of RV $Z_k\widetilde{Y}_{(1)}$:

$$\begin{aligned}\breve{a}_2^{(n)} &= (4^{B-2})\frac{(n-1)(n+1)^2}{(n+2)},\\ \breve{a}_1^{(n)} &= -(2^{B-1})\frac{(n-1)(n+1)}{(n+2)},\\ \breve{a}_0^{(n)} &= \left(\frac{3}{2}\right)\frac{(n+1)}{(n+2)}.\end{aligned} \tag{48}$$

Replacing (17), (47), and (48) in (42), final closed-form approximation to estimate the achievable rate per beam with CDI and SIR CQI feedback is obtained.

4.3.1. Low-SNR Regime. Schedulers that rely on SIR CQI to perform user selection do not provide any multiuser diversity gain when Tx power is low. This is because they do not consider channel norm information to carry out decisions.

4.3.2. High-SNR Regime. Because the reported SIR CQI is not perfect, the achievable rate still has an interference-limited behavior in this situation. However, the corresponding asymptotic upper bound grows logarithmically with the number of users. The closed-form expression for this upper bound is obtained replacing SNR gain and fading figure approximations in (22). After some manipulations, final expression

$$\lim_{P\to\infty} C_{(n)}^{\mathrm{SIR}}(P) \approx \log_2(e)\left(\frac{n+3/2}{n+2}\right) + \log_2[2^B(n+1)] \tag{49}$$

results, which reduces to $B + \log_2(n)$ when the number of participating users is large. Therefore, multiuser selection policy based on SIR CQI provides a logarithmic increase in limiting achievable rate [10]. This is in contrast to previous findings, where system rate improvement due to multiuser diversity effect was only by a factor of a double logarithm with respect to the number of users.

5. Short-Term Fairness: Concepts and Performance Metric

Fairness in wireless networks indicates how equally radio resources are allocated among mobile users. Fairness should always be evaluated within a window in time. Those scheduling algorithms that obtain high fairness over a relatively short-time window are denoted as short-term fair, while the algorithms that obtain high fairness over an infinite-time window are denoted as asymptotically fair. The provision of short-term fairness characteristics for any multiuser diversity scheme is important because networking protocols usually have timers at different protocol layers that interact with each other in an unpredictable manner. An expiration of a timer is a bad event for an end-to-end connection. Such an event is usually interpreted as an indicator of congestion and loss of connectivity [22]. Thus, short-term fairness is always desirable for any packet scheduling procedures that reap multiuser diversity gain.

Several measures of fairness have been introduced in literature. Perhaps the simplest indicator is the so-called *Jain's fairness index* (JFI), introduced in [3] and used in recent papers such as [23] to characterize fairness behavior over a finite horizon:

$$F_k(W) = \frac{\mathbb{E}_W^2\{R_k\}}{\mathbb{E}_W\{R_k^2\}} = \frac{\mathbb{E}_W^2\{R_k\}}{\mathbb{E}_W^2\{R_k\} + \mathrm{Var}\{R_k\}}, \tag{50}$$

where R_k is an RV that describes the amount of resource allocated to user k, $\mathbb{E}_W\{R_k\}$ is the expectation calculated within a time window of length W (time slots), and $\mathrm{Var}\{R_k\}$ is the corresponding variance.

Jain's fairness index has several properties that makes it a suitable fairness measure. For example, the index is continuous and bounded between zero and unity. Moreover, JFI does not depend on the amount of the shared resource and on the number of participating users. The boundedness of JFI aids intuitive understanding of the fairness index. Even though an ideal fair distribution of common resources would result in an index of 1, values above 0.95 are typically considered to indicate excellent fairness properties.

Resource allocation can be measured either in terms of the number of time slots assigned to a given user (within a window), or in terms of the throughput that was experienced by the user in these allocated time slots. However, here we only focus on the latter definition since achieving time-slot fairness in the presence of time-varying channels does not necessary imply a fair allocation of throughput in the assigned time slots. Throughput fairness curves for the different scheduling procedures introduced so far are presented in Section 6.3 with the goal of quantifying the short-term fairness performance that is sacrificed at the expense of obtaining additional multiuser diversity gain in our virtual MIMO system sum rate.

6. Performance Evaluation: Throughput-Fairness Behavior in Virtual MIMO

The actual virtual MIMO system sum rates for three different scheduling procedures and two CQI definitions are studied in this section based on intermediate results derived in Sections 3 and 4. The schedulers select at each time a pair of users that report orthogonal CDI codewords and differ with respect to their usage of CQI in scheduling decisions. Note that in those situations where scheduler fails to find a set of semiorthogonal users, the BS may either schedule transmission to a single user or resign the channel use at that time instant. Even though the former approach is most reasonable for a real-world system implementation, in this work we focus on the latter since we want to provide a representative characterization for TDMA and SDMA schemes when they work independently, leaving aside complex interactions between them that makes sum rate performance difficult to analyze.

6.1. Virtual MIMO System Sum Rate with CDI and No CQI. In this part, we consider the case of scheduling a pair of semiorthogonal users when no CQI is available at BS to perform user selection. We work on a simple case, known as pure ORR scheduler [12], where both primary and secondary users are selected in RR. Note that the performance in this case is equivalent to the one observed in case of PF scheduler when window size is tuned to optimize short-term fairness (large throughput tracking window). As expected, achievable sum rate is given by

$$C^{\mathrm{ORR}}(P) = 2C_k(P), \tag{51}$$

where closed-form approximation for $C_k(P)$ was derived in Section 3.3. Virtual MIMO system sum rate for pure ORR scheme is analyzed in Section 6.3. However, we now focus on analyzing the throughput behavior of an individual user when its selection does not take into account CQI reports.

Figure 1 shows the achievable rate (per beam) for pure ORR scheme when the CDI is represented by CL TD modes 1 and 2. The curves correspond to analytical approximation (51) (dashed curves) along with its corresponding asymptotic upper bounds (dash-dotted lines). The achievable rate for TDMA-RR is also included in these plots (solid lines). To make a fair comparison, Tx power in case of TDMA-RR is equal to the power per beam in case of pure ORR scheduling. As expected, the achievable rate for both TDMA-RR and pure ORR tends to be identical as Tx power decreases.

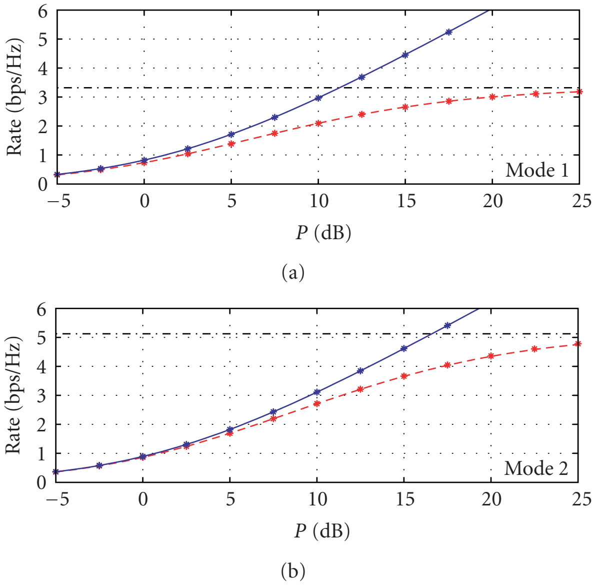
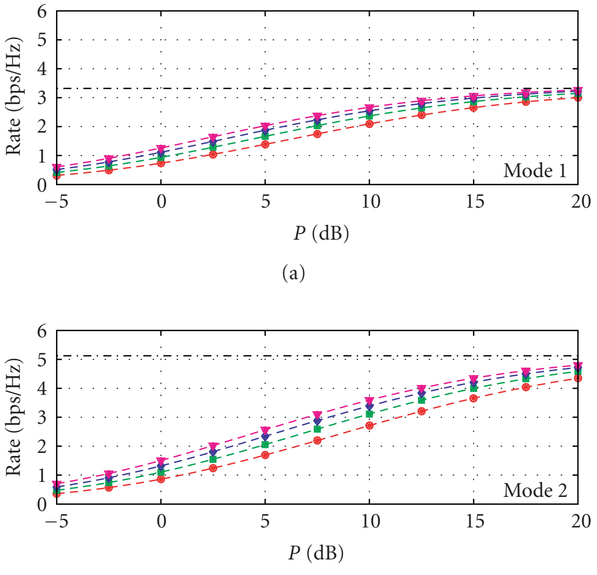


Figure 1: Achievable rate per beam for two-antenna mode 1 and 2 in the presence of Rayleigh fading and constant Tx power. Solid curves refer to TDMA-RR when total Tx power is normalized to $P/2$. Dashed curves refer to achievable rate per beam for spatial multiplexing with no CQI, and dash-dotted curves represent the asymptotic upper bound behavior presented in (22). In all cases, point values ("$*$") were simulated to verify the analytical results.

6.2. Virtual MIMO System Sum Rate with both CDI and CQI. Simple hybrid ORR proposals, known as ORR-Norm and ORR-SIR depending on the type of CQI that mobiles report, were introduced in [12] as improved versions of pure ORR scheme. These hybrid schedulers guarantee a certain degree of fairness by selecting the primary user according to its waiting time in transmission and exploit multiuser diversity in the selection of the secondary semiorthogonal user. Thus, achievable sum rate in this situation is now given by

$$\begin{aligned} C^{\mathrm{ORR}}_{\mathrm{CQI}}&(P,K) \\ &= C_k(P) \\ &\quad + \sum_{n=1}^{K-1}\left\{\left[\binom{K-1}{n}\left(2^{-B}\right)^n\left(1-2^{-B}\right)^{K-n-1}\right]C^{\mathrm{CQI}}_{(n)}(P)\right\}. \end{aligned} \tag{52}$$

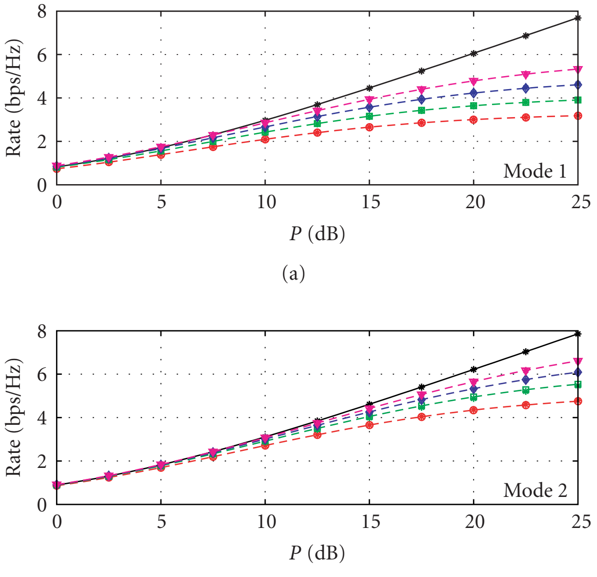
The first term in (52) represents the achievable rate for the primary user selected in RR (Section 3.3), while the second term approximates the achievable rate for the secondary user



(a)

(b)

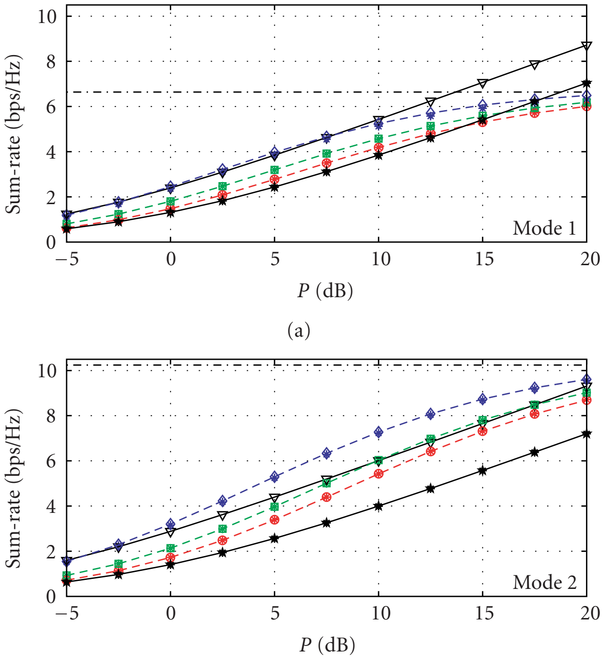
Figure 2: Achievable rate per beam for two-antenna mode 1 and 2 in the presence of Rayleigh fading and constant Tx power. Dashed curves refer to achievable rate per beam for spatial multiplexing with channel norm CQI and different number of users reporting identical CDI ($n = 1, 2, 4, 8$). Dash-dotted curves represent the asymptotic upper bound behavior presented in (22). In all cases, point values ("$*$") were simulated to verify the analytical results.



(a)

(b)

Figure 3: Achievable rate per beam for two-antenna mode 1 and 2 in the presence of Rayleigh fading and constant Tx power. Solid curves refer to TDMA-RR when total Tx power is normalized to $P/2$. Dashed curves refer to achievable rate per beam for spatial multiplexing with SIR CQI and different number of users reporting identical CDI ($n = 1, 2, 4, 8$). In all cases, point values ("$*$") were simulated to verify the analytical results.



(a)

(b)

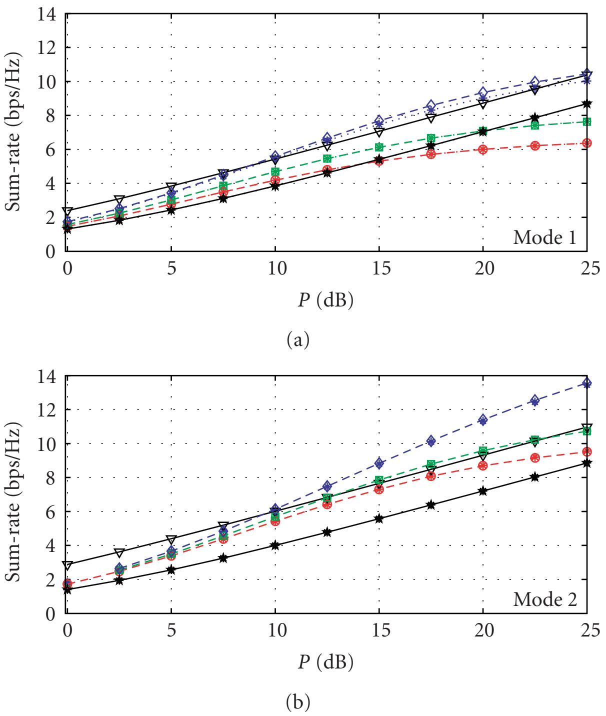
Figure 4: Virtual MIMO system sum-rate for two-antenna mode 1 ($K = 16$) and mode 2 ($K = 64$) in the presence of Rayleigh fading, constant Tx power, and channel norm CQI. Solid curves with stars ("$\star$") refer to TDMA-RR, while solid lines with triangles ("$\triangledown$") correspond to TDMA-BUS. Dashed curves with circles ("$\circ$"), dashed curves with squares ("$\square$") and dashed curves with diamonds ("$\diamond$") refer to pure ORR (ZFBF-PF throughput lower bound), ORR-Norm and ZFBF-SUS (ZFBF-PF throughput upper bound), respectively. Dash-dotted curves represent the asymptotic upper bound presented in (22). In all cases, point values ("$*$") were simulated to verify the analytical results.

selected according to the channel norm CQI (Section 4.2) and SIR CQI (Section 4.3).

Throughput upper bound for ZFBF-PF scheme is achieved when users instantaneous rates are not normalized by their average throughput before performing selection (small throughput tracking window). This is equivalent to choosing the set of users that maximize sum rate at each time slot without considering short-term fairness issues. It has already been observed in Section 2.2 that SUS procedure provides a simple way to obtain a set of semiorthogonal users with large CQI. Based on this, achievable sum rate in this situation can be represented by

$$\begin{aligned} &C^{\mathrm{SUS}}_{\mathrm{CQI}}(P,K)\\ &\quad\lessapprox C^{\mathrm{CQI}}_{(K)}(P)\\ &\qquad+\sum_{n=1}^{K-1}\left\{\left[\binom{K-1}{n}\left(2^{-B}\right)^{n}\left(1-2^{-B}\right)^{K-n-1}\right]C^{\mathrm{CQI}}_{(n)}(P)\right\}. \end{aligned} \tag{53}$$

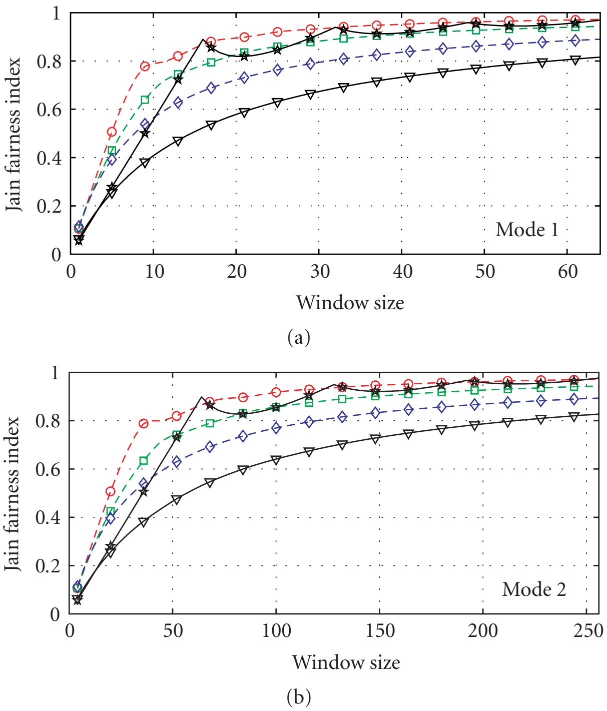
The first term in (53) represents the achievable rate of the user with the best CQI among all active users, while



(a)

(b)

FIGURE 5: Virtual MIMO system sum-rate for two-antenna mode 1 ($K = 16$) and mode 2 ($K = 64$) in the presence of Rayleigh fading, constant Tx power, and SIR CQI. Solid curves with stars ("$\star$") refer to TDMA-RR, while solid lines with triangles ("$\bigtriangledown$") correspond to TDMA-BUS. Dashed curves with circles ("$\circ$"), dashed curves with squares ("$\square$") and dashed curves with diamonds ("$\diamond$") refer to pure ORR (ZFBF-PF throughput lower bound), ORR-SIR and ZFBF-SUS (ZFBF-PF throughput upper bound), respectively. In all cases, point values ("$\ast$") were simulated to verify the analytical results.



(a)

(b)

FIGURE 6: Virtual MIMO system throughput fairness index for two-antenna mode 1 ($K = 16$) and mode 2 ($K = 64$) in the presence of Rayleigh fading, constant Tx power ($P = 5$ dB), and channel norm CQI. Solid curves with stars ("$\star$") refer to TDMA-RR, while solid lines with triangles ("$\bigtriangledown$") correspond to TDMA-BUS. Dashed curves with circles ("$\circ$"), dashed curves with squares ("$\square$") and dashed curves with diamonds ("$\diamond$") refer to pure ORR (ZFBF-PF fairness upper bound), ORR-Norm and ZFBF-SUS (ZFBF-PF fairness lower bound), respectively.

the second term approximates the achievable rate of the user with the largest CQI among all users that satisfy orthogonality criterion (with respect to the first selected user). Note that final closed-form expression in this case is actually a tight upper bound because now independence assumption between ordered statistics of individual users rates in both terms is no longer valid. Virtual MIMO system sum rates for both hybrid ORR and ZFBF-SUS (both CQI definitions) are analyzed in Section 6.3. We now focus on the achievable rate of an individual user when its selection takes advantage of CQI reports.

Figures 2 and 3 show the achievable rate (per beam) when users are selected based on channel norms CQI and SIR CQI, respectively. These curves correspond to analytical approximations (29) and (42) (dashed curves), along with their simulated point values ("$\ast$"). Again, it is observed that the proposed approximations follow simulated values well for different numbers of users in both CL TD modes. As expected, the use of SIR CQI instead of channel norm CQI provides a better performance at high-SNR regimes.

6.3. Tradeoff Analysis of Throughput and Fairness in Virtual MIMO Systems. We are now ready to analyze the interaction between overall system throughput and short-term throughput fairness that the different channel-aware scheduling procedures introduced so far are able to provide. In this context, Figures 4 and 5 present the actual virtual MIMO system sum rate for pure ORR, hybrid ORR-CQI, and ZFBF-SUS schemes when both channel norm CQI and SIR CQI are exploited, respectively. These curves correspond to analytical approximations (51), (52), and (53), along with their simulated point values ("$\ast$"). In addition, performances of TDMA-RR and TDMA-BUS (i.e., the TDMA scheme that selects the user with the highest channel gain at each time) are also included for the sake of comparison. To complement these plots, Figures 6 and 7 show the short-term fairness behavior for these schemes when using the fairness index introduced in (50) as a performance measure for different time-window horizons.

When analyzing these curves, it is straightforward to observe that, as expected, those schemes that reap higher multiuser diversity gain require a larger window size to achieve a certain degree of throughput fairness. Even though interesting tradeoffs between throughput and fairness can be reported when comparing these figures, perhaps the most important conclusion to highlight is that the simultaneous transmission to a set of smartly selected users provides

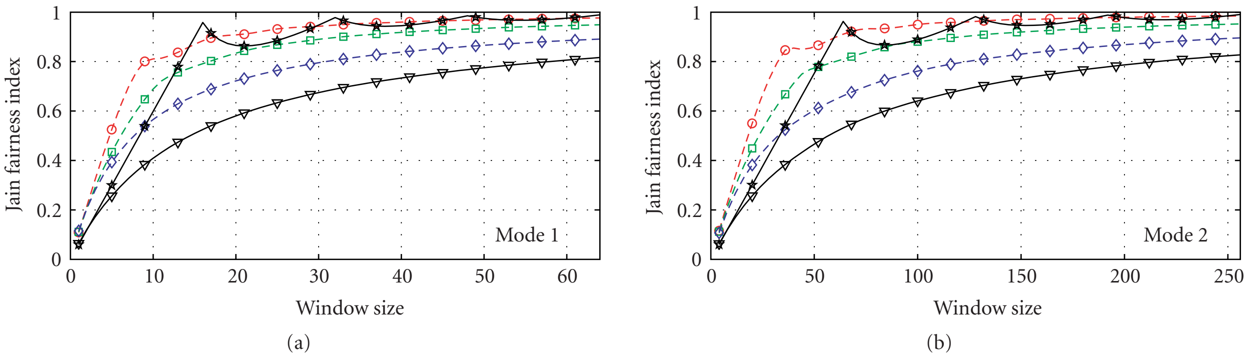


Figure 7: Virtual MIMO system throughput fairness index for two-antenna mode 1 ($K = 16$) and mode 2 ($K = 64$) in the presence of Rayleigh fading, constant Tx power ($P = 15$ dB), and SIR CQI. Solid curves with stars ("⋆") refer to TDMA-RR, while solid lines with triangles ("▽") correspond to TDMA-BUS. Dashed curves with circles ("∘"), dashed curves with squares ("□") and dashed curves with diamonds ("⋄") refer to pure ORR (ZFBF-PF fairness upper bound), ORR-SIR and ZFBF-SUS (ZFBF-PF fairness lower bound), respectively.

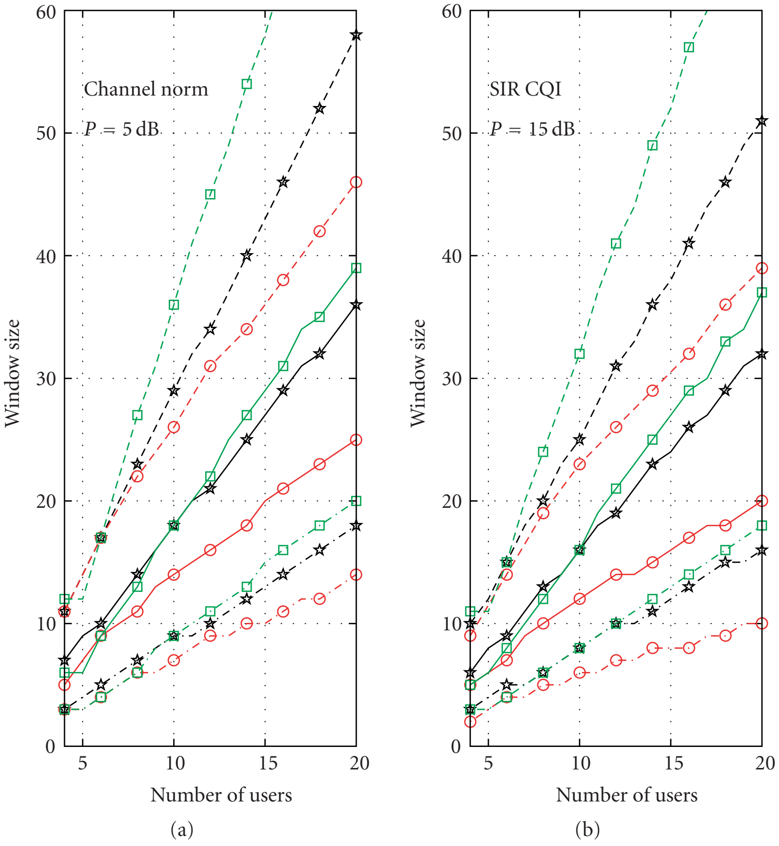


Figure 8: Required window size to achieve short-term throughput fairness with two-antenna mode 1 in the presence of Rayleigh fading, constant Tx power, and both CQI definitions. Curves with circles ("∘"), stars ("⋆") and squares ("□") refer to pure ORR, TDMA-RR and ORR-CQI, respectively. In all cases, dash-dotted curves represent a fairness index of 0.8, solid lines correspond to a fairness index of 0.9, and dashed curves refer to a fairness index of 0.95.

a better performance both, in terms of throughput and fairness, when compared to an analogous TDMA scheme. For example, when comparing TDMA-RR and pure ORR schemes, it is noticed that the latter provides as much as 15% (35%) more throughput in case of CL TD mode 1 (mode 2) without affecting considerably the short-term fairness degree that the former provides. Similar results are obtained when comparing TDMA-BUS with ZFBF-SUS, but in this situation some throughput gain is traded off with an increase of the short-term fairness. It is important to highlight that the amount of CDI feedback does not impact considerably on the fairness of the schemes introduced so far if the ratio between the total number of users and the number of CDI codewords remains constant;

however, it actually has a direct effect on the multiuser diversity gain that these scheduling procedures provide , particularly when dealing with SIR CQI in high Tx power region.

Finally, Figure 8 shows the window size that is required to attain a certain level of short-term fairness as a function of the number of users in case of CL TD mode 1 with channel norm CQI ($P = 5\,\text{dB}$) and SIR CQI ($P = 15\,\text{dB}$), respectively. In these figures, dash-dotted curves represent a fairness index of 0.8, solid lines correspond to a fairness index of 0.9, and dashed curves refer to a fairness index of 0.95. Only TDMA-RR, pure ORR, and ORR-CQI schemes are included. This is because both ZFBF-SUS and TDMA-BUS do not provide acceptable fairness levels within practical window sizes. According to these curves, the time window that is required to achieve a certain degree of fairness grows linearly with the number of users. As expected, the slope of the curves depends not only on the requested fairness level, but also on the scheduling scheme. Note that both Tx power and CQI definition have a weaker effect on fairness performance. In all cases, pure ORR is the scheme with the best behavior. Note that the gap between TDMA-RR and ORR-CQI tends to grow as required level of fairness increases; however, for fairness levels up to 0.9, performance difference between these two schemes is almost negligible. We highlight that similar behavior is observed in case of CL TD mode 2.

7. Conclusions

In this paper, we investigated the tradeoff between maximizing system throughput and achieving throughput fairness in virtual MIMO downlink systems with quantized channel direction information and different types of channel quality information in the transmitter. We proposed a new theoretical approach to derive closed-form approximations to quantify throughput performance when combining different scheduling rules with zero-forcing beamforming. The short-term fairness analysis of the different schemes was performed using Jain's fairness index as performance metric. The advantages and disadvantages of the different schemes were highlighted by visualizing our closed-form expressions.

In our proposed theoretical model both desired signal energy and mutual interference in reception are modeled with first-order corrected versions of chi-square distributions, with characterization parameters obtained based on the first two raw statistics of these signals. The derived expressions were validated using existing 3GPP physical layer signaling structures. Our analysis revealed that simple scheduling procedures allow to reap a large fraction (in the order of 80%) of the sum-rate performance that greedy scheduling provides. This overall throughput performance was obtained without affecting considerably the optimal short-term fairness behavior that the end user would perceive.

Appendices

A. Error Correction

When approximating a generic distribution $f(\gamma)$ (with unknown closed-form formula) by a χ^2-distribution with r degrees of freedom and mean η, the error

$$\varepsilon(\gamma) = f(\gamma) - \frac{1}{\Gamma(r/2)}\left(\frac{r}{2\eta}\right)^{r/2}\gamma^{r/2-1}e^{-[r/(2\eta)]\gamma} \tag{A.1}$$

results. We shall express this error in terms of the raw moments $\mathbb{E}\{\gamma^n\}$ and the generalized Laguerre polynomials

$$\begin{aligned} L_k^{(\alpha)}(u) &= \frac{u^{-\alpha}e^{u}}{k!}\frac{\partial^k}{\partial u^k}\left(e^{-u}u^{k+\alpha}\right) \\ &= \sum_{i=0}^{k}\binom{k+\alpha}{k-i}\frac{(-u)^i}{i!}. \end{aligned} \tag{A.2}$$

These polynomials are orthogonal over the entire real line with respect to the weighting function $u^{\alpha}e^{-u}$; therefore,

$$\int_0^{\infty} u^{\alpha}e^{-u}L_k^{(\alpha)}(u)L_l^{(\alpha)}(u)\,du = \frac{(k+\alpha)!}{k!}\delta_{kl}, \tag{A.3}$$

where δ_{kl} is the Kronecker delta function. The orthogonality property stated above is equivalent to saying that if γ is a χ^2-distribution with r degrees of freedom and mean η, then

$$\mathbb{E}\{L_k^{(\alpha)}(\beta\gamma)L_l^{(\alpha)}(\beta\gamma)\} = \begin{cases} \dfrac{(k+\alpha)!}{\alpha!k!}, & k = l, \\ 0, & k \neq l \end{cases} \tag{A.4}$$

with $\alpha = r/2 - 1$ and $\beta = r/(2\eta)$. Hence, the error can be written as a series

$$\varepsilon(\gamma) = \frac{\beta}{\alpha!}(\beta\gamma)^{\alpha}e^{-\beta\gamma}\left[\sum_{k=2}^{+\infty} C_k^{(\alpha)}L_k^{(\alpha)}(\beta\gamma)\right]. \tag{A.5}$$

Series starts with $k = 2$ because moments of $\varepsilon(\gamma)$ of order up to 1 are null. In following sections, we show how can coefficients $C_k^{(\alpha)}$ be expressed in terms of the (known) raw moments of γ.

A.1. First-Order Correction for Exponential PDF Approximation. Let us first concentrate on the first-order error corrected version for $f(\gamma)$ when fading figure $\mathcal{F} \approx 1$. This approximation is obtained retaining the first nonzero term of the sum in (A.5), that is,

$$f(\gamma) \approx \frac{\beta}{\alpha!}(\beta\gamma)^{\alpha}e^{-\beta\gamma}\left[1 + C_2^{(\alpha)}L_2^{(\alpha)}(\beta\gamma)\right]. \tag{A.6}$$

Since in this case the exponential distribution (i.e., χ^2-distribution with $r = 2$ degrees of freedom) is the most suitable approximation, we have that $\alpha = 0$ and $\beta = 1/\eta$. It follows from (A.2) that

$$L_2^{(0)}(\beta\gamma) = \frac{1}{2}\left[(\beta\gamma)^2 - 4(\beta\gamma) + 2\right]. \tag{A.7}$$

Therefore, we only need to determine $C_2^{(0)}$. In order to do so, we have that

$$\begin{aligned}\int_0^\infty & L_2^{(0)}(\beta\gamma)\varepsilon(\gamma)d\gamma \\ &= \int_0^\infty L_2^{(0)}(\beta\gamma)\beta e^{-\beta\gamma}\sum_{k=2}^{\infty}\left[C_k^{(0)}L_k^{(0)}(\beta\gamma)\right]d\gamma \\ &= C_2^{(0)}\int_0^\infty L_2^{(0)}(\beta\gamma)L_2^{(0)}(\beta\gamma)\beta e^{-\beta\gamma}\,d\gamma \\ &\quad+\sum_{k=3}^{\infty}C_k^{(0)}\int_0^\infty L_2^{(0)}(\beta\gamma)L_k^{(0)}(\beta\gamma)\beta e^{-\beta\gamma}\,d\gamma.\end{aligned} \tag{A.8}$$

Orthogonality property introduced in (A.4) states that integral in the first term equals 1, while all integrals in the sum are null. Based on these considerations, it is possible to see that

$$\int_0^\infty L_2^{(0)}(\beta\gamma)\varepsilon(\gamma)d\gamma = C_2^{(0)}. \tag{A.9}$$

Following an alternative analysis, that is, replacing $L_2^{(0)}(\beta\gamma)$ by expression (A.7), we also have that

$$\begin{aligned}\int_0^\infty L_2^{(0)}(\beta\gamma)\varepsilon(\gamma)d\gamma = \frac{1}{2}\int_0^\infty(\beta\gamma)^2\varepsilon(\gamma)d\gamma - 2\int_0^\infty(\beta\gamma)\varepsilon(\gamma)d\gamma \\ +\int_0^\infty\varepsilon(\gamma)d\gamma.\end{aligned} \tag{A.10}$$

The last two integrals vanish because the moments of $\varepsilon(\gamma)$ of order up to 1 are null. Therefore,

$$\begin{aligned}\int_0^\infty & L_2^{(0)}(\beta\gamma)\varepsilon(\gamma)d\gamma \\ &= \frac{1}{2}\left[\int_0^\infty(\beta\gamma)^2 f(\gamma)d\gamma - \int_0^\infty(\beta\gamma)^2\beta e^{-\beta\gamma}\,d\gamma\right] \\ &= \frac{\beta^2}{2}\left[\mathbb{E}\{\gamma^2\} - \frac{2}{\beta^2}\right].\end{aligned} \tag{A.11}$$

Combining (A.9) and (A.11), $C_2^{(0)} = (1/2)\beta^2\mathbb{E}\{\gamma^2\}-1$ results. Replacing it in (A.6), final first-order corrected expression when fitting $f(\gamma)$ as an exponential RV with parameter $\beta^{-1} = \mathbb{E}\{\gamma\}$ results as follows:

$$\begin{gathered} f(\gamma) \approx \beta e^{-\beta\gamma}\left(a_2\gamma^2 + a_1\gamma + a_0\right), \\ a_2 = \frac{1}{4}\beta^4\mathbb{E}\{\gamma^2\} - \frac{1}{2}\beta^2,\ \ a_1 = -\beta^3\mathbb{E}\{\gamma^2\}+2\beta,\ \ a_0 = \frac{1}{2}\beta^2\mathbb{E}\{\gamma^2\}.\end{gathered} \tag{A.12}$$

A.2. First-Order Correction for χ^2-Distribution with Four Degrees of Freedom PDF Approximation. In this section, we work on the first-order error corrected version for $f(\gamma)$ when fading figure $\mathcal{F} \approx 2$. Again, this approximation is obtained retaining the first nonzero term of the sum in (A.5) considering $\alpha = 1$ and $\beta = 2/\eta$. Note that now the most suitable χ^2-distribution to approximate $f(\gamma)$ should have $r = 4$ degrees of freedom. Therefore, approximation

$$f(\gamma) \approx \beta^2\gamma e^{-\beta\gamma}\left[1 + C_2^{(1)}L_2^{(1)}(\beta\gamma)\right] \tag{A.13}$$

results. One more time, it is possible to derive from (A.2) that

$$L_2^{(1)}(\beta\gamma) = \frac{1}{2}\left[(\beta\gamma)^2 - 6(\beta\gamma) + 6\right]. \tag{A.14}$$

Keeping in mind that we need to obtain $C_2^{(1)}$, it can be observed that

$$\begin{aligned}\int_0^\infty & L_2^{(1)}(\beta\gamma)\varepsilon(\gamma)d\gamma \\ &= \int_0^\infty L_2^{(1)}(\beta\gamma)\beta^2\gamma e^{-\beta\gamma}\sum_{k=2}^{\infty}\left[C_k^{(1)}L_k^{(1)}(\beta\gamma)\right]d\gamma \\ &= C_2^{(1)}\int_0^\infty L_2^{(1)}(\beta\gamma)L_2^{(1)}(\beta\gamma)\beta^2\gamma e^{-\beta\gamma}d\gamma \\ &\quad+\sum_{k=3}^{\infty}C_k^{(1)}\int_0^\infty L_2^{(1)}(\beta\gamma)L_k^{(1)}(\beta\gamma)\beta^2\gamma e^{-\beta\gamma}d\gamma.\end{aligned} \tag{A.15}$$

According to the orthogonality property introduced in (A.4), the integral in the first term is now equal to 3, while all the other integrals inside the sum remain null. Based on this, we find that

$$\int_0^\infty L_2^{(1)}(\beta\gamma)\varepsilon(\gamma)d\gamma = 3C_2^{(1)}. \tag{A.16}$$

As an alternative approach, we now replace $L_2^{(1)}(\beta\gamma)$ by expression (A.14). Therefore,

$$\begin{aligned}\int_0^\infty L_2^{(1)}(\beta\gamma)\varepsilon(\gamma)d\gamma = \frac{1}{2}\int_0^\infty(\beta\gamma)^2\varepsilon(\gamma)d\gamma - 3\int_0^\infty(\beta\gamma)\varepsilon(\gamma)d\gamma \\ +3\int_0^\infty\varepsilon(\gamma)d\gamma.\end{aligned} \tag{A.17}$$

The last two integrals vanish because the moments of $\varepsilon(\gamma)$ of order up to 1 are null. Therefore,

$$\begin{aligned}\int_0^\infty & L_2^{(1)}(\beta\gamma)\varepsilon(\gamma)d\gamma \\ &= \frac{1}{2}\left[\int_0^\infty(\beta\gamma)^2 f(\gamma)d\gamma - \int_0^\infty(\beta\gamma)^2\beta^2\gamma e^{-\beta\gamma}\,d\gamma\right] \\ &= \frac{\beta^2}{2}\left[\mathbb{E}\{\gamma^2\} - \frac{6}{\beta^2}\right].\end{aligned} \tag{A.18}$$

Combining (A.16) and (A.18), we are able to arrive at $C_2^{(1)} = (1/6)\beta^2\mathbb{E}\{\gamma^2\} - 1$. Replacing this value in (A.13) allows us to conclude that the first-order error corrected version when approximating $f(\gamma)$ with a χ^2-distribution with 4 degrees of freedom, and parameter $\beta^{-1} = (1/2)\mathbb{E}\{\gamma\}$ is equal to

$$\begin{aligned} f(\gamma) &\approx \beta^2 \gamma e^{-\beta\gamma}\left(b_2\gamma^2 + b_1\gamma + b_0\right), \\ b_2 &= \frac{1}{12}\beta^4\mathbb{E}\{\gamma^2\} - \frac{1}{2}\beta^2, \\ b_1 &= -\frac{1}{2}\beta^3\mathbb{E}\{\gamma^2\} + 3\beta, \\ b_0 &= \frac{1}{2}\beta^2\mathbb{E}\{\gamma^2\} - 2. \end{aligned} \tag{A.19}$$

B. SNR Gains and Fading Figures for W-CDMA Closed-Loop Transmit-Diversity Modes

Let us first compute the SNR gains when the best/worst Tx weight is selected for transmission. Working on (11) taking into account that $|w_{1,k}| = \sqrt{1-\widehat{\alpha}_k^2}$ and $|w_{2,k}| = \widehat{\alpha}_k$,

$$\begin{aligned} \mathbb{E}\{X_k\} &= \left(1-\widehat{\alpha}_k^2\right)\mathbb{E}\left\{|h_k|_{(2)}^{2}\right\} + \widehat{\alpha}_k^2\mathbb{E}\left\{|h_k|_{(1)}^{2}\right\} \\ &\quad + 2\sqrt{1-\widehat{\alpha}_k^2}\,\widehat{\alpha}_k\mathbb{E}\left\{|h_k|_{(1)}|h_k|_{(2)}\right\}\mathbb{E}\{\cos\varphi_k\}, \\ \mathbb{E}\{Y_k\} &= \widehat{\alpha}_k^2\mathbb{E}\left\{|h_k|_{(2)}^{2}\right\} + \left(1-\widehat{\alpha}_k^2\right)\mathbb{E}\left\{|h_k|_{(1)}^{2}\right\} \\ &\quad - 2\sqrt{1-\widehat{\alpha}_k^2}\,\widehat{\alpha}_k\mathbb{E}\left\{|h_k|_{(1)}|h_k|_{(2)}\right\}\mathbb{E}\{\cos\varphi_k\} \end{aligned} \tag{B.1}$$

result, where $|h_k|_{(1)}$ and $|h_k|_{(2)}$ are the first- and second-order statistic of two i.i.d. Rayleigh RVs with mean $\mathbb{E}\{|h_k|\} = \sqrt{\pi/4}$, respectively. Similarly, in case of second-order raw moments,

$$\begin{aligned} &\mathbb{E}\{X_k^2\} \\ &\quad = \left(1-\widehat{\alpha}_k^2\right)^2\mathbb{E}\left\{|h_k|_{(2)}^{4}\right\} + 2\left(1-\widehat{\alpha}_k^2\right)\widehat{\alpha}_k^2\mathbb{E}\left\{|h_k|_{(1)}^{2}|h_k|_{(2)}^{2}\right\} \\ &\qquad \times\left[1+2\mathbb{E}\{\cos^2\varphi_k\}\right] + 4\sqrt{1-\widehat{\alpha}_k^2}\,\widehat{\alpha}_k\mathbb{E}\{\cos\varphi_k\} \\ &\qquad \times\left[\left(1-\widehat{\alpha}_k^2\right)\mathbb{E}\left\{|h_k|_{(1)}|h_k|_{(2)}^{3}\right\} + \widehat{\alpha}_k^2\mathbb{E}\left\{|h_k|_{(1)}^{3}|h_k|_{(2)}\right\}\right] \\ &\qquad + \widehat{\alpha}_k^4\mathbb{E}\left\{|h_k|_{(1)}^{4}\right\}, \\ &\mathbb{E}\{Y_k^2\} \\ &\quad = \widehat{\alpha}_k^4\mathbb{E}\left\{|h_k|_{(2)}^{4}\right\} + 2\left(1-\widehat{\alpha}_k^2\right)\widehat{\alpha}_k^2\mathbb{E}\left\{|h_k|_{(1)}^{2}|h_k|_{(2)}^{2}\right\} \\ &\qquad \times\left[1+2\mathbb{E}\{\cos^2\varphi_k\}\right] - 4\sqrt{1-\widehat{\alpha}_k^2}\,\widehat{\alpha}_k\mathbb{E}\{\cos\varphi_k\} \\ &\qquad \times\left[\widehat{\alpha}_k^2\mathbb{E}\left\{|h_k|_{(1)}|h_k|_{(2)}^{3}\right\} + \left(1-\widehat{\alpha}_k\right)^2\mathbb{E}\left\{|h_k|_{(1)}^{3}|h_k|_{(2)}\right\}\right] \\ &\qquad + \left(1-\widehat{\alpha}_k^2\right)^2\mathbb{E}\left\{|h_k|_{(1)}^{4}\right\} \end{aligned} \tag{B.2}$$

result. In this situation, RV φ_k is uniformly distributed on $[-\pi/2^{N_p}, \pi/2^{N_p})$, where N_p is the number of bits used to quantize phase angles. Therefore, $\mathbb{E}\{\cos\varphi_k\} = (4/\pi)\sin(\pi/4)$ and $\mathbb{E}\{\cos^2\varphi_k\} = 1/2 + 1/\pi$ in case of CL TD mode 1, and $\mathbb{E}\{\cos\varphi_k\} = (8/\pi)\sin(\pi/8)$ and $\mathbb{E}\{\cos^2\varphi_k\} = (1/2) + \sqrt{2}/\pi$ in case of CL TD mode 2. Similarly, $\widehat{\alpha}_k = \sqrt{0.5}$ and $\widehat{\alpha}_k = \sqrt{0.2}$ for both CL TD modes 1 and 2, respectively. Single raw moments and product raw moments of the order statistics of RV $|h_k|$ are obtained by using recurrence relations (13.4) and (13.7) of [24]. Replacing all these results, we find that

$$\begin{aligned} \mathcal{G}_x &= 1 + \sqrt{\frac{1}{2}}, \\ \mathcal{G}_y &= 1 - \sqrt{\frac{1}{2}}, \\ \mathbb{E}\{X_k^2\} &= 2 + \frac{1}{\pi} + 3\sqrt{\frac{1}{2}}, \\ \mathbb{E}\{Y_k^2\} &= 2 + \frac{1}{\pi} - 3\sqrt{\frac{1}{2}}, \end{aligned} \tag{B.3}$$

assuming CL TD mode 1, and

$$\begin{aligned} \mathcal{G}_x &= 1.3 + 1.6\sin\left(\frac{\pi}{8}\right), \\ \mathcal{G}_y &= 0.7 - 1.6\sin\left(\frac{\pi}{8}\right), \\ \mathbb{E}\{X_k^2\} &= 2.9 + \left(4.8 + \frac{3.84}{\pi}\right)\sin\frac{\pi}{8} + \frac{1.28}{\pi}\sqrt{\frac{1}{2}}, \\ \mathbb{E}\{Y_k^2\} &= 1.1 - \left(4.8 - \frac{3.84}{\pi}\right)\sin\frac{\pi}{8} + \frac{1.28}{\pi}\sqrt{\frac{1}{2}}, \end{aligned} \tag{B.4}$$

considering CL TD mode 2. Combining these raw moments in (8), fading figure expressions

$$\mathcal{F}_x = \frac{3/2 + 2\sqrt{1/2}}{1/2 + 1/\pi + \sqrt{1/2}}, \qquad \mathcal{F}_y = \frac{3/2 - 2\sqrt{1/2}}{1/2 + 1/\pi - \sqrt{1/2}}, \tag{B.5}$$

result assuming CL TD mode 1, and

$$\begin{aligned} &\mathcal{F}_x \\ &\quad = \frac{2.97 + 4.16\sin(\pi/8) - 1.28\sqrt{1/2}}{-0.07 + 3.84(1/6 + 1/\pi)\sin(\pi/8) + 1.28(1 + 1/\pi)\sqrt{1/2}}, \\ &\mathcal{F}_y \\ &\quad = \frac{1.77 - 2.24\sin(\pi/8) - 1.28\sqrt{1/2}}{-0.67 + 3.84(-2/3 + 1/\pi)\sin(\pi/8) + 1.28(1 + 1/\pi)\sqrt{1/2}}, \end{aligned} \tag{B.6}$$

considering CL TD mode 2.

C. Useful Closed-Form Expression

Our aim is to compute the integral

$$\begin{aligned} L_n(\beta, c) &= \int_0^\infty \log_e(\gamma + c)\beta(\beta\gamma)^n e^{-\beta\gamma}\,d\gamma, \\ & n = 0, 1, \ldots; \quad \beta > 0; \quad c > 0. \end{aligned} \tag{C.1}$$

Let us use formula (8.356.4) of [25] and integrate (C.1) by parts. Then, we find that

$$L_n(\beta,c) = n!\log_e(c) + \int_0^\infty \frac{\Gamma(n+1,\beta\gamma)}{\gamma+c}d\gamma. \tag{C.2}$$

Here, we have by (6.5.3), (6.5.2), (6.5.13), and (6.5.11) of [19] that

$$\int_0^\infty \frac{\Gamma(n+1,\beta\gamma)}{\gamma+c}d\gamma = n!\sum_{k=0}^{n}\frac{1}{k!}\int_0^\infty \frac{(\beta\gamma)^k e^{-\beta\gamma}}{\gamma+c}d\gamma. \tag{C.3}$$

Then, by using (3.383.10) of [25] and (6.5.9) of [19], we obtain

$$\int_0^\infty \frac{(\beta\gamma)^k e^{-\beta\gamma}}{\gamma+c}d\gamma = k!e^{\beta c}E_{k+1}(\beta c). \tag{C.4}$$

After combining the last three formulas, we get the desired result:

$$\begin{gathered} L_n(\beta,c) = n!\left[\log_e(c) + e^{\beta c}\sum_{k=0}^{n}E_{k+1}(\beta c)\right], \\ n = 0,1,\ldots; \quad \beta > 0; \quad c > 0. \end{gathered} \tag{C.5}$$

References

[1] R. W. Heath Jr., M. Airy, and A. J. Paulraj, "Multiuser diversity for MIMO wireless systems with linear receivers," in *Proceedings of the 35th Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1194–1199, Pacific Grove, Calif, USA, November 2001.

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

[3] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Tech. Rep. TR-301, Digital Equipment Corporation, Maynard, Mass, USA, September 1984.

[4] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[5] H. Viswanathan, S. Venkatesan, and H. Huang, "Downlink capacity evaluation of cellular networks with known-interference cancellation," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 802–811, 2003.

[6] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, 2006.

[7] A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, "Efficient use of side information in multiple-antenna data transmission over fading channels," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1423–1436, 1998.

[8] D. J. Love, R. W. Heath Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2735–2747, 2003.

[9] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2562–2579, 2003.

[10] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1478–1491, 2007.

[11] 3GPP, "Physical layer procedures (FDD)," Technical Specification TS 25.214, V5.6.0, 3GPP, Valbonne, France, 2003.

[12] A. A. Dowhuszko, G. Corral-Briones, J. Hämäläinen, and R. Wichman, "Achievable sum-rate analysis of practical multiuser scheduling schemes with limited feedback," in *Proceedings of IEEE International Conference on Communications (ICC '07)*, pp. 4381–4386, Glasgow, Scotland, June 2007.

[13] 3GPP, "Physical layer aspects of UTRA high speed downlink packet access," Tech. Rep. TSG-RAN TR 25.858 V 5.0.0, 3GPP, Valbonne, France, 2002.

[14] R. J. McEliece and W. E. Stark, "Channels with block interference," *IEEE Transactions on Information Theory*, vol. 30, no. 1, pp. 44–53, 1984.

[15] S. Sanayei and A. Nosratinia, "Opportunistic beamforming with limited feedback," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 2765–2771, 2007.

[16] J. Diaz, O. Simeone, and Y. Bar-Ness, "Sum-rate of MIMO broadcast channels with one bit feedback," in *Proceedings of IEEE International Symposium on Information Theory (ISIT '06)*, pp. 1944–1948, Seattle, Wash, USA, July 2006.

[17] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proceedings of the 51st IEEE Vehicular Technology Conference (VTC '00)*, vol. 3, pp. 1854–1858, Tokyo, Japan, May 2000.

[18] M. Nakagami, "The m-distribution—a general formula for intensity distribution of rapid fading," in *Statistical Methods in Radio Wave Propagation*, pp. 581–635, McGraw-Hill, New York, NY, USA, 1958.

[19] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, USA, 1970.

[20] A. A. Dowhuszko, G. Corral-Briones, J. Hämäläinen, and R. Wichman, "Outage probability analysis of practical multiuser scheduling schemes with limited feedback," in *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC '07)*, pp. 1036–1040, Dublin, Ireland, April 2007.

[21] S. Zhou, Z. Wang, and G. B. Giannakis, "Quantifying the power loss when transmit beamforming relies on finite-rate feedback," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1948–1957, 2005.

[22] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling policies for wireless systems with short term fairness constraints," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '03)*, vol. 1, pp. 533–537, San Francisco, Calif, USA, December 2003.

[23] V. Hassel, M. R. Hanssen, and G. E. Øien, "Spectral efficiency and fairness for opportunistic round robin scheduling," in *Proceedings of IEEE International Conference on Communications (ICC '06)*, vol. 2, pp. 784–789, Istanbul, Turkey, June 2006.

[24] N. Balakrishnan and C. Rao, Eds., *Handbook of Statistics 16: Order Statistics: Theory & Methods*, Elsevier, Amsterdam, The Netherlands, 1998.

[25] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, NY, USA, 2007.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 271540, 13 pages
doi:10.1155/2009/271540

Research Article

Throughput versus Fairness: Channel-Aware Scheduling in Multiple Antenna Downlink

Eduard A. Jorswieck,[1] Aydin Sezgin,[2] and Xi Zhang[3]

[1] *Communications Laboratory, Faculty of Electrical Engineering and Information Technology, Dresden University of Technology, D-01062 Dresden, Germany*
[2] *Department of Electrical Engineering & Computer Science, Henry Samueli School of Engineering, University of California, Irvine, CA 92697, USA*
[3] *ACCESS Linnaeus Center, Royal Institute of Technology, SE-100 44 Stockholm, Sweden*

Correspondence should be addressed to Eduard A. Jorswieck, jorswieck@ifn.et.tu-dresden.de

Received 1 July 2008; Accepted 23 December 2008

Recommended by Alagan Anpalagan

Channel aware and opportunistic scheduling algorithms exploit the channel knowledge and fading to increase the average throughput. Alternatively, each user could be served equally in order to maximize fairness. Obviously, there is a tradeoff between average throughput and fairness in the system. In this paper, we study four representative schedulers, namely the maximum throughput scheduler (MTS), the proportional fair scheduler (PFS), the (relative) opportunistic round robin scheduler (ORS), and the round robin scheduler (RRS) for a space-time coded multiple antenna downlink system. The system applies TDMA based scheduling and exploits the multiple antennas in terms of spatial diversity. We show that the average sum rate performance and the average worst-case delay depend strongly on the user distribution within the cell. MTS gains from asymmetrical distributed users whereas the other three schedulers suffer. On the other hand, the average fairness of MTS and PFS decreases with asymmetrical user distribution. The key contribution of this paper is to put these tradeoffs and observations on a solid theoretical basis. Both the PFS and the ORS provide a reasonable performance in terms of throughput and fairness. However, PFS outperforms ORS for symmetrical user distributions, whereas ORS outperforms PFS for asymmetrical user distribution.

1. Introduction

The optimal strategy for maximizing the sum capacity with perfect channel state information (CSI) of a cellular single-input single-output (SISO) multiuser channel is to allow only the user having the best channel conditions in terms of SNR to transmit at each time slot (TDMA). This result in [1] has induced the notion of multiuser diversity [2], that is, the achievable capacity of the system increases with the number of the users. The corresponding scheduling policy is called maximum throughput scheduler (MTS). Subsequently, TDMA-based channel-aware scheduling schemes which consider temporal fairness [3] or stringent rate constraints under energy efficiency [4] are developed.

A major disadvantage of MTS is its unfairness toward users at the cell edge. On the other hand, the most fair but channel unaware scheduler is the round robin scheduler (RRS) [5], that is, all transmissions take place in a strict numerical order. The MTS and RRS leave room for various channel aware schedulers that lie in between these two. In order to increase the fairness for users at the cell edge, the so-called proportional fair scheduler (PFS) can be applied. The PFS weights the instantaneous transmission rates by their averages to find the best user and achieves equal activity probability for all users [6]. Yet another scheduler, which is referred to as opportunistic round robin scheduling (ORS), was introduced in [7]. It is a combination of the RRS and MTS. The comparison of different schedulers with respect to different performance criteria is a highly viable research area. For instance, in [8], the throughput guarantee violation probability is approximated and simulated for different schedulers in different channel models. The asymptotic throughput of channel-aware schedulers is analyzed in [9].

In order to quantitatively measure the impact of the scheduler on the fairness, different measures are proposed in the literature [10–12]. The Jain fairness index (JFI) defined in [10], also known as the global fairness index (GFI) [13], provides a single number between zero and one that measures the fairness even for resource scheduling in finite windows. The average fairness defined in [11] is developed from an information theoretic point of view. The worst-case delay as it is used in, for example, [12] measures the average number of transmissions needed until all users were active at least m times.

Obviously, there exists a tradeoff between average throughput and average fairness [14]. In this paper, we study this tradeoff for the four scheduling algorithms MTS, RRS, PFS, and ORS. The main novelty lies in the systematic approach to this problem using majorization theory. This tool helps understanding the impact of user distributions within the cell on the system performance and on the average worst-case delay. The application of majorization theory allows to analytically and qualitatively assess the advantages and disadvantages of the four channel-aware schedulers. The contributions of the paper are as follows.

(1) In Section 2.5, closed form expressions for the four scheduler for arbitrary nonsymmetrical user distributions are derived.

(2) The impact of the user distribution on the average sum rate is analyzed in Section 3, and it is shown that the average sum rate is increased with asymmetrical user distributions for MTS. For all other schedulers (RRS, PFS, and ORS), it decreases.

(3) Different fairness measures and their properties are discussed in Section 4. Furthermore, we study the impact of the user distribution and its connection to the service probabilities.

(4) The asymptotic performance for high SNR or large number of users is analyzed in Section 5.

(5) In Section 6, the sum rate of MTS, RRS, and PFS under a fixed rate constraint is derived, and the impact of user distributionis characterized.

(6) In Section 7, we illustrate the theoretical results with numerical single-cell multiuser simulations.

The paper is concluded in Section 7. Parts of the results for single-antenna transmitter are presented without proofs in [15]. The impact of interferer locations on the downlink performance of the system is studied in [16].

2. System Model and Preliminaries

In this section, we present the system model, the channel model, the measure of the user distribution based on majorization, the high-SNR performance measures, and the four scheduler. Our approach to the cross-layer analysis of these scheduling algorithms is physical layer oriented.

2.1. System Model. In the signal model, there are K mobile users which are served by a base station in downlink transmission. The base station has multiple antennas (n_T), the mobiles have one antenna each. Denote the channels to the users as $\mathbf{h}_1, \ldots, \mathbf{h}_K$. The base applies an OSTBC [17, 18] in order to exploit spatial diversity without spatial feedback overhead. Spatial feedback contains information about the spatial signatures of the user channels, whereas channel quality information contains scalar values . The data stream vectors $\mathbf{d}_1, \ldots, \mathbf{d}_K$ of dimension $1 \times M$ of the K users are weighted by a power allocation $p_1, \ldots, p_K$ and added before they come into the OSTBC as $\hat{x}_1, \ldots, \hat{x}_M$. The output of the OSTBC is a vector $\mathbf{x} = [x_1, \ldots, x_{n_T}]$ of dimension $1 \times n_T$ (compare to system model in [19]). The code rate is given by $r_c = M/n_T$. Note that the framework can be extended also to other code classes [20].

Each mobile first performs channel matched filtering according to the effective OSTBC channel. Afterward, the received signal at user k of stream n is given by

$$y_{k,n} = a_k \sum_{l=1}^{K} \overline{x}_{l,n} + n_{k,n}, \quad 1 \le n \le M, \tag{1}$$

with fading coefficients $\alpha_k = a_k^2 = \|\mathbf{h}_k\|^2/n_T$, transmit stream n intended for user l as $\overline{x}_{l,n}$ and noise for stream n as $n_{k,n}$. There are M parallel streams for each mobile. However, all streams have the same properties in terms of a_k and noise statistics. Therefore, we restrict our attention without loss of generality to the first stream $n = 1$ and omit the index in the following. Let p_k be the power allocated to user k within one block, that is, $p_k = \mathbb{E}[|x_k|^2]$. We assume a short-term power constraint, that is, $\sum_{k=1}^{K} p_k \le P$. The noise power at the receivers is σ^2. The transmit power is distributed uniformly over the n_T transmit antennas, and each data stream has an effective power p_k/n_T. We incorporate this weighting into the transmit SNR given by $\rho = P/n_T\sigma^2$.

The mobiles feed back their scalar channel quality indicators, that is, their fading coefficient $a_1, \ldots, a_K$ to the base and we assume these numbers are perfectly known at the base station. As such, the base has perfect information about the channel norm but not about the complete fading vectors.

2.2. Channel Model. The channel vectors $\mathbf{h}_1, \ldots, \mathbf{h}_K$ are modeled as independently zero-mean complex Gaussian distributed vectors with covariance matrix $c_k\mathbf{I}$ in rich multipath environment. The variance c_k depends mainly on the distance of the user to the base, and it is called average channel power. Therefore, the fading coefficients $\alpha_1, \ldots, \alpha_K$ are independently χ^2-distributed with n_T complex degrees of freedom weighted by the average channel power $c_1, \ldots, c_K$, that is, using independent standard $\chi^2_{n_T}$-distributed random variables $w_1, \ldots, w_K$, the fading coefficients are expressed as $\alpha_k = c_k w_k$.

2.3. Measure of User Distribution. The distance of the mobile k to the base station is determined by the average channel power c_k. In the following, we refer to the vector of average channel powers $\mathbf{c} = [c_1, \ldots, c_K]$ as the user distribution. In order to guarantee a fair comparison between different user distributions, we constrain the sum variance to be equal to the number of users, that is, $\sum_{k=1}^{K} c_k = K$. Without loss of generality, we order the users in a nonincreasing way according to their fading variances, that is, $c_1 \geq c_2 \geq \cdots \geq c_K$. The constraint regarding the sum of the fading variances verifies that we compare scenarios in which the channel carries the same average sum power. We need the following definitions [21].

Definition 1. For two vectors $\mathbf{x}, \mathbf{y} \in R^n$, one says that the vector $\mathbf{x}$ majorizes the vector $\mathbf{y}$ and writes $\mathbf{x} \succ \mathbf{y}$ if $\sum_{k=1}^{m} x_k \geq \sum_{k=1}^{m} y_k$ for $m = 1, \ldots, n-1$ and $\sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y_k$ (note that sometimes majorization is defined by the sum of the *smallest m* components [22]).

The next definition describes a function Φ which is applied to the vectors $\mathbf{x}$ and $\mathbf{y}$ with $\mathbf{x} \succ \mathbf{y}$.

Definition 2. A real-valued function Φ defined on $\mathcal{A} \subset R^n$ is said to be *Schur convex* on $\mathcal{A}$ if from $\mathbf{x} \succ \mathbf{y}$ on $\mathcal{A}$ follows $\Phi(\mathbf{x}) \geq \Phi(\mathbf{y})$. Similarly, Φ is said to be *Schur concave* on $\mathcal{A}$ if from $\mathbf{x} \succ \mathbf{y}$ on $\mathcal{A}$ follows $\Phi(\mathbf{x}) \leq \Phi(\mathbf{y})$.

Majorization is a useful tool to study the impact of vectors which can be partially ordered. The common monotony properties of scalar functions correspond to the Schur-convex property of vector functions. The reason for the term "Schur-convex" instead of "Schur-monotone" is that every symmetric and convex vector function is Schur-convex. Majorization is a large and active area of research in linear algebra, with entire books [21] devoted to its theory and application.

It is worth mentioning that majorization induces only a partial order on vectors with more than two components, that is, not all possible vectors can be compared with each other. This is due to the fact that vectors with more than two components cannot be totally ordered. However, a sufficient number of vectors can be compared. Also, the extreme cases can be used for comparison with any other vector. For more information about this measure of user distribution and its application see [23, Section 4.2.1].

2.4. High-SNR Measures $\mathcal{S}_\infty$ and $\mathcal{L}_\infty$. The quantitative performance is analyzed using the high-SNR offset concept from [24]. Denote by $C(\rho)$ the average throughput as a function of the SNR. The two high-SNR measures are introduced as follows:

$$\mathcal{S}_\infty = \lim_{\rho \to \infty} \frac{C(\rho)}{\log(\rho)},$$
$$\mathcal{L}_\infty = \lim_{\rho \to \infty} \left(\log(\rho) - \frac{C(\rho)}{\mathcal{S}_\infty} \right). \tag{2}$$

The measures $\mathcal{S}_\infty$ and $\mathcal{L}_\infty$ are referred to as high-SNR slope and the high-SNR power offset, respectively. At high SNR, the average throughput behaves like $C(\rho) = \mathcal{S}_\infty((\rho[\mathrm{dB}]/3\mathrm{dB}) - \mathcal{L}_\infty) + O(1)$. For convenience, these high-SNR measures are defined in 3 dB units. For further discussion, see [24, Section 2]. These two high-SNR measures are useful if two systems are compared which differ either in their multiplexing gain, that is, the slope of the average throughput curve at high SNR, or which have equal $\mathcal{S}_\infty$ but are shifted at high SNR.

2.5. Types of (Channel Aware) Scheduling. Since the base station has only partial CSI in form of the channel norm, we restrict all scheduling strategies to TDMA-based scheduling. From the single-antenna downlink, it is well known that if perfect CSI is available at the base station, the sum rate is maximized by single-user transmission to the best user only [1], that is, TDMA achieves the sum capacity. This result leads to the notion of multiuser diversity and the concept of opportunistic communication [2]. This scheduler is called MTS, and the achievable average sum rate is given by

$$R_{\text{sum}}^{\text{MT}} = \mathbb{E}\left[\log\left(1 + \rho \max_{1 \leq k \leq K} \|\mathbf{h}_k\|^2\right) \right]. \tag{3}$$

Note that the average sum rate of the MTS can be written in integral representation as

$$R_{\text{sum}}^{\text{MT}} = \int_0^\infty \frac{\rho}{1+\rho t} \left[1 - \prod_{k=1}^{K} \left(1 - \frac{\Gamma(n_T, (t/c_k))}{\Gamma(n_T)} \right) \right] dt, \tag{4}$$

using the incomplete gamma function $\Gamma(a, z) = \int_z^\infty \exp(-t) t^{a-1} dt$. The case with single-antenna base and symmetrically distributed users ($\mathbf{c} = \mathbf{1}$) is studied in [25]. The MTS is unfair from a user perspective because mobiles at the cell edge have less probability to be served.

The opposite type of scheduler is the round robin scheduler (RRS). It is not channel aware but it minimizes the average worst-case delay, that is, the average time until every user has been served at least once. The average sum rate is given by

$$\begin{aligned} R_{\text{sum}}^{\text{RR}} &= \mathbb{E}\left[\frac{1}{K} \sum_{k=1}^{K} \log\left(1 + \rho \|\mathbf{h}_k\|^2\right) \right] \\ &= \mathbb{E}\left[\frac{1}{K} \sum_{k=1}^{K} \log\left(1 + \rho c_k w_k\right) \right]. \end{aligned} \tag{5}$$

Note that (5) can be rewritten for $n_T = 1$ in closed form as

$$R_{\text{sum}}^{\text{RR}} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{Ei}\left(1, \frac{1}{\rho c_k}\right) \exp\left(\frac{1}{\rho c_k}\right), \tag{6}$$

where the exponential integral is given by $\mathrm{Ei}(a, x) = \int_1^\infty \exp(-tx) t^{-a} dt$.

These two schedulers are the two most extreme cases. The MTS maximizes the average sum rate, whereas the RRS minimizes the average worst-case delay. A compromise between the two is the proportional fair scheduler (PFS) [2]. For the analysis, we use the so-called relative SNR scheduler. The user is served which has the highest ratio of the instantaneous rate to average rate. Hence, the achievable sum rate is given by

$$R_{\text{sum}}^{\text{PF}} = \mathbb{E}\left[\log\left(1 + \rho\|\mathbf{h}_{k^*}\|^2\right)\right] \\ \text{with } k^* = \arg \max_{1 \le k \le K} \frac{\|\mathbf{h}_k\|^2}{c_k}. \tag{7}$$

In reality, the average transmission rate is updated from transmission interval to transmission interval. Here, we use the ergodic formulation of the scheduler (let the window length $t_c \to \infty$). Note that (7) can be rewritten as

$$R_{\text{sum}}^{\text{PF}} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\log\left(1 + \rho c_k \max_{1 \le l \le K} w_l\right)\right], \tag{8}$$

because the scheduling probability of all users is equal to $1/K$. For $n_T = 1$, (8) can be rewritten in closed form as

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{K}(-1)^{l-1}\binom{K}{l}\operatorname{Ei}\left(1, \frac{l}{\rho c_k}\right)e^{(l/\rho c_k)}. \tag{9}$$

Another interesting channel-aware scheduler is proposed in [7]. The one-round version [26] of the relative opportunistic round robin scheduler (ORS) guarantees the same average worst-case delay as the RRS but exploits a certain amount of multiuser diversity. It consists of K rounds and initializes the set of available users $\mathscr{S}$ with $\mathscr{S} = \{1, \ldots, K\}$. Within each step, the relative best user $\max_{k \in \mathscr{S}}(\|h_k\|^2/c_k)$ out of the set of available users is picked and removed from the set. After K steps, it is guaranteed that all users were active at least once.

For our analysis, we need the representation in the following lemma.

Lemma 1. *The average sum rate of the ORS* (13) *can be written as*

$$R_{\text{sum}}^{\text{OR}} = \int_0^{\infty}\left[1 - \frac{1}{K^2}\sum_{n=1}^{K}\sum_{i=1}^{K}\left(1 - \frac{\Gamma(n_T, (t/c_i))}{\Gamma(n_T)}\right)^n\right] \\ \cdot \frac{\rho}{1+\rho t}dt. \tag{10}$$

Proof. The CDF of the relative ORS is derived for $n_T = 1$ in [27, Equation (6)] and is given by

$$P(t) = \frac{1}{K^2}\sum_{n=1}^{K}\sum_{i=1}^{K}\left(1 - e^{-(t/c_i)}\right)^n. \tag{11}$$

For general $n_T > 1$, it reads

$$P(t) = \frac{1}{K^2}\sum_{n=1}^{K}\sum_{i=1}^{K}\left(1 - \frac{\Gamma(n_T, (t/c_i))}{\Gamma(n_T)}\right)^n. \tag{12}$$

We use the integration by parts rule $\int_a^b f(x)g'(x)dx = |f(x)g(x)|_a^b - \int_a^b f'(x)g(x)dx$. Now, identify $f(x) = \log(1 + \rho x)$ and $g(x)' = p(x)$, respectively, with the pdf of the relative ORS $p(x)$. Choose carefully $g(x) = P(x) - 1$ to assure existence of the first part. Then, we obtain finally the representation in (10). □

The sum rate performance for $n_T = 1$ can be further simplified as in [27, Equation (8)] to obtain the closed form expression

$$R_{\text{sum}}^{\text{OR}} = \frac{1}{K^2}\sum_{n=1}^{K} n \sum_{i=1}^{K}\sum_{j=0}^{n-1}\binom{n-1}{j}(-1)^j \\ \cdot \frac{e^{((1+j)/c_i)}}{1+j}\operatorname{Ei}\left(1, \frac{1+j}{c_i}\right). \tag{13}$$

With the sum rate expressions in (4), (5), (8), and (10), we are now ready for the analysis of the user distribution $\mathbf{c}$ in the next section.

3. Analysis of Sum Rate Performance

In this section, we analyze the impact of the user distribution on the sum rate performance of the four scheduler. One main question is whether the standard assumption about a symmetric user distribution, which is made often for simplification, leads to an upper or lower bound on the real system throughput. First, we present the theoretical results, and then we discuss their meaning in the paper context.

3.1. Schur-Convexity and Schur-Concavity Properties. The following result is provided in [28] for $n_T = 1$ and restated and proved here for $n_T > 1$. It states that a more asymmetrical user distribution increases the average sum rate with MTS.

Theorem 1. *Let* $\mathbf{c}$ *and* $\mathbf{d}$ *be two different average user powers. The average sum rate of the MTS is Schur-convex with respect to user powers* $\mathbf{c}$ *and* $\mathbf{d}$, *that is,*

$$\mathbf{c} \succcurlyeq \mathbf{d} \Longrightarrow R_{\text{sum}}^{\text{MT}}(\mathbf{c}) \ge R_{\text{sum}}^{\text{MT}}(\mathbf{d}). \tag{14}$$

The proof can be found in [28, Theorem 1] for the single-antenna $n_T = 1$ case. We present in Appendix A the more general proof for convenience.

The impact of the user distribution on the performance of the RRS is analyzed in the next result.

Theorem 2. *The average sum rate of the RRS is Schur-concave with respect to the vector of average user powers* $\mathbf{c}$, *that is,*

$$\mathbf{c} \succcurlyeq \mathbf{d} \Longrightarrow R_{\text{sum}}^{\text{RR}}(\mathbf{c}) \le R_{\text{sum}}^{\text{RR}}(\mathbf{d}). \tag{15}$$

Proof. Define the average sum rate as a function of $\mathbf{c}$ as

$$R_{\text{sum}}^{\text{RR}}(\mathbf{c}) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\log\left(1 + \rho c_k w_k\right)\right], \tag{16}$$

and check Schur's condition [23] directly

$$\frac{\partial R_{\text{sum}}^{\text{RR}}(\mathbf{c})}{\partial c_1} - \frac{\partial R_{\text{sum}}^{\text{RR}}(\mathbf{c})}{\partial c_2} \\ = \mathbb{E}\left[\frac{\rho w_1}{1 + \rho c_1 w_1}\right] - \mathbb{E}\left[\frac{\rho w_2}{1 + \rho c_2 w_2}\right] \le 0. \tag{17}$$

□

The impact of the user distribution on the performance of PFS is derived analogously in Theorem 3.

Theorem 3. *The average sum rate of the PFS is Schur-concave with respect to the vector of average user powers* **c***, that is,*

$$\mathbf{c} \succcurlyeq \mathbf{d} \Longrightarrow R_{\text{sum}}^{\text{PF}}(\mathbf{c}) \le R_{\text{sum}}^{\text{PF}}(\mathbf{d}). \tag{18}$$

Proof. Start from the representation in (8) and check Schur's condition

$$\begin{aligned}\frac{\partial R_{\text{sum}}^{\text{PF}}(\mathbf{c})}{\partial c_1} &- \frac{\partial R_{\text{sum}}^{\text{PF}}(\mathbf{c})}{\partial c_2}\\ &= \frac{1}{K}\mathbb{E}\left[\frac{\rho c_1 \max_{1\le l\le K} w_l}{1+\rho c_1 \max_{1\le l\le K} w_l}\right]\\ &\quad - \frac{1}{K}\mathbb{E}\left[\frac{\rho c_2 \max_{1\le l\le K} w_l}{1+\rho c_2 \max_{1\le l\le K} w_l}\right] \le 0.\end{aligned} \tag{19}$$

□

Finally, the impact of the user distribution on the sum rate performance of ORS is characterized in the next result which is proved in Appendix B.

Theorem 4. *The average sum rate of the ORS is Schur-concave with respect to the vector of average user power* **c***, that is,*

$$\mathbf{c} \succcurlyeq \mathbf{d} \Longrightarrow R_{\text{sum}}^{\text{OR}}(\mathbf{c}) \le R_{\text{sum}}^{\text{OR}}(\mathbf{d}). \tag{20}$$

3.2. Discussion of Schur Properties. Let us restate the results from the last section in words. The sum rate of MTS improves with more asymmetrically distributed users. The sum rate of RRS, ORS, and PFS decreases with more asymmetrically users. Hence, the four results indicate that the common assumption about symmetrically distributed users leads to the following.

(1) A lower bound to the sum rate performance of MTS.

(2) An upper bound to the sum rate performance of RRS, ORS, and PFS.

This implies that a correct analysis even in terms of the sum rate does always require assumptions on the user distribution. In conclusion, there is only one scheduler which improves for asymmetrically distributed users, namely, the MTS. The average sum rates of the other scheduler, PFS, ORS, and RRS, decrease with more asymmetrically distributed user.

4. Fairness Analysis

In this section, the fairness properties of the four schedulers are analyzed. First, the average worst-case delay is proposed as a proper physical layer motivated delay measure. The impact of the service probabilities of the users on the worst-case delay is studied. Then, two other common fairness measures are reviewed, namely, Jain's fairness index and the dispersion. It is shown that all three measures are Schur-convex functions with respect to the service probabilities of the users. Finally, the connection between user distribution and service probability and delay is discussed.

4.1. Analysis of Average Worst-Case Delay. In order to capture the fairness of the different scheduler, the average worst-case delay is considered. The average worst-case delay $\mathbb{E}[D_{m,K}]$ measures the average number of transmissions that are needed until all K users have been active at least m times. We define $D_1 = \mathbb{E}[D_{1,K}]$.

The two most fair schedulers are the RRS and ORS. Both have an average worst-case delay of mK because all users are guaranteed to be active within a block of K transmissions. Especially, it takes K transmissions until every users has transmitted exactly once, that is,

$$D_1^{\text{RRS}} = D_1^{\text{ORS}} = K. \tag{21}$$

The PFS normalizes the users channels. Therefore, the probability that user k being active is, independently of k, $1 \le k \le K$, equal to $1/K$. Especially, it is independent of the user distribution **c**. The result from [29] applies for $m = 1$:

$$D_1^{\text{PFS}} = K\int_0^\infty 1 - (1-\exp(-x))^K dx. \tag{22}$$

Note that (22) can be written as

$$D_1^{\text{PFS}} = K(\Psi(K+1)+\gamma), \tag{23}$$

with the Ψ-function [30, 6.3] and Euler's constant γ [30, 6.1.3].

The analysis of the MTS is more difficult. Rewrite the average worst-case delay [12, Section 3.3] without dropping probability as

$$D_1^{\text{MTS}} = n\int_0^\infty \left(1 - \prod_{k=1}^{K}\left(1 - \frac{\Gamma(m, d_k t)}{\Gamma(m)}\right)\right) dt. \tag{24}$$

For $m = 1$, the expression in (24) says how many packets are transmitted on average until every user has at least transmitted one. The coefficients d_k in (24) are related to the probability that user k is chosen $\pi_k = d_k/K$. For the MTS, we prove the following result.

Theorem 5. *The average worst-case delay* $\mathbb{E}[D_{1,K}]$ *is Schur-convex with respect to* **d***, that is,*

$$\mathbf{d}_1 \succcurlyeq \mathbf{d}_2 \longrightarrow D_1^{\text{MTS}}(\mathbf{d}_1) \ge D_1^{\text{MTS}}(\mathbf{d}_2). \tag{25}$$

Proof. In order to check Schur's condition, [23] consider

$$\begin{aligned}\frac{\partial \mathbb{E}[D_{1,K}](\mathbf{d})}{\partial d_1} &- \frac{\partial \mathbb{E}[D_{1,K}](\mathbf{d})}{\partial d_2}\\ &= n\int_0^\infty \prod_{l=3}^{K}(1-\exp(-d_l t))g(t,d_1,d_2)dt,\end{aligned} \tag{26}$$

with $g(t,d_1,d_2) = t\exp(-d_2 t)(1-\exp(-d_1 t)) - t\exp(-d_1 t)(1-\exp(-d_2 t)) \ge 0$ for all $d_1 \ge d_2$, and $t \ge 0$. It follows that the integral in (24) is greater than or equal to zero. □

Theorem 5 formally states the intuitive fact that the average worst-case delay grows if some users are less frequent active on average. If the probability that user k is active is equal to $1/K$, independently of k, then the expression in (24) is minimized. Note that a similar analysis has been performed in the different context of birthday matching in [31].

4.2. Jain's Fairness Index and Dispersion. In [10], a quantitative measure of fairness is introduced. It is called Jain's fairness index (JFI) or global fairness index (GFI) [13]. Define x_k as the amount of a resource that is distributed to user k. Then, JFI is defined as [10, Equation (2)]

$$\mathrm{JFI} = \frac{\left((1/K)\sum_{k=1}^{K} x_k\right)^2}{(1/K)\sum_{k=1}^{K} x_k^2}. \tag{27}$$

Let us specialize this general definition to the case in which one resource is one transmission. The JFI is averaged over L transmissions [27]

$$\mathrm{JFI}(L) = \frac{\mathbb{E}_L\left((1/K)\sum_{k=1}^{K} x_k\right)^2}{\mathbb{E}_L(1/K)\sum_{k=1}^{K} x_k^2}. \tag{28}$$

Denote by π_k the probability that user k is active within L transmissions, then $x_k = \pi_k L$. Collect $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$. Let $L \to \infty$ to obtain the long-term average JFI as

$$\mathrm{JFI} = \frac{\left((1/K)\sum_{k=1}^{K} \pi_k\right)^2}{(1/K)\sum_{k=1}^{K} \pi_k^2}. \tag{29}$$

Note that $\sum_{k=1}^{K}\pi_k = 1$, and hence (29) leads to the dispersion of $\mathbf{p}$:

$$\mathrm{Dsp}(\boldsymbol{\pi}) = \frac{1}{\sum_{k=1}^{K}\pi_k^2}. \tag{30}$$

Interestingly, this measure of fairness is closely related to majorization theory. The function in (30) is symmetric and concave in $\boldsymbol{\pi}$ and therefore Schur concave [23, Proposition 2.8]. A function is called symmetric if the argument vector can be arbitrarily permuted without changing the value of the function.

Corollary 1. *The dispersion is a Schur-concave function of the vector $\boldsymbol{\pi}$, that is,*

$$\boldsymbol{\pi}_1 \succcurlyeq \boldsymbol{\pi}_2 \Longrightarrow \mathrm{Dsp}(\boldsymbol{\pi}_1) \leq \mathrm{Dsp}(\boldsymbol{\pi}_2). \tag{31}$$

4.3. Connection of User Distribution, Service Probability, and Delay. From the results in the last sections, it follows that the impact of the user location on the different fairness measures depends on the resulting service probability vector $\boldsymbol{\pi}$. Therefore, we have to map the user distribution vector $\mathbf{c}$ to the service probability vector $\boldsymbol{\pi}$. The concrete mapping depends on the chosen scheduler. For PFS, the service probabilities of all users are equal to $\pi_k = 1/K$ and thus independent of $\mathbf{c}$.

In order to apply majorization theory to the analysis of the average worst-case delay as a function of the user distribution, we have to transfer the partial order for user distributions to the partial order for probability that a user k is picked.

Define the vector of probabilities that user k is picked $\boldsymbol{\pi}$ as a function of the user distribution $\mathbf{c}$, that is,

$$\begin{aligned} \pi_k(\mathbf{c}) &= \Pr\left[c_k w_k \geq \max_{l \neq k} c_l w_l\right] \\ &= \sum_{\pi \in \mathcal{P} \setminus k} \int_{a_{\pi_{K-1}}=0}^{\infty} \int_{a_{\pi_{K-2}}=a_{\pi_{K-1}}}^{\infty} \cdots \\ &\quad \cdot \int_{a_k=a_{\pi_1}}^{\infty} \prod_{k=1}^{K} \frac{a_k^{n_T-1} e^{-(a_k/\Gamma(n_T)c_k)}}{c_k} d\mathbf{a}. \end{aligned} \tag{32}$$

The RHS in (32) contains all possible disjunct events, that is, all permutations, such that $c_k w_k \geq c_{\pi_1} w_{\pi_1} \geq c_{\pi_2} w_{\pi_2} \geq \cdots \geq c_{\pi_{K-1}} w_{\pi_{K-1}}$. The sum over all probabilities, that is, integrals with certain limits, gives the probability that user k is picked.

Unfortunately, the next result is an impossibility result. It shows that it is not possible to say that if $\mathbf{c} \succcurlyeq \mathbf{d}$ then automatically $\boldsymbol{\pi}(\mathbf{c}) \succcurlyeq \boldsymbol{\pi}(\mathbf{d})$.

Corollary 2. *The mapping from the vector of user distributions to the vector of service probabilities is not order preserving with respect to the partial order majorization.*

Proof. We provide a counterexample. Consider the user distribution vectors $\mathbf{c}= [5, 3, 2]^T$ and $\mathbf{d}= [4, 4, 2]^T$ and $n_T = 1$. The resulting activity probabilities computed according to (32) are given by $\boldsymbol{\pi}(\mathbf{c}) = [0.6428, 0.1786, 0.1786]^T$ and $\boldsymbol{\pi}(\mathbf{d}) = [0.4167, 0.4167, 0.1666]^T$. Majorization cannot be used to compare these two vectors because $\pi_1(\mathbf{c}) > \pi_2(\mathbf{d})$ but $\pi_1(\mathbf{c}) + \pi_2(\mathbf{c}) < \pi_1(\mathbf{d}) + \pi_2(\mathbf{d})$. □

Even though the connection between user distribution and service probability is not order preserving with respect to the partial order of majorization, it does not imply that the average worst-case delay is not a Schur-convex or Schur-concave function of the user distribution. Due to the complicated dependency of the average worst-case delay and the user distribution via (32), the following observation is stated as a conjecture.

Conjecture 1. *The average worst-case delay of MTS as a function of the user distribution is Schur-convex, that is,* $\mathbf{c} \succcurlyeq \mathbf{d} \Rightarrow \mathbb{E}[D_{1,K}(\mathbf{c})] \geq \mathbb{E}[D_{1,K}(\mathbf{d})]$.

5. Asymptotic Characterizations

In this section, we characterize the average sum rate of the different scheduling schemes for high SNR or for a large number of users. The scaling laws of the schemes are derived as a function of the user distribution. These results provide more quantitative but closed form expressions for the sum rate performance of the four schedulers.

5.1. High-SNR Behavior. The high-SNR slope $\mathcal{S}_\infty$ as defined in (2) for all four scheduling schemes is equal to one because

$$\begin{aligned}\mathcal{S}_\infty &= \lim_{\rho\to\infty} \frac{\int_0^\infty \log(1+\rho x) pdf(x)dx}{\log(\rho)} \\ &= \int_0^\infty \lim_{\rho\to\infty} \frac{\log(1+\rho x)}{\log(\rho)} pdf(x)dx \\ &= \int_0^\infty pdf(x)dx = 1.\end{aligned} \tag{33}$$

It is allowed to swap integration and limit by applying the dominated convergence theorem. In general, any TDMA scheme could have at most a high-SNR slope of one. The high-SNR power offset is different for the four schedulers. It is derived in the following result.

Theorem 6. *The high-SNR power offset is characterized for four cases as follows.*

(1) *For MTS, the high-SNR power offset is bounded from below and above by*

$$\begin{aligned}&\gamma + \log\left(\Gamma(1+n_T)^{1/n_T}\right) - \sum_{k=1}^{K}(-1)^{k-1}\binom{Kn_T}{k}\log(k) \\ &\quad \geq \mathcal{L}_\infty{}^{\mathrm{MT}} \geq \gamma - \log\left(Kn_T\right).\end{aligned} \tag{34}$$

For $n_T = 1$, the lower bound in (34) is equal to the lower bound result in [23, Theorem 2].

(2) *For RRS, the high-SNR power offset as a function of the user distribution is given by*

$$\mathcal{L}_\infty{}^{\mathrm{RR}}(\mathbf{c}) = \frac{1}{K}\sum_{k=1}^{K} -\Psi(n_T) - \log\left(c_k\right). \tag{35}$$

For $n_T = 1$, we obtain the closed form expression (compare to [15])

$$\mathcal{L}_\infty{}^{\mathrm{RR}}(\mathbf{c}) = \frac{1}{K}\sum_{k=1}^{K} \gamma - \log\left(c_k\right). \tag{36}$$

(3) *For PFS, the high-SNR power offset as a function of the user distribution is given by*

$$\mathcal{L}_\infty{}^{\mathrm{PF}}(\mathbf{c}) = -\Psi(n_T) - \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{K}(-1)^{l-1}\binom{K}{l}\log\left(\frac{l}{c_k}\right). \tag{37}$$

(4) *For ORS, the high-SNR power offsets as a function of the user distribution is given by*

$$\begin{aligned}\mathcal{L}_\infty{}^{\mathrm{OR}}(\mathbf{c}) &= \frac{1}{K^2}\sum_{n=1}^{K} n \sum_{k=1}^{K}\sum_{j=0}^{n-1}\binom{n-1}{j}\frac{(-1)^j}{1+j} \\ &\quad \cdot \left(\gamma + \log\left(\frac{1+j}{c_k}\right)\right).\end{aligned} \tag{38}$$

The proof of Theorem 6 follows similar lines as in [32, Theorem 2] and is, therefore omitted. Note that the Schur convexity of (36) can be directly observed and this approves the result in (15). However, in (37) and (38), the Schur convexity cannot be directly observed because of the alternating sum.

The high-SNR power offsets fulfill the following inequality chain:

$$\mathcal{L}_\infty{}^{\mathrm{MT}} \leq \left\{\mathcal{L}_\infty{}^{\mathrm{PF}}, \mathcal{L}_\infty{}^{\mathrm{OR}}\right\} \leq \mathcal{L}_\infty{}^{\mathrm{RR}}. \tag{39}$$

The order of PFS and ORS depends on the user distribution and number of antennas at the base station scenario. Note that the average worst-case delay does not scale with the SNR.

5.2. Scaling with Number of Users. First, consider the case in which the users are symmetrically distributed, that is, $\mathbf{c} = \mathbf{1}$. The scaling behavior with $K \to \infty$ for fixed SNR ρ can be easily shown by considering a simple upper and lower bounds on the average sum rate. The average sum rate of RR does not scale with K at all.

Corollary 3. *For symmetrically distributed users $\mathbf{c} = \mathbf{1}$, the average sumrates of MTS, PFS, and ORS scale for large K with $\log(K)$, that is,*

$$\begin{aligned}\lim_{K\to\infty}\frac{R_{\mathrm{sum}}^{\mathrm{MT}}(K)}{\log(K)} &= \lim_{K\to\infty}\frac{R_{\mathrm{sum}}^{\mathrm{PF}}(K)}{\log(K)} \\ &= \lim_{K\to\infty}\frac{R_{\mathrm{sum}}^{\mathrm{OR}}(K)}{\log(K)} = 1.\end{aligned} \tag{40}$$

The case in which the users are not symmetrically distributed is discussed in the numerical results section. The scaling of the average worst-case delay with the number of users is also of interest and is thus studied in Corollary 4. It follows directly from (21) and (23).

Corollary 4. *For symmetrically distributed users, the average worst-case delay scales linearly with K for RRS and ORS. For MTS and PFS, it scales as K $\log(K)$, that is,*

$$\begin{aligned}\lim_{K\to\infty}\frac{D_1^{\mathrm{RRS}}(K)}{K} &= \lim_{K\to\infty}\frac{D_1^{\mathrm{ORS}}(K)}{K} = 1, \\ \lim_{K\to\infty}\frac{D_1^{\mathrm{MTS}}(K)}{K\log(K)} &= \lim_{K\to\infty}\frac{D_1^{\mathrm{PFS}}(K)}{K\log(K)} = 1.\end{aligned} \tag{41}$$

The case in which the users are not symmetrically distributed is discussed also in the numerical results section. Note that the scaling law for MTS and PFS in (41) is the best case as shown in Theorem 5, the case in which the users are symmetrically distributed offers the lowest average worst-case delay.

6. Fixed Rate Allocation and Long-Term Power Constraint

In this section, we consider a certain communication scenario which leads to a slightly modified performance function on the physical layer. Usually, the traffic is divided into classes (see, e.g., traffic classes in [33]) which require a certain SNR level to guarantee successful delivery of the user contents. In the following, we study the behavior of the sum rate under fixed rate allocations for the three schedulers (MTS, RRS, and PFS) as a function of the user distribution for comparison with the sum rate behavior from the last section.

Let us assume that we have only one fixed transmission rate R_0 available, and each scheduled user obtains its information packet with that rate. Therefore, a certain SNR is needed for successful transmission. Denote the long-term sum transmit power constraint at the base station as P_ℓ, that is,

$$\mathbb{E}_{a_1,\ldots,a_k}\left[\sum_{k=1}^{K} p_k(a_1,\ldots,a_k)\right] \leq P_\ell. \tag{42}$$

We consider the three schedulers MTS, RRS, and PFS. The power allocation at the base station for all three schedulers is channel inversion under the long-term power constraint.

Theorem 7. *The achievable sum rate for fixed rate transmission of the RRS is given by*

$$R^{\mathrm{RR}}_{\mathrm{sum},fx} = \frac{1}{K}\sum_{k=1}^{K}\log\left(1+\frac{\rho P_\ell}{\mathbb{E}\left[(1/c_k w_k)\right]}\right). \tag{43}$$

The achievable sum rate for fixed rate transmission of the MTS is given by

$$R^{\mathrm{MT}}_{\mathrm{sum},fx} = \log\left(1+\frac{\rho P_\ell}{\mathbb{E}\left[(1/\max_{1\le k\le K} c_k w_k)\right]}\right). \tag{44}$$

Finally, the sum rate for fixed rate transmission of the PFS is given by

$$R^{\mathrm{PF}}_{\mathrm{sum},fx} = \frac{1}{K}\sum_{k=1}^{K}\log\left(1+\frac{\rho P_\ell}{\mathbb{E}\left[(1/c_k \max_{1\le k\le K} w_k)\right]}\right). \tag{45}$$

Proof. We will use one framework to derive the achievable sum rate for fixed rate transmission [34]. Denote the instantaneous channel power of the scheduled user as ζ. Then, the instantaneous achievable rate is $\log(1+\rho\zeta p(\zeta))$ with power $p(\zeta)$ allocated. This instantaneous rate should be equal to the fixed rate R_0 under the average power constraint in (42). We solve

$$R_0 = \log\left(1+\rho\zeta p(\zeta)\right) \tag{46}$$

for $p(\zeta)$ and normalize the constant c_P with respect to the long-term power constraint to obtain the optimal power allocation

$$p(\zeta) = \frac{c_P}{\zeta} = \frac{P_\ell}{\zeta}\frac{1}{\mathbb{E}[1/\zeta]}. \tag{47}$$

Equation (47) is simply channel inversion with long-term power constraint, that is,

$$\mathbb{E}[p(\zeta)] = P_\ell\mathbb{E}\left[\frac{1}{\zeta}\right]\frac{1}{\mathbb{E}[1/\zeta]} = P_\ell. \tag{48}$$

Inserting (47) into (46) yields

$$R_0 = \log\left(1+\rho\frac{P_\ell}{\mathbb{E}[1/\zeta]}\right). \tag{49}$$

Then expressions in (43), (44), and (45) follow when we use the effective channels ζ after scheduling. □

The impact of the user location on the sum rate performances is characterized in the following corollary.

Corollary 5. *The sum rate of RRS with fixed rate constraint is Schur concave with respect to* **c**. *The sum rate of PFS with fixed rate constraint is Schur concave with respect to* **c**.

The sum rates with fixed rate constraint and long-term power constraint for RRS and PFS show the same behavior as the sum rate with short-term power constraint.

Proof. We verify indirectly Schur's condition for the RRS and PFS and thereby leave the expectation unsolved. Both sum rates R_0^{PF} and R_0^{RR} can be written as functions of the user distribution **c**

$$\psi(\mathbf{c}) = \frac{1}{K}\sum_{k=1}^{K}\log\left(1+\frac{\rho c_k P_\ell}{\mathbb{E}[x]}\right) \tag{50}$$

for some random variable x. The function in (50) is symmetric with respect to **c**. The sum of concave functions in c_k is Schur-concave (see, e.g., [23, Proposition 2.7] or [21, 3.C.1]). □

Regarding the impact of the user distribution on the MTS sum rate with fixed rates, we observe that the behavior depends on the number of antennas and number of users. We leave this for future research.

7. Numerical Simulations

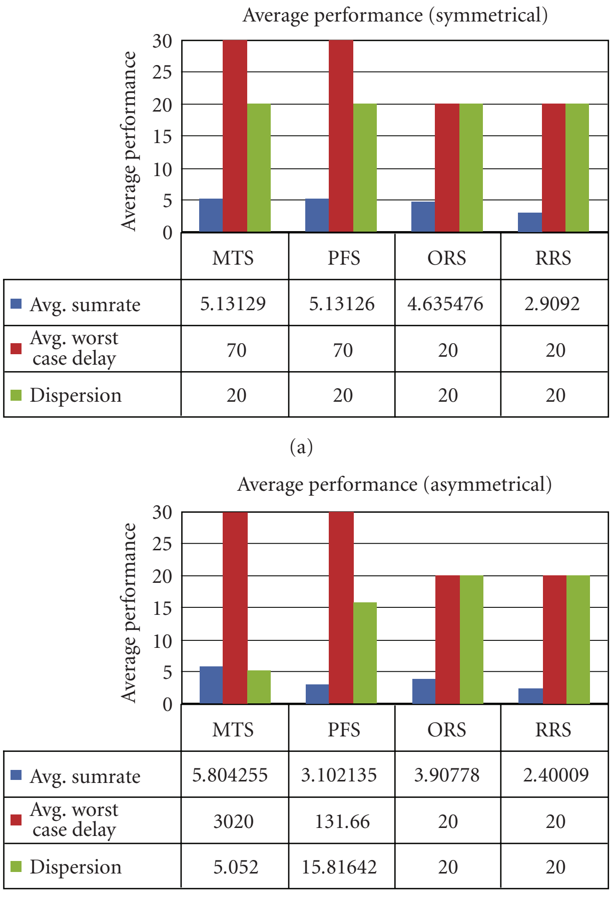
In this section, we present illustrations which validate and explain the theoretical results from the last sections. The performance for the case with symmetrically distributed users $\mathbf{c} = \mathbf{1}$ is compared to the case with asymmetrically users. For the asymmetrically user distribution, we choose the exponential decaying model

$$c_k = \exp(-tk), \quad \text{and normalize} \sum_{k=1}^{K} c_k = K. \tag{51}$$

For $K = 20$ and $t = 0.2$, we obtain the user distribution

$$\begin{aligned}\mathbf{c} = [&3.6930, 3.0236, 2.4755, 2.0268, 1.6594, 1.3586,\\ &1.1123, 0.9107, 0.7456, 0.6105, 0.4998, 0.4092,\\ &0.3350, 0.2743, 0.2246, 0.1839, 0.1505, 0.1232,\\ &0.1009, 0.0826].\end{aligned} \tag{52}$$

In the numerical simulations, for each data point, 100 000 Monte Carlo runs are performed to compute the averages.



	MTS	PFS	ORS	RRS
🟦 Avg. sumrate	5.13129	5.13126	4.635476	2.9092
🟥 Avg. worst case delay	70	70	20	20
🟩 Dispersion	20	20	20	20

(a)

	MTS	PFS	ORS	RRS
🟦 Avg. sumrate	5.804255	3.102135	3.90778	2.40009
🟥 Avg. worst case delay	3020	131.66	20	20
🟩 Dispersion	5.052	15.81642	20	20

(b)

Figure 1: Average sum rate, worst-case delay, and dispersion for $K = 20$ symmetrically and asymmetrically distributed users.

7.1. General Results.

In Figure 1, the average sum rate, the average worst-case delay, and the dispersion are shown for the four studied schedulers. In Figure 1(a), the users are symmetrically distributed, that is, $\mathbf{c} = \mathbf{1}$, whereas in Figure 1(b), the users are asymmetrically distributed according to the model in (51) with $t = 0.2$. The results in Figure 1 illustrate the following observations. The average sum rate of MTS increases with more asymmetrically distributed users (compare to (14)), while the average sum rate of all three other schedulers decreases (compare to (15), (18), and (20)). However, PFS outperforms ORS for the symmetrical scenario, whereas it is the other way round for the asymmetrical scenario. Another observation is that the average worst-case delay is more differentiated than the dispersion. This underlines that the average worst-case delay is better suited for fairness analysis than the JFI-based dispersion. Finally, the average worst-case delay for the asymmetrical scenario of the PFS and ORS tends to grow without bound. Therefore, taking the tradeoff between fairness and average sum rate into account, the PFS and ORS perform reasonable well. PFS is advantageous in symmetric scenarios whereas ORS performs better in asymmetric scenarios.

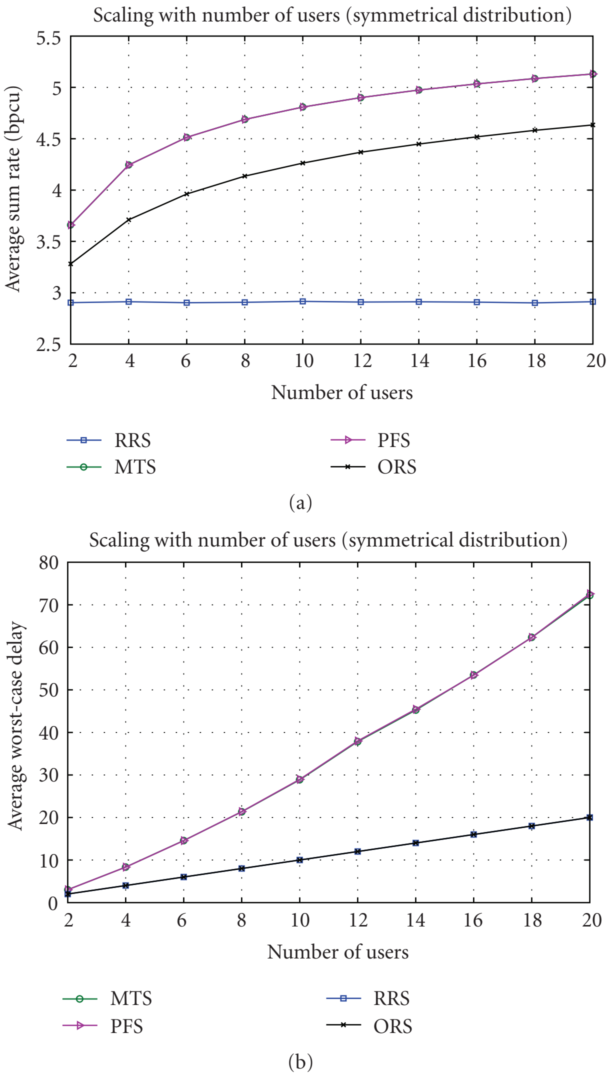


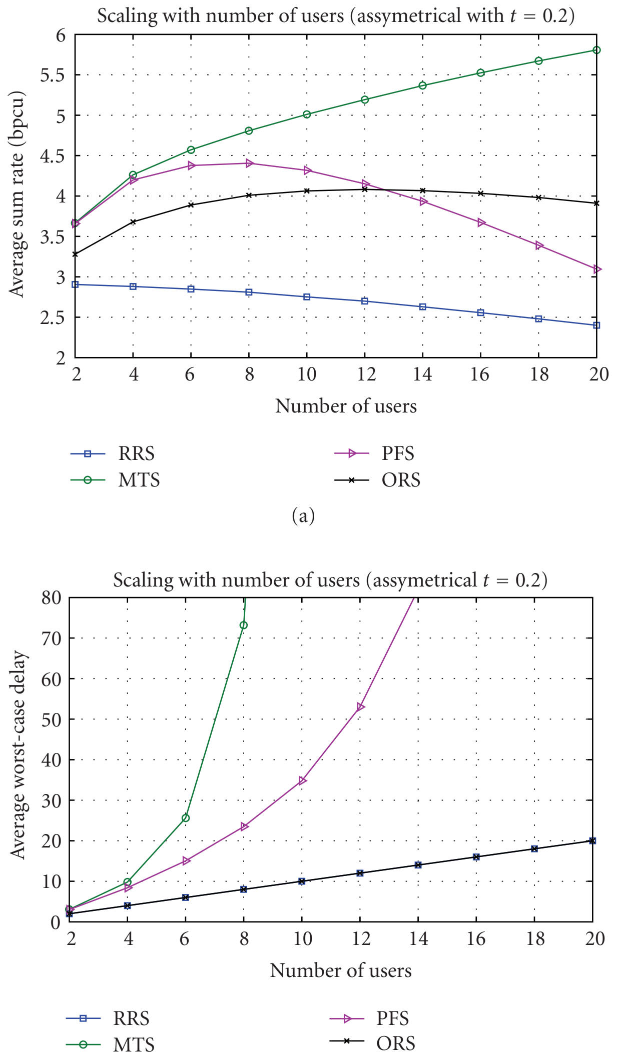
Figure 2: Average sum rate and worst-case delay versus number of users for symmetrically distributed users.

7.2. Scaling with Number of Users.

In Figures (2) and (3), we show the average performance of the four scheduling algorithms for symmetrically distributed as well as asymmetrically distributed users. The derived scaling laws in (40) and (41) are confirmed. The interesting observation is that for the asymmetrical case, PFS outperforms OFS for a small number of users, whereas it is the other way round for large number of users.

The average worst case delay for MTS and PFS increases with asymmetrical user distribution as predicted in Theorem 5. As soon as a single c_k approaches zero, the average worst-case delay approaches infinity. The round-based schedulers RRS and ORS are robust against the asymmetrical user distribution.

The main observation in this section is that for practical scenarios in which fairness is important as well as users are randomly distributed within the cell, ORS clearly outperforms PFS. Note that the results presented here hold for a



(a)

(b)

Figure 3: Average sum rate and worst-case delay versus number of users for asymmetrically distributed users.

static scenario in which we place the users only once inside the cell and simulate the small-scale fading. Mobility as well as traffic models is left for further research.

7.3. Multiple Antenna Case—OSTBC. The application of OSTBC yields to a tradeoff between the code rate and the number of degrees of freedom of the channel gain. The code rate r_C decreases with the number of antennas, whereas the number of degrees of freedom of the χ^2 distributed channel gain increases. For an OSTCB with n_T transmit antennas, it is shown in [35] that the maximum achievable code rate is given by

$$r_C(n_T) = \frac{\lfloor (n_T + 1)/2 \rfloor + 1}{2\lceil (n_T + 1)/2 \rceil}. \tag{53}$$

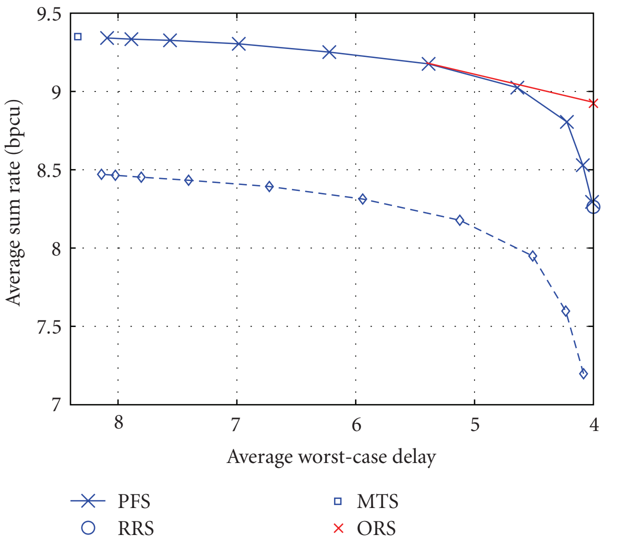


Figure 4: Average sum rate/worst-case delay tradeoff, $n_T = \{1, 2\}$; $K = 4$; SNR = 20 dB.

The code rate $r_C(n_T)$ starts at $r_C(1) = r_C(2) = 1$ and decreases to $\lim_{n_T \to \infty} r_C(n_T) = 1/2$. Therefore, we restrict the numerical simulations to the case $n_T = 2$.

In Figure 4, the achievable average sum rate versus average worst-case delay tradeoff is shown for a two antenna BS with four users at SNR = 20 dB for the four schedulers. The PFS is operated at ten window length operating points $t_c = 2^k$, $k = 1, \ldots, 10$. The RRS has lowest delay, whereas the MTS has largest delay but best performance. The closure of the convex hull of all operating points gives the achievable sum rate/delay region. The dashed line shows the single-antenna case. It can be observed that two antennas increase average sum rate as well as decrease the average worst-case delay significantly. Note that no additional (spatial) feedback is required to achieve this gain.

8. Conclusions

In this paper, we proposed an approach to analyze qualitatively the tradeoff between system throughput and fairness in a multiuser multiple antenna downlink transmission system. Four representative (three of them channel aware) schedulers were studied for different user distributions using majorization theory. The sum rate of MTS improves with asymmetrical user distribution, whereas the sum rate of all other schedulers improves with symmetrical user distribution. MTS and RRS serve as upper and lower bounds on throughput and lower and upper bounds on worst-case delay, respectively. The throughput-delay tradeoff of the four schedulers is characterized; if fairness as well as performance is important, the optimal choice will depend on the user distribution and the number of users. Finally, the gain of using multiple antennas without increased feedback overhead at the base station is illustrated.

Appendices

A. Proof of Theorem 1

Proof. In the proof, we verify Schur's condition directly. Therefore, we need the first derivative of $R_{\text{sum}}^{\text{MT}}$ with respect to c_1 and c_2 given as

$$\begin{aligned}\frac{\partial R_{\text{sum}}^{\text{MT}}}{\partial c_1} &= \int_0^\infty \frac{\rho t}{1+\rho t}\prod_{k=3}^{K}\left(1-\frac{\Gamma(n_T,(t/c_k))}{\Gamma(n_T)}\right)\\ &\quad\cdot\left(1-\frac{\Gamma(n_T,(t/c_2))}{\Gamma(n_T)}\right)\frac{(t^{n_T-1}/c_1)}{c_1^2\Gamma(n_T)}\exp\left(\frac{-t}{c_1}\right)dt,\\ \frac{\partial R_{\text{sum}}^{\text{MT}}}{\partial c_2} &= \int_0^\infty \frac{\rho t}{1+\rho t}\prod_{k=3}^{K}\left(1-\frac{\Gamma(n_T,(t/c_k))}{\Gamma(n_T)}\right)\\ &\quad\cdot\left(1-\frac{\Gamma(n_T,(t/c_1))}{\Gamma(n_T)}\right)\frac{(t^{n_T-1}/c_2)}{c_1^2\Gamma(n_T)}\exp\left(\frac{-t}{c_2}\right)dt.\end{aligned} \tag{A.1}$$

Define the two functions

$$\begin{aligned}f(\rho,t,\mathbf{c}) &= \frac{\rho t}{1+\rho t}\prod_{k=3}^{K}\left(1-\frac{\Gamma(n_T,(t/c_k))}{\Gamma(n_T)}\right),\\ g(t,c_1,c_2) &= \left(1-\frac{\Gamma(n_T,(t/c_2))}{\Gamma(n_T)}\right)\frac{(t/c_1)^{n_T-1}}{c_1^2\Gamma(n_T)}\exp\left(-\frac{t}{c_1}\right)\\ &\quad-\left(1-\frac{\Gamma(n_T,(t/c_1))}{\Gamma(n_T)}\right)\frac{(t/c_2)^{n_T-1}}{c_2^2\Gamma(n_T)}\exp\left(-\frac{t}{c_2}\right),\end{aligned} \tag{A.2}$$

in order to express the difference of the first derivatives of the sum rate of the MTS as

$$\frac{\partial R_{\text{sum}}^{\text{MT}}}{\partial c_1}-\frac{\partial R_{\text{sum}}^{\text{MT}}}{\partial c_2} = \int_0^\infty f(\rho,t,\mathbf{c})g(t,c_1,c_2)dt. \tag{A.3}$$

The following properties of the functions f and g are easily verified; f is monotonic increasing from zero to one. The function g is $g(t=0)=0$, has one zero at $t^*: g(t^*)=0$, and is negative for all $t<t^*$ and positive for all $t>t^*$. Therefore, we can lower bound (A.3) by using the zero t^* as

$$\frac{\partial R_{\text{sum}}^{\text{MT}}}{\partial c_1}-\frac{\partial R_{\text{sum}}^{\text{MT}}}{\partial c_2} \geq f(\rho,t^*,\mathbf{c})\int_0^\infty g(t,c_1,c_2)dt. \tag{A.4}$$

Finally, the integral in (A.4) can be computed in closed form

$$\begin{aligned}\int_0^\infty g(t,c_1,c_2)dt &= \frac{1}{2}\frac{1}{c_1c_2\Gamma(1+n_T)\sqrt{\pi}}\\ &\quad\cdot\left\{2\sqrt{\pi}\Gamma(n_T+1)[c_2-c_1]+\Gamma(n_T+1/2)4^{n_T}\left(\frac{c_1}{c_2}\right)^{n_T}\right.\\ &\quad\cdot\left[c_1\cdot{}_2F_1\left(n_T,2n_T;1+n_T;-\left(\frac{c_1}{c_2}\right)\right)\right.\\ &\quad\left.\left.-c_2\cdot{}_2F_1\left(n_T,2n_T;1+n_T;-\left(\frac{c_2}{c_1}\right)\right)\right]\right\},\end{aligned} \tag{A.5}$$

where ${}_2F_1(a,b;c;z)$ is the Gauss hypergeometric function [30, Chapter 15]. For single-antenna BS, we set $n_T=1$ to obtain

$$G(c_1,c_2,1)=0, \tag{A.6}$$

which is in perfect agreement with the result and its proof in [28]. Since, the function $G(c_1,c_2,n_T)$ is monotonic increasing with n_T, this implies that

$$\frac{\partial R_{\text{sum}}^{\text{MT}}}{\partial c_1}-\frac{\partial R_{\text{sum}}^{\text{MT}}}{\partial c_2} \geq f(\rho,t^*,\mathbf{c})G(c_1,c_2,n_T)\geq 0, \tag{A.7}$$

which verifies Schur's condition for Schur convexity. □

B. Proof of Theorem 4

Proof. The proof is similar to the proof in Appendix A. The difference is that we have two sums in the integral instead of the product. Starting from the representation in (10), the difference of the first partial derivatives with respect to c_1 and c_2, respectively, is computed

$$\begin{aligned}\frac{\partial R_{\text{sum}}^{\text{OR}}}{\partial c_1} &= \int_0^\infty \frac{\rho}{1+\rho t}\frac{1}{K^2}\\ &\quad\cdot\sum_{k=1}^{K}\frac{(1-(\Gamma(n_T,t/c_1)/\Gamma(n_T)))^k k t^{n_T}\exp(-t/c_1)}{\Gamma(n_T)-\Gamma(n_T,t/c_1)c_1^{n_T+1}}dt,\\ \frac{\partial R_{\text{sum}}^{\text{OR}}}{\partial c_2} &= \int_0^\infty \frac{\rho}{1+\rho t}\frac{1}{K^2}\\ &\quad\cdot\sum_{k=1}^{K}\frac{(1-(\Gamma(n_T,t/c_2)/\Gamma(n_T)))^k k t^{n_T}\exp(-t/c_2)}{\Gamma(n_T)-\Gamma(n_T,t/c_2)c_2^{n_T+1}}dt.\end{aligned} \tag{B.1}$$

Define the two functions

$$\begin{aligned}\phi(\rho,t) &= \frac{\rho}{1+\rho t},\\ \gamma&(t,c_1,c_2,k,n_T)\\ &= \frac{(1-(\Gamma(n_T,t/c_1)/\Gamma(n_T)))^k k t^{n_T}\exp(-t/c_1)}{\Gamma(n_T)-\Gamma(n_T,t/c_1)c_1^{n_T+1}}\\ &\quad-\frac{(1-(\Gamma(n_T,t/c_2)/\Gamma(n_T)))^k k t^{n_T}\exp(-t/c_2)}{\Gamma(n_T)-\Gamma(n_T,t/c_2)c_2^{n_T+1}},\end{aligned} \tag{B.2}$$

in order to rewrite the difference of the first derivatives as

$$\begin{aligned}\Delta &= \frac{\partial R_{\text{sum}}^{\text{OR}}}{\partial c_1}-\frac{\partial R_{\text{sum}}^{\text{OR}}}{\partial c_2}\\ &= \frac{1}{K^2}\sum_{k=1}^{K}\int_0^\infty \phi(\rho,t)\gamma(t,c_1,c_2,k,n_T)dt.\end{aligned} \tag{B.3}$$

The properties of the functions ϕ and γ are as follows. ϕ is monotonic decreasing with respect to t, and γ has similar properties as the function g in the proof in Appendix A.

$\gamma(t=0)=0$, it has on zero at $t^*: g(t^*)=0$, it is negative for all $t<t^*$ and positive for all $t>t^*$. Therefore, we obtain an upper bound on Δ in (B.3) as

$$\Delta \leq \frac{1}{K^2}\sum_{k=1}^{K}\phi(\rho,t^*)\int_0^\infty \gamma(t,c_1,c_2,k,n_T)\,dt = 0, \qquad \text{(B.4)}$$

because $\int_0^\infty \gamma(t,c_1,c_2,k,n_T)\,dt = 0$. This verifies Schur's condition for Schur concavity and completes the proof. □

References

[1] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proceedings of IEEE International Conference on Communications (ICC '95)*, vol. 1, pp. 331–335, Seattle, Wash, USA, June 1995.

[2] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, Cambridge, UK, 2005.

[3] T. Issariyakul and E. Hossain, "ORCA-MRT: an optimization-based approach for fair scheduling in multirate TDMA wireless networks," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2823–2835, 2005.

[4] X. Wang, A. G. Marques, and G. B. Giannakis, "Power-efficient resource allocation and quantization for TDMA using adaptive transmission and limited-rate feedback," *IEEE Transactions on Signal Processing*, vol. 56, no. 9, pp. 4470–4485, 2008.

[5] F. Halsall, *Data Communications, Computer Networks and Open Systems*, Electronic Systems Engineering, Addison-Wesley, Reading, Mass, USA, 4th edition, 1996.

[6] L. Yang and M.-S. Alouini, "Performance analysis of multiuser selection diversity," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 6, pp. 1848–1861, 2006.

[7] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling policies for wireless systems with short term fairness constraints," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '03)*, vol. 1, pp. 533–537, San Francisco, Calif, USA, December 2003.

[8] V. Hassel, G. E. Øien, and D. Gesbert, "Throughput guarantees for wireless networks with opportunistic scheduling: a comparative study," *IEEE Transactions on Wireless Communications*, vol. 6, no. 12, pp. 4215–4220, 2007.

[9] G. Song and Y. Li, "Asymptotic throughput analysis for channel-aware scheduling," *IEEE Transactions on Communications*, vol. 54, no. 10, pp. 1827–1834, 2006.

[10] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Research Report TR-301, DEC, New York, NY, USA, September 1984.

[11] R. Elliott, "A measure of fairness of service for scheduling algorithms in multiuser systems," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE '02)*, vol. 3, pp. 1583–1588, Winnipeg, Canada, May 2002.

[12] M. Sharif and B. Hassibi, "Delay considerations for opportunistic scheduling in broadcast fading channels," *IEEE Transactions on Wireless Communications*, vol. 6, no. 9, pp. 3353–3363, 2007.

[13] N. Golmie, *Coexistence in Wireless Networks*, Cambridge University Press, Cambridge, UK, 2007.

[14] Z. Han and K. J. R. Liu, *Resource Allocation for Wireless Networks: Basics, Techniques, and Applications*, Cambridge University Press, Cambridge, UK, 2008.

[15] E. A. Jorswieck, A. Sezgin, and X. Zhang, "Framework for analysis of opportunistic schedulers: average sum rate vs. average fairness," in *Proceedings of the 4th Workshop on Resource Allocation in Wireless Networks (RAWNET '08)*, Berlin, Germany, March 2008.

[16] A. Sezgin, E. Jorswieck, and M. Charafeddine, "Interaction between scheduling and user locations in an OSTBC coded downlink system," in *Proceedings of the 7th International ITG Conference on Source and Channel Coding (SCC '08)*, Ulm, Germany, January 2008.

[17] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, 1998.

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

[19] E. Jorswieck, B. Ottersten, A. Sezgin, and A. Paulraj, "Guaranteed performance region in fading orthogonal space-time coded broadcast channels," *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, Article ID 268979, 12 pages, 2008.

[20] A. Sezgin and E. Jorswieck, "On the performance of partial feedback based orthogonal block coding," in *Proceedings of the 62nd IEEE Vehicular Technology Conference (VTC '05)*, vol. 3, pp. 1504–1508, Dallas, Tex, USA, September 2005.

[21] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Application*, vol. 143 of *Mathematics in Science and Engineering*, Academic Press, New York, NY, USA, 1979.

[22] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[23] E. Jorswieck and H. Boche, "Majorization and matrix-monotone functions in wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 3, no. 6, pp. 553–701, 2006.

[24] A. Lozano, A. M. Tulino, and S. Verdú, "High-SNR power offset in multiantenna communication," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4134–4151, 2005.

[25] C.-J. Chen and L.-C. Wang, "A unified capacity analysis for wireless systems with joint multiuser scheduling and antenna diversity in Nakagami fading channels," *IEEE Transactions on Communications*, vol. 54, no. 3, pp. 469–478, 2006.

[26] M. Johansson, "Diversity-enhanced equal access—considerable throughput gains with 1-bit feedback," in *Proceedings of the 5th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC '04)*, pp. 6–10, Lisbon, Portugal, July 2004.

[27] V. Hassel, M. R. Hanssen, and G. E. Øien, "Spectral efficiency and fairness for opportunistic round robin scheduling," in *Proceedings of IEEE International Conference on Communications (ICC '06)*, vol. 2, pp. 784–789, Istanbul, Turkey, July 2006.

[28] E. A. Jorswieck and H. Boche, "Throughput analysis of cellular downlink with different types of channel state information," in *Proceedings of IEEE International Conference on Communications (ICC '06)*, vol. 4, pp. 1526–1531, Istanbul, Turkey, July 2006.

[29] D. J. Newman and L. Shepp, "The double dixie cup problem," *The American Mathematical Monthly*, vol. 67, no. 1, pp. 58–61, 1960.

[30] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, NY, USA, 1970.

[31] M. L. Clevenson and W. Watkins, "Majorization and the birthday inequality," *Mathematics Magazine*, vol. 64, no. 3, pp. 183–188, 1991.

[32] E. A. Jorswieck, P. Svedman, and B. Ottersten, “Performance of TDMA and SDMA based opportunistic beamforming,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 11, pp. 4058–4063, 2008.

[33] T. Bonald and A. Proutire, “On the traffic capacity of cellular data networks,” in *Proceedings of the 24th International Conference on Thermoelectrics (ICT ’05)*, Clemson, SC, USA, June 2005.

[34] E. A. Jorswieck and H. Boche, “Delay-limited capacity: multiple antennas, moment constraints, and fading statistics,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 12, pp. 4204–4208, 2007.

[35] X.-B. Liang, “Orthogonal designs with maximal rates,” *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2468–2503, 2003.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 801613, 10 pages
doi:10.1155/2009/801613

Research Article

Optimal and Fair Resource Allocation for Multiuser Wireless Multimedia Transmissions

Zhangyu Guan, Dongfeng Yuan, and Haixia Zhang

Wireless Mobile Communications and Transmission Laboratory. (WMCT), Shandong University, Jinan, 250100, China

Correspondence should be addressed to Dongfeng Yuan, dfyuan@sdu.edu.cn

Received 30 June 2008; Revised 18 December 2008; Accepted 20 February 2009

Recommended by Kwang-Cheng Chen

This paper presents an optimal and fair strategy for multiuser multimedia radio resource allocation (RRA) based on coopetition, which suggests a judicious mixture of competition and cooperation. We formulate the co-opetition strategy as sum utility maximization at constraints from both Physical (PHY) and Application (APP) layers. We show that the maximization can be solved efficiently employing the well-defined Layering as Optimization Decomposition (LOD) method. Moreover, the coopetition strategy is applied to power allocation among multiple video users, and evaluated through comparing with existing- competition based strategy. Numerical results indicate that, the co-opetition strategy adapts the best to the changes of network conditions, participating users, and so forth. It is also shown that the coopetition can lead to an improved number of satisfied users, and in the meanwhile provide more flexible tradeoff between system efficiency and fairness among users.

1. Introduction

Radio resource allocation (RRA) for multimedia services has drawn a lot of attention because of its capability of offering an efficient way to handle the resources. In previous research, much attention has been paid to system efficiency improvement, that is, maximizing system utility [1–8]. It is shown that the Nash Bargaining Solution (NBS), a well-defined notion in game theory, can be used to maximize the sum of Peak Signal-to-Noise Ratios (PSNRs) in rate allocation for collaborative video transmissions [1]. Optimal resource allocation for multiuser wireless transmissions is studied in [2] from an information theoretic perspective, and it is shown that sum rate maximization (SRM) is suboptimal when taking video quality into account. This work has been extended to joint power and subcarrier allocation for mutiuser video transmission in multi-carrier systems [3]. In [4], Application (APP), MAC, and Physical (PHY) layers are jointly optimized using Cross-Layer Design (CLD) for streaming video delivery in a multiuser wireless environments, and two objective functions are introduced, that is, minimizing the sum of mean square error (MSE) of all video users, maximizing the sum of PSNRs. As a continuous work of [4, 5] proposed an application-driven cross-layer optimization strategy and discussed the challenges in CLD for multiuser multimedia services. Two Layering, as Optimization Decomposition (LOD) methods, dual decomposition and gradient projection-based decomposition, are used in [6, 7] for downlink utility maximization (DUM) assuming utility functions at APP layer are concave, increasing, and differentiable. The maximization of weighted sum of data rates in cross-layer resource allocation is addressed in [8], and an improved conjugate gradient method under given power constraint is presented as well.

In the work mentioned above, all the resource allocation methods try to maximize the global utility function. There are also several resource allocations that run in a distributive way, for instance, ReSerVation Protocol (RSVP) was used to allocate bandwidth among multiple multimedia streams over internet based on the Traffic SPECifications (TSPECs) [9]; air time fairness allocates transmission time proportionally to TSPECs to eliminate the passive impact of cross-layer strategies employed in different transmitters [10]. Proportional fairness was introduced [11] to allocate resources based on users' rate requirements, and further applied to rate controlling [12]. In [1], the Kalai-Smorodinsky Bargaining

Solution (KSBS) was used to allocate rates amongst multiple video users such that the utility achieved by each user is proportional to the maximum utility achievable.

Both maximization based and distributive policies work in a competitive way as explained by the following two examples. Utility maximization can actually be viewed as a process in which all users compete for resources according to the criteria that the Highest Quality Improvement the Highest Possibility Resources (HQIHPR) [2]. Using KSBS, users compete for resources to make efficient use of the resource and achieve higher utility. The disadvantage of these competitive policies is that they do not consider user's quality of service (QoS) satisfication degree, meaning that they are not suitable for multimedia services. To address this disadvantage, we propose an optimal and fair policy for multimedia resource allocation, which introduces a judicious mixture of competition and cooperation, such that user's QoS satisfication degree is taken into account. The idea behind this judicious mixture is Co-opetition, a concept from economic [13]. Co-opetition has been employed in decentralized resource management [14] and collaborative multimedia resource allocation in our preliminary work [15]. It is shown that co-opetition can provide better tradeoff between system efficiency and fairness.

Main contribution of this paper relies on the proposal of a novel co-opetition strategy for RRA in multimedia services, which is both optimal and fair. In this paper, optimal represents sum utility maximization (SUM) subject to the constraints on individual utility. It is worth to mention that the value of optimal sum utility might be smaller than that achieved by the unconstrained SUM, due to the constraints. Fair is defined to describe that, compared to unconstrained SUM, our strategy can result in fairer resource allocation. The additional fairness from our strategy comes from the individual utility constraint. Recall that the unconstrained SUM allocates resources in a competitive way, which has no constraint on individual utility. Our co-opetition strategy suggests a judicious mixture of competition and cooperation in resource allocation. We formulate the co-opetition strategy mathematically and solve it efficiently using LOD method. This mathematical formulation would help to get a better insight into the essential of competition and cooperation behaviors of users in RRA. We apply our strategy to wireless resource allocation for multiuser video transmissions and evaluate its performance by comparing with existing competition based mechanisms.

The rest of this paper is organized as follows. In Section 2, we formulate the co-opetition strategy, and in Section 3 we implement it by employing LOD method. In Section 4, we apply the co-opetition strategy to power allocation amongst multiple video users together with numerical results for performance evaluation. Conclusion is drawn in Section 5.

2. Problem Setup

We consider RRA over a downlink transmission with N users. We assume that the resource available at PHY layer is denoted by X. Denote $\mathcal{R} \subset \mathcal{R}_{0,+}^N$ as the rate region achievable at PHY layers, and assume that $\mathcal{R}$ is convex and compact. Convexity assumption means that time-sharing mode is enabled at PHY layer. Let $U_n(r_n), r_n \in \mathcal{R}_{0,+}$ denote the user n's utility function, which is assumed to be concave, increasing, and differentiable. An example of utility is PSNR for video services [16]. Each user has a minimum desired rate, denoted by r_{0n}, which should be at least guaranteed. That means

$$r_n \geq r_{0n}, \tag{1}$$

otherwise, user n would not be served. A competition strategy should be employed to develop our co-opetition strategy. In this paper, we focus on optimization-based strategy, that is, sum utility maximization (SUM). Investigation based on distributive and competition-based strategies will be accommodated in our future work. For SUM, system utility function $U : \mathcal{R}_{0,+}^N \to \mathcal{R}_{0,+}$ is defined as

$$U(\vec{r}) = \sum_{n=1}^{N} U_n(r_n), \tag{2}$$

where $\vec{r} = (r_1, \ldots, r_N)$. Hence, SUM can be written as

$$\max_{\vec{r}\in\mathcal{R}} U(\vec{r}), \quad \text{s.t. } r_n \geq r_{0n}. \tag{3}$$

To allow co-opetition, we first define the notion of satisfied user. A user is called satisfied user if its achieved QoS is above or equal to predefined QoS threshold, U_{th}. Then the basic idea of co-opetition can be described as follows. During the process of RRA, in which all users compete for resources to achieve SUM, users who have achieved U_{th} stop competing temporarily, until all resources have been allocated or all users have been satisfied. Denote rate required by user n to achieve U_{th} with $r_{n,\text{th}}$, and denote $\vec{r}_{\text{th}}$ as $(r_{1,\text{th}}, \ldots, r_{N,\text{th}})$. We distinguish the following two cases.

(1) If $\vec{r}_{\text{th}} \in \mathcal{R}$, co-opetition allocates resources such that the minimum utility of all users is U_{th}, that is, $U_n \geq U_{\text{th}}, \forall n$.

(2) If $\vec{r}_{\text{th}} \notin \mathcal{R}$, co-opetition allocates resources such that the maximum utility of all users is U_{th}, that is, $U_n \leq U_{\text{th}}, \forall n$.

Thus, our co-opetition strategy reads

$$\begin{aligned} &\max_{\vec{r}\in\mathcal{R}} U(\vec{r}), \\ &\quad \text{s.t. } r_n \geq r_{0n}, \\ &\qquad U_n \geq U_{\text{th}}, \forall n, \text{ if } \vec{r}_{\text{th}} \in \mathcal{R}, \\ &\qquad U_n \leq U_{\text{th}}, \forall n, \text{ if } \vec{r}_{\text{th}} \notin \mathcal{R}. \end{aligned} \tag{4}$$

Introducing U_{th} provides better tradeoff between system efficiency and fairness. For example, for video services in which PSNR is chosen as a QoS metric, U_{th} can be set corresponding to PSNR = 35 dB, above which user could achieve good video quality and user's video satisfaction degree increases very slowly as PSNR increases. In this case, rate, which can translate to resources at PHY layer, is more important to unsatisfied users. In the following, we investigate how the LOD method is used to solve (4) efficiently.

3. LOD Method

LOD is a well-defined technique for network utility maximization (NUM) by decomposing the NUM into a set of subproblems coupled with each other. Each subproblem is associated with a protocol layer, in which it can be solved separately [17].

3.1. Rewrite Co-opetition Strategy. We assume it is known whether $\vec{r}_{\text{th}}$ can be achieved or not. In the case of $\vec{r}_{\text{th}} \in \mathcal{R}$, $U_n \geq U_{\text{th}}$ translates into $r_n \geq r_{n,\text{th}}$, and $U_n \leq U_{\text{th}}$ translates into $r_n \leq r_{n,\text{th}}$ otherwise. We also assume that

$$r_{n,\text{th}} > r_{0n} \tag{5}$$

always satisfies. Then constraints in (4) can be rewritten as

$$\begin{aligned} &\vec{r}_{\text{th}} \leq \vec{r} \leq \infty, \quad \text{if } \vec{r}_{\text{th}} \in \mathcal{R}, \\ &\vec{r}_0 \leq \vec{r} \leq \vec{r}_{\text{th}}, \quad \text{if } \vec{r}_{\text{th}} \notin \mathcal{R}, \end{aligned} \tag{6}$$

where $\vec{r} = (r_1, \dots, r_N), \vec{r}_0 = (r_{01}, \dots, r_{0N})$(In the case of $\vec{r}_0 \notin \mathcal{R}$, total resource available cannot guarantee all users the minimum resource required, and some users will deny to be served. In this paper, we assume the minimum resource of all users can be always guaranteed, that is, $\vec{r}_0 \in \mathcal{R}$.) . We observe that, no matter $\vec{r}_{\text{th}} \in \mathcal{R}$ or not, the constraint has the same form of

$$\vec{r}_{\text{low}} \leq \vec{r} \leq \vec{r}_{\text{upp}}, \tag{7}$$

with $\vec{r}_{\text{low}} = (r_{l1}, \dots, r_{lN}), \vec{r}_{\text{upp}} = (r_{u1}, \dots, r_{uN})$. Hence, (4) can be rewritten as

$$\max_{\vec{r} \in \mathcal{R}} U(\vec{r}), \quad \text{s.t. } \vec{r}_{\text{low}} \leq \vec{r} \leq \vec{r}_{\text{upp}}. \tag{8}$$

3.2. Dual Decomposition. To solve (8) with LOD, (8) is firstly modified by introducing an additional variable $\vec{s}$, then the primal function (8) reads

$$\begin{aligned} &\max_{\vec{s}} U(\vec{s}), \\ &\text{s.t. } \vec{r}_{\text{low}} \leq \vec{s} \leq \vec{r}, \\ &\quad \vec{r} \leq \vec{r}_{\text{upp}}, \\ &\quad \vec{r} \in \mathcal{R}. \end{aligned} \tag{9}$$

After introducing the Lagrangian factors

$$\begin{aligned} \vec{\lambda} &= (\lambda_1, \dots, \lambda_N)^{\text{T}}, \\ \vec{\lambda\prime} &= (\lambda_{1'}, \dots, \lambda_N')^{\text{T}}, \end{aligned} \tag{10}$$

the Lagrangian function of (9) is written as

$$L\left(\vec{s}, \vec{r}, \vec{\lambda}, \vec{\lambda}'\right) = U(\vec{s}) + \left(\vec{\lambda}^{\text{T}}, \vec{\lambda\prime}^{\text{T}}\right) \begin{pmatrix} \vec{r} - \vec{s} \\ \vec{s} - \vec{r}_{\text{low}} \end{pmatrix} \tag{11}$$

with $\vec{\lambda} \geq 0, \vec{\lambda\prime} \geq 0$. Thus, the dual function is

$$g\left(\vec{\lambda}, \vec{\lambda}'\right) = \sup_{\vec{s}} L\left(\vec{s}, \vec{r}, \vec{\lambda}, \vec{\lambda}'\right), \tag{12}$$

The maximization in (9) can be solved by searching the optimum $\vec{\lambda}$ and $\vec{\lambda\prime}$ such that the dual function is minimized, that is,

$$\min_{\vec{\lambda}, \vec{\lambda\prime}} g\left(\vec{\lambda}, \vec{\lambda}'\right). \tag{13}$$

Based on the analysis afore, (12) can be decomposed into two subproblems as

$$g\left(\vec{\lambda}, \vec{\lambda\prime}\right) = g_{\text{A}}\left(\vec{\lambda}, \vec{\lambda\prime}\right) + g_{\text{P}}\left(\vec{\lambda}\right), \tag{14}$$

where

$$g_{\text{A}}\left(\vec{\lambda}, \vec{\lambda\prime}\right) = \max_{\vec{s}}\left(U(\vec{s}) + \left(\vec{\lambda\prime}^{\text{T}} - \vec{\lambda}^{\text{T}}\right)\vec{s} - \vec{\lambda\prime}^{\text{T}}\vec{r}_{\text{low}}\right), \tag{15}$$

$$g_{\text{P}}\left(\vec{\lambda}\right) = \max_{\substack{\vec{r} \in \mathcal{R}, \\ \vec{r} \leq \vec{r}_{\text{upp}}}} \vec{\lambda}^{\text{T}}\vec{r}. \tag{16}$$

For given $\vec{\lambda}$ and $\vec{\lambda\prime}$, the above two-maximization can be solved independently at APP layer for (15) and at PHY layer for (16). So far, we have transformed the original maximization, (8), into its dual problem.

3.3. Solving (13), (15) *and* (16). As mentioned above, for each fixed $\vec{\lambda}$ and $\vec{\lambda\prime}$, (15) and (16) have to be solved. Denote $G(\vec{s})$ as the item to be maximized in (15), that is,

$$G(\vec{s}) = U(\vec{s}) + \left(\vec{\lambda\prime}^{\text{T}} - \vec{\lambda}^{\text{T}}\right)\vec{s} - \vec{\lambda\prime}^{\text{T}}\vec{r}_{\text{low}}. \tag{17}$$

Then $G(\vec{s})$ is continuous and differentiable, and further denote S_0 as set of $\vec{s} = (s_1, \dots, s_N)$ such that

$$S_0 = \left\{ \vec{s} \,\middle|\, \frac{\partial G(\vec{s})}{\partial s_n} = 0, \; n = 1, \dots, N \right\}. \tag{18}$$

Then (15) can be solved via efficiently selecting the optimum $\vec{s}^*$, such that

$$\vec{s}^* = \arg\max_{\vec{s} \in S_0} G(\vec{s}). \tag{19}$$

Maximization of (16) refers to weighted sum rate maximization (WSRMax) at constraint of maximizing individual rate for certain PHY layer setup. $\vec{r} \in \mathcal{R}$ is a general constraint usually corresponding to given power or bandwidth. $\vec{r} \leq \vec{r}_{\text{upp}}$ can be translated into individual constraint. Recall that, $\mathcal{R}$ is

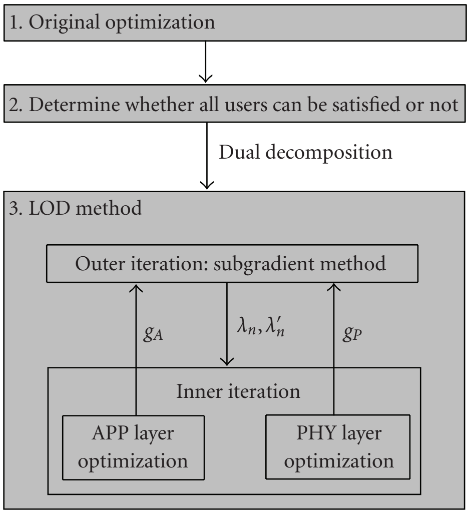


Figure 1: Illustration of the implement of co-opetition strategy.

assumed to be convex and compact, thus the domain of (16), denoted with $\mathcal{R}'$,

$$\mathcal{R}' = \mathcal{R} \cap \left\{\vec{r} \,\middle|\, \vec{r} \le \vec{r}_{\text{upp}}\right\}, \tag{20}$$

is also convex and compact. WSRMax over $\mathcal{R}'$ is a well-researched problem and there are many efficient solutions for a wide range of PHY layer setups [3, 8, 18].

Hereafter, we assume that for each $\vec{\lambda}$ and $\vec{\lambda\prime}$, (15) and (16) can be solved efficiently. Then the optimum $\vec{\lambda}$ and $\vec{\lambda\prime}$ can be determined, for example, using either sub-gradient method, cutting plane method or ellipsoid method [19]. In Section 5, we would show how to solve (13), (15) and (16) more concretely through power allocation.

3.4. Determining Whether $\vec{\mathbf{r}}_{\mathbf{th}} \in \mathbf{R}$ *or Not.* Note that is $\vec{r}_{\text{th}}$ not necessarily achievable. Whether $\vec{r}_{\text{th}} \in \mathcal{R}$ or not can be determined by userwisely computing the minimum resource required to achieve $\vec{r}_{\text{th}}$. Fortunately again there are several solutions available for different scenarios. For example, in [20] a generic procedure, CLARA, was presented for cross-layer resource minimization subject to a set of constraints on the overall QoS. [21] proposed an iterative algorithm which monotonically converges to the unique allocation with optimal sum power efficiency. This is actually another hot topic as opposed to utility maximization in this paper, namely, cost minimization to achieve certain QoS.

3.5. Summery of LOD Method. In this Section, we have mapped our co-opetition strategy, (4), to a standard constrained optimization over convex domain, that is, (8). Moreover, importantly, through applying the LOD, many well-researched solutions are available which make our co-opetition strategy more applicable. Finally, since the resource allocation in this paper can be formulated as a convex optimization, the LOD method has worst-case polynomialtime complexity [17]. It will be shown that the LOD method converges within limited iterations. Figure 1 is a brief description to apply the co-opetition strategy. We investigate how co-opetition can be applied to power allocation in detail.

4. RRA Using Co-Opetition

In this Section, we first describe the system scenario, and then illustrate the co-opetition strategy in detail. Finally, numerical results are presented for performance evaluation through comparing with competition-based strategy.

4.1. System Setup. We consider downlink N-user video transmission in a cell with a base-station (BS) which acts as the central spectrum manager (CSM). At APP layer, users transmit same or different video sequences. We choose PSNR as user's utility as it is the only widely accepted video QoS metric and choose the rate-distortion (RD) model proposed in [16] to describe user's average RD behavior as this model applies well to the state-of-the-art video encoder [22]. Then user's utility can be defined as

$$U_n(r_n) = 10 \log \frac{255^2 (r_n - R_{0n})}{D_{0n}(r_n - R_{0n}) + \mu_n}, \tag{21}$$

where R_{0n}, D_{0n} and μ_n are sequence parameters, which are dependent on video sequence characteristics, such as spatial and temporal resolution, delay constraints as well as the percentage of INTRA coded macro-blocks [1, 16]. D_{0n} is the minimum rate that should be at least guaranteed for user n, therefore in this work we assume that $r_n > R_{0n}$.

At PHY layer, the BS has limited transmit power, P_{tot}. Let $\vec{P} = (P_1, \dots, P_N)$ represent the power allocated to all the users, thus we have $\sum_{n=1}^{N} P_n \le P_{\text{tot}}$. Each user is assumed to experience an AWGN channel, whose capacity, $C_n(P_n)$, is given by

$$C_n(P_n) = B \cdot \log_2\left(1 + \frac{P_n}{\sigma^2_{\text{n},n}}\right), \tag{22}$$

where B and $\sigma^2_{\text{n},n}$ denote bandwidth available and receiver noise power, respectively.

It is assumed that private information of each user, including $R_{0n}, D_{0n}, \mu_n, \sigma^2_{\text{n},n}$, are sent to CSM, where power allocation is made. Then CSM sends back the decision of power allocated to each user. Note that, more complicated PHY layer setups can also be taken into account, such as multicarrier and multiple antennas systems over Rayleigh fading channels. However, employing simple PHY layer setup would help to highlight the focus of this paper, investigating optimal and fair criteria for RRA. It is worth mentioning that the co-opetition strategy can be easily extended to other scenarios.

4.2. Co-Opetition Strategy.

4.2.1. CO-opetition Formulation. According to the common sense in the field of video signal processing, the PSNR threshold can be set to different values, such as 40 dB, 35 dB, or 32 dB, representing perfect, good and acceptable video quality, respectively. The PSNR threshold can also be set dynamically according to the total resources available, the number of users, and so forth. As an illustration, we choose QoS threshold as PSNR = 35 dB corresponding to good video quality, that is, $U_{\text{th}} = 35$ dB in (4). Denote $\vec{P}_{\text{th}}$ as $(P_{1,\text{th}},\dots,P_{N,\text{th}})$ representing power required by users to achieve PSNR of 35 dB. Using co-opetition strategy, if $\text{sum}(\vec{P}_{\text{th}}) \le P_{\text{tot}}$($\text{sum}(\vec{P}_{\text{th}})$ means calculating the sum of all members in $\vec{P}_{\text{th}}$, i.e., $\sum_{n=1}^{N} P_{n,\text{th}}$.) , the lower and upper bounds of achievable PSNR are set at $U_{\text{low}} = 35$dB and $U_{\text{upp}} = \infty$, respectively, and $U_{\text{low}} = -\infty$ and $U_{\text{upp}} = 35$ dB otherwise. Correspondingly, when we have $\text{sum}(\vec{P}_{\text{th}}) \le P_{\text{tot}}$, lower and upper bounds of rates are $\vec{r}_{\text{low}} = (r_{1,\text{th}},\dots,r_{N,\text{th}})$ and $\vec{r}_{\text{upp}} = \infty$, respectively, and $\vec{r}_{\text{low}} = (R_{01},\dots,R_{0\text{N}})$ and $\vec{r}_{\text{upp}} = (r_{1,\text{th}},\dots,r_{N,\text{th}})$ otherwise. In this paper, it is easy to calculate $P_{n,\text{th}}, r_{n,\text{th}}$ corresponding to PSNR threshold, for both (21) and (22) are invertible and monotonic increasing functions. Thus, given PSNR threshold, $\text{sum}(\vec{P}_{\text{th}}) \le P_{\text{tot}}$ or not can be easily determined, and consequently, both $\vec{r}_{\text{low}}$ and $\vec{r}_{\text{upp}}$ are known.

Given each user's utility definition in (21) and (22), system utility writes

$$U_{\text{s}}\left(\vec{P}\right) = 10\sum_{n=1}^{N} \log \frac{255^2(C_n(P_n) - R_{0n})}{D_{0n}(C_n(P_n) - R_{0n}) + \mu_n}, \tag{23}$$

where $C_n(P_n)$ refers to as r_n. We assume that capacity approaching channel codes is employed at PHY layer. Then our co-opetition strategy writes

$$\begin{aligned} \max \quad & U_{\text{s}}\left(\vec{P}\right), \\ \text{s.t.} \quad & \sum_{n=1}^{N} P_n \le P_{\text{tot}}, \\ & \vec{r}_{\text{low}}, \le \vec{C} \le \vec{r}_{\text{upp}} \end{aligned} \tag{24}$$

where $\vec{C} = (C_1(P_1),\dots,C_N(P_N))$. Note that (24) has the same form as (8). The first constraint on the sum of the power (24) corresponds to $\vec{r} \in \mathcal{R}$ in (8).

4.2.2. The Implement of Co-opetition. Using LOD, maximization of (24) can be decomposed into

$$\begin{aligned} \max_{\vec{c}} & \sum_{n=1}^{N} 10\log \frac{255^2(c_n - R_{0n})}{D_{0n}(c_n - R_{0n}) + \mu_n} \\ & + \sum_{n=1}^{N} ((\lambda'_n - \lambda_n)c_n - \lambda'_n r_{n,\text{low}}) \end{aligned} \tag{25}$$

where $\vec{c} = (c_1,\dots,c_N)$, and

$$\begin{aligned} & \max B\sum_{n=1}^{N} \lambda_n \log_2\left(1 + \frac{P_n}{\sigma_{\text{n},n}^2}\right), \\ & \text{s.t.} \sum_{n=1}^{N} P_n \le P_{\text{tot}} \\ & P_n \le P_{n,\text{upp}}, \ \forall n \end{aligned} \tag{26}$$

where $P_{n,\text{upp}}$ is defined as the upper bound of transmit power of user n corresponding to $r_{n,\text{upp}}$.

The optimum variable of (25), $\vec{c}^* = (c_1^*,\dots,c_N^*)$, can be obtained by simply making the partial derivative ofg_A and let it equal to 0,

$$D_{0n}(c_n - R_{0n})^2 + \mu_n(c_n - R_{0n}) - \frac{10\mu_n}{(\lambda_n - \lambda'_n)\ln 10} = 0, \quad \forall n. \tag{27}$$

Then we have

$$c_n^* = R_{0n} + \frac{\sqrt{\mu_n^2 + 4D_{0n} \cdot \text{tmp}} - \mu_n}{2D_{0n}}, \tag{28}$$

where $\text{tmp} = 10\mu_n/(\lambda_n - \lambda'_n)$.

As mentioned in Section 3.3, (26) can be solved at PHY layer by the weighted sum rate maximization with thecon-straints of total and individual power. Note that $C_n(P_n)$ in (22) is concave and increasing with respect to P_n, thus the item to be maximized in (26) is also concave increasing. The domain of (26) is formed by two linear inequalities, each of which forms a convex domain together with $P_n \ge 0, \forall n$. Thus the domain of (26) is also convex, and (26) is accessible to conventional convex optimization techniques, such as feasible direction method and projected gradient method. In this paper the feasible increasing direction method is employed (see the Appendix for details).

So far, given fixed $\vec{\lambda}, \vec{\lambda'}$, two subproblems, (25) and (26), have been solved. We denote the optimal values of them with $g_A^*(\vec{\lambda}, \vec{\lambda'})$ and $g_P^*(\vec{\lambda})$, respectively. In the following, the optimum $\vec{\lambda}, \vec{\lambda'}$, denoted by $\vec{\lambda}^*, \vec{\lambda'}^*$, will be determined such that the sum of $g_A^*(\vec{\lambda}, \vec{\lambda'})$ and $g_P^*(\vec{\lambda})$ is minimized, that is,

$$\left(\vec{\lambda}^*, \vec{\lambda'}^*\right) = \arg\min_{\vec{\lambda}, \vec{\lambda'}} g_A^*\left(\vec{\lambda}, \vec{\lambda'}\right) + g_P^*\left(\vec{\lambda}\right). \tag{29}$$

Note that, the dual function might not be differentiable or, in other words, (29) is not accessible to classical computational method, such as steepest descent method. In this paper we employ the sub-gradient method, which applies to both differentiable and nondifferentiable dual functions. Much like the feasible increasing direction method, sub-gradient method also searches the optimal $\vec{\lambda}$ and $\vec{\lambda'}$ iteratively. The main iteration writes

$$\begin{pmatrix} \vec{\lambda}^{k+1} \\ \vec{\lambda'}^{k+1} \end{pmatrix} = \begin{pmatrix} \vec{\lambda}^{k} \\ \vec{\lambda'}^{k} \end{pmatrix} - \alpha^k \vec{g}^k, \tag{30}$$

Table 1: test video sequences (videoID, video type, temporal level (TL), frame rate).

ID	Video sequence	μ	D_0	R_0
1	Foreman (CIF, TL = 4, 30 Hz)	5232400	0	0
2	Coastguard (CIF, TL = 4, 30 Hz)	6329700	4.3	0
3	Mobile (CIF, TL = 4, 30 Hz)	38230000	1	44040
4	Foreman (QCIF, TL = 4, 30 Hz)	2653300	0	19614
5	Foreman (CIF, TL = 4, 15 Hz)	2760000	1	20720
6	Foreman (CIF, TL = 2, 30 Hz)	4610000	3	55080

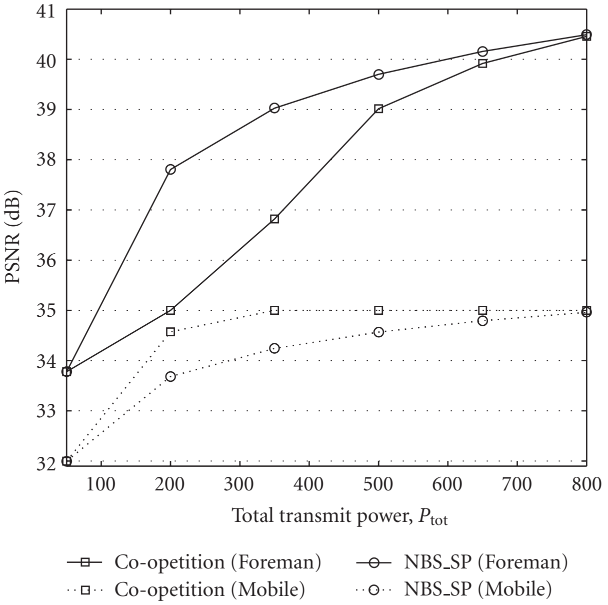


Figure 2: Plot of individual PSNRs achieved by the co-opetition, NBS SP. User 1: Foreman (CIF, TL= 4, 30 Hz), user 2: Mobile (CIF, TL = 4, 30 Hz).

where α^k is the step-size which can be set as constant, and $\vec{g}^k$ denotes the sub-gradient at $(\vec{\lambda}^k, \vec{\lambda'}^k)$. Note that, $\vec{P} = (P_1, \dots, P_N)^{\mathrm{T}}$ at $(\vec{\lambda}^k, \vec{\lambda'}^k)$ rightly forms a sub-gradient, so the sub-gradient can be obtained almost without any cost.

4.3. Numerical Results. In this subsection, the proposed co-opetition strategy (co-opetition) is evaluated by comparing with the strategy proposed in [1], which allocates resources using the Nash bargaining Solution of Same bargaining Power (NBS_SP). For the sake of comparison, we use the same test sequences as those in [1], and we list the parameters in Table I for reader's convenience.

4.3.1. Comparison in Terms of Individual PSNR. In this experiment we focus on individual PSNRs in the case of two users. At APP layer, user 1 transmits Foreman sequence of CIF resolution at 30 Hz, and user 2 transmits Mobile sequence of CIF resolution at 30 Hz. At PHY layer, we set the bandwidth to $B = 250$ kHz, and let the receiver noise power to be $\sigma_{n,1}^2 = 50$ and $\sigma_{n,2}^2 = 1$ for user 1 and user 2, respectively.

```
Set $k = 1$ and $P_n^k = 0,\ \forall n$, Precision $\varepsilon = 10^{-4}$
Repeat:
    Determine $\nabla g_P^k$ using(A.1)
    Determine $\vec{d}^k$ according (A.4) and(A.5)
    Determine $\alpha^k$ using(A.6)
    Compute $\vec{P}^{k+1}$ using(A.8)
Until: $|(\nabla g_P^k)^{\mathrm{T}} \vec{d}^k| \le \varepsilon$.
```

Algorithm 1: Feasible increasing direction method.

Total transmit power P_{tot} varies from 50 to 800. Figure 2 shows the individual PSNRs achieved by these two schemes. If NBS_SP is employed, user 1 can achieve higher PSNR that user 2 or, in other words, it is very hard for user 2 to achieve satisfying video quality (PSNR $\ge$ 35). In the case of $P_{\text{tot}} \ge 200$, user 1 can always be satisfied. Note in this case, user 1's video satisfaction degree increases very slowly as the PSNR increases, but significantly for user 2. Taking this observation into account, co-opetition imposes individual constraint on each user (see (4)). For example, with $P_{\text{tot}} = 200$, which can not satisfy two users simultaneously, co-opetition decreases user 1's PSNR to 35 dB, and consequently, user 2's PSNR achieves an improvement about 1 dB. If have $350 \le P_{\text{tot}} \le 650$, user 2's PSNR is improved such that user 2 is just satisfied. Note, in these two cases, co-opetiton keeps user 1 satisfied, while user 2 either be satisfied or achieve much QoS improvement. It is worth to mention that, under a given total transmit power constraint, NBS_SP can achieve higher total PSNR of two users than that in co-opetition. This is because the NBS_SP maximizes the sum of PSNRs without taking the individual PSNR constraints into account. The co-opetition works in quite a different way. It maximizes the sum of PSNRs under the constraints of individual PSNR. Therefore, the co-opetition is not only optimal (As stated in Section 1, in this paper the optimal means sum utility maximization under certain constraints, differing from unconstrained optimization.) , but also fairer than NBS_SP. This argument is further verified with other experiments

4.3.2. Comparison in Terms of the Number of Satisfied Users and Minimum PSNRs. We study a more complicated scenario with nine users, each transmitting a sequence randomly selected from Table 1. They also experience different

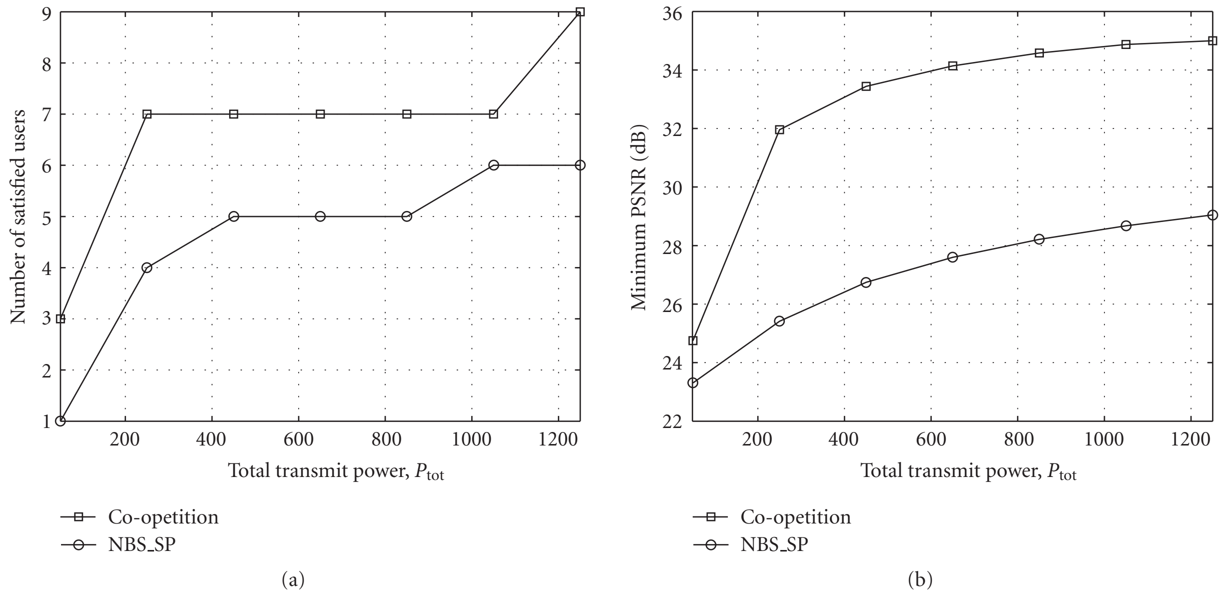


Figure 3: Plot of the number of satisfied users (a) and minimum PSNRs (b) achieved by co-opetition and NBS_SP in the case of nine users. Id of sequences transmitted are 3, 6, 1, 3, 5, 1, 3, 2, 2, respectively. These sequences are randomly selected from Table 1. Bandwidth B is set to 400 KHz for all users, and the receiver noise power are set to 16, 7, 5, 1, 19, 12, 24, 12, 11, respectively, again by random generation.

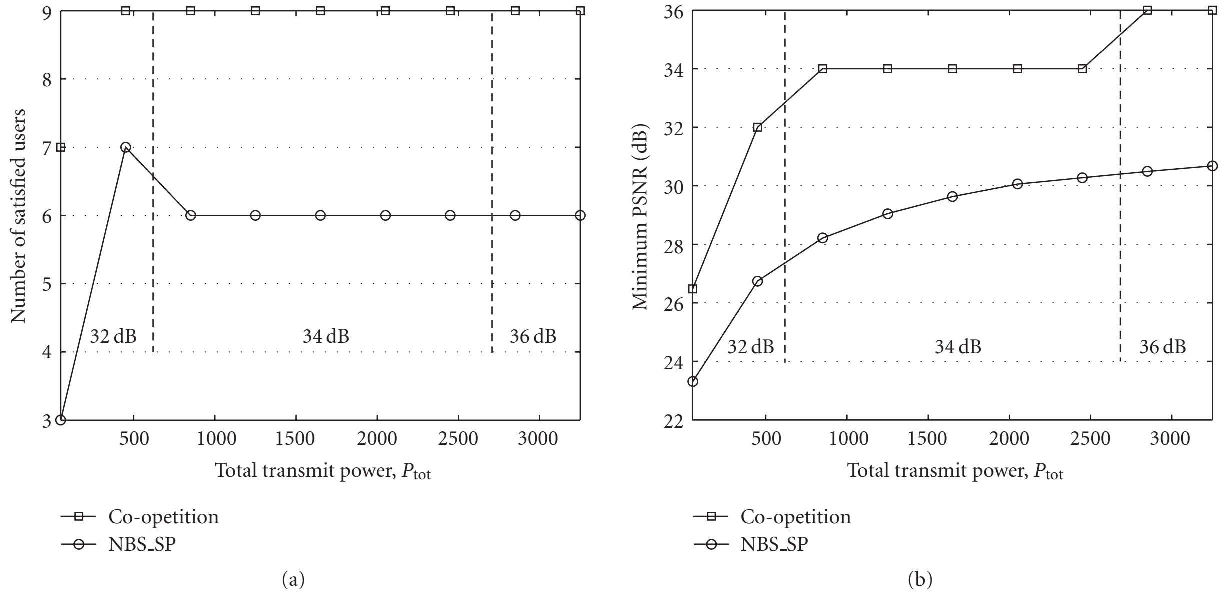


Figure 4: Plot of the number of satisfied users (a) and minimum PSNRs (b) achieved by NBS_SP and adaptive co-opetition. System setup is the same as that of Figure 3. 32 dB, 34 dB, and 36 dB refer to PSNR thresholds corresponding to different P_{tot}.

receiver noises randomly generated from 0 to 25. Figure 3 shows the number of satisfied users and the minimum PSNRs achieved by NBS_SP and co-opetition. We observe that, the co-opetition always outperforms the NBS_SP. For example, in the case of $P_{\text{tot}} = 1250$, co-opetition can make all users satisfied, but only 6 users satisfied by NBS_SP. With respect to the minimum PSNR, which is an important criteria evaluating system in the worst case, improvement of around 6 dB can be achieved when $P_{\text{tot}} \geq 200$. Note that, NBS_SP can only make minimum PSNRs from about 25 dB to 29 dB, corresponding to poor video quality, while above 32 dB for co-opetition leading to acceptable video quality. Recall that, the co-opetition implies a judicious mixture of competition and cooperation. Through competition, the best system efficiency can be achieved. However, pure competition, for example, NBS_SP, might make very high PSNRs for some users, for example, users transmitting simple video content or having good channel quality, but low PSNRs for the others. This disadvantage is eliminated by co-opetition through introducing cooperation among users.

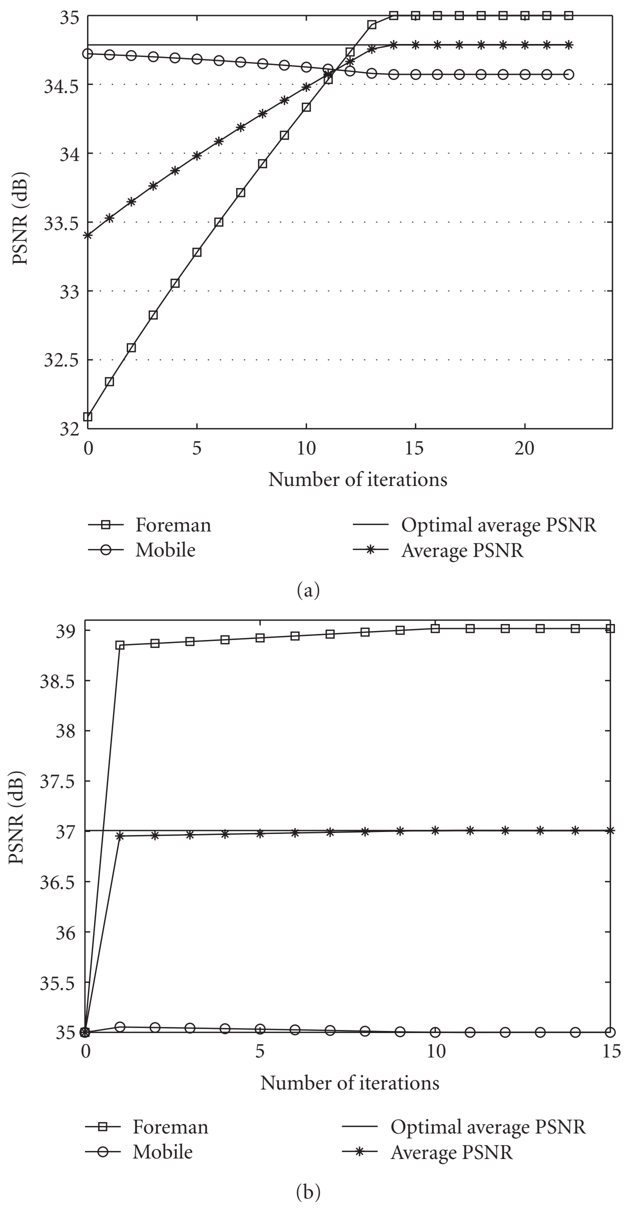


FIGURE 5: Plot of individual PSNRs and average PSNR. User 1: Foreman (CIF, TL = 4, 30 Hz), user 2: Mobile (CIF, TL = 4, 30 Hz). (a): $P_{\text{tot}} = 200$ and (b): $P_{\text{tot}} = 500$.

Again, this experiment indicates that co-opetition provides a good tradeoff between system efficiency and fairness.

4.3.3. Adaptive Co-opetiton Strategy. In previous experiments, the threshold PSNR is fixed to be 35 dB. In order to consider more fairness in resource allocation, adaptive threshold can be employed. As an illustration, we present a simple method to set the threshold PSNR. More optimal and fair scheme for determining the threshold PSNR will be investigated in our future work. We employ PSNR = 32 dB, 34 dB and 36 dB to represent acceptable, good and very good quality, respectively. Denote resources required by the three levels with R_a, R_g, R_v, then threshold PSNR, PSNR_{th}, can be determined as follows

$$\begin{aligned}
\text{PSNR}_{\text{th}} &= 32\,\text{dB}, \quad \text{if } R_{\text{tot}} < R_g, \\
\text{PSNR}_{\text{th}} &= 34\,\text{dB}, \quad \text{if } R_g \leq R_{\text{tot}} \leq R_v, \qquad (31)\\
\text{PSNR}_{\text{th}} &= 36\,\text{dB}, \quad \text{if } R_g \leq R_{\text{tot}} \leq R_v,
\end{aligned}$$

where R_{tot} is denote as total resources available.

Same system setup as that in previous experiment is used. We observe from Figure 4(a) that, co-opetition employing adaptive PSNR_{th} still outperforms the NBS_SP. Moreover, adaptive PSNR_{th} is more concerned with fairness than that using fixed threshold. For example, in the case of low resource, for example, $P_{\text{tot}} \leq 500$, $\text{PSNR}_{\text{th}} = 32$ dB is selected. Consequently, an improvement of about 3 dB and 2 dB can be achieved for the minimum PSNRs compared to NBS_SP and co-opetition using fixed threshold (see Figure 3(b)), respectively. Note, these improvements are significantly important for users having low PSNRs. Although these improvements come from further decreasing the maximum achievable PSNR, it can provide fairer resource allocation. For instance, in Figure 4(a), it is very easy for all users to achieve similar quality level using co-opetition. Moreover, PSNR_{th} can also be set to a very high level, for example, 36 dB in the case of $P_{\text{tot}} > 2500$. An important advantage of this is that all users can be guaranteed high video quality, but cannot by fixed PSNR threshold and NBS_SP.

4.3.4. Optimality Verification. Our co-opetition is also optimal. As stated in Section 1, optimal means sum utility maximization (SUM) under individual constraints. The optimality is verified by experimental analysis in the case of two users. Results of two examples of them are shown in Figure 5(a) and Figure 5(b). System setup is the same as that in Figure 2. The optimal average PSNRs are achieved by exhaustive search. Recall that the LOD method consists of inner and outer iterations. In each inner iteration, the power allocation is initiated corresponding to (R_{01}, R_{02}) for Figure 5(a) and $(r_{1,\text{th}}, r_{2,\text{th}})$ for Figure 5(b). In the outer iteration, the values of $\vec{\lambda}$ and $\vec{\lambda\prime}$ are initialized randomly. Figures 5(a) and 5(b) show the results of outer iterations.

From these two figures, we can see that our strategy is optimal under individual constraints. In Figure 2, $P_{\text{tot}} = 200$ cannot satisfy two users simultaneously. Therefore the PSNR of user 1 is pegged at the threshold PSNR = 35 dB. The optimal average PSNR can be achieved after 14 iterations. In Figure 5(b), $P_{\text{tot}} = 500$ can make satisfying PSNR for both the two users. We observe that, user 2's PSNR has only little fluctuation, and converges to the threshold. At the optimal power allocation, both the two users' PSNRs are above or equal to the threshold. All these coincide with the results in Figure 2.

4.3.5. Summarization. To summarize, threshold PSNR plays importantly in adaptive/nonadaptive co-opetition strategies. It provides radio resource allocation (RRA) with more flexible tradeoff between system efficiency and fairness among users.

5. Conclusion

In this paper, we have presented an optimal and fair co-opetition strategy for multiuser multimedia RRA. Following contributions and conclusions have been made and drawn

(1) We formulate the co-opetition strategy as sum utility maximization under constraints from both APP and PHY layers. APP layer constraints imply that co-opetition takes the QoS satisfaction degree into account in RRA.

(2) We show that the co-opetition strategy can be implemented efficiently through applying the LOD method. Therefore the co-opetition strategy can easily apply to real time multimedia services.

(3) We apply the co-opetition strategy to power allocation among multiple video users. Numerical results indicate that co-opetition can result in an improved number of satisfied users and significant improvement in minimum PSNRs as well. A simple method for adaptively determining threshold PSNR is also presented, such that fairer resource allocation can be achieved.

(4) We conclude that co-opetition, that is, mixture of cooperation and competition, is more applicable to multiuser multimedia RRA than pure competition based strategy. Co-opetition strategy is not only optimal, but also fair.

Our future work is to design more feasible co-opetition strategy for different system setups, including multicarrier and multiple antennas systems. We also wish to extend our preliminary work to future heterogenous network, in which users not necessarily run in a collaborative way.

Appendix

Feasible Increasing Direction Method

Feasible Increasing direction method iteratively searches the optimum variable, $\vec{P}^* = (P_1^*,\dots,P_N^*)$, by in each iteration selecting a feasible increasing direction and update step size. Denote $\vec{P^k} = (P_1^k,\dots,P_N^k)$ as power allocation in the k_{th} iteration, then $\vec{P^k}$ satisfies the constraints in (26). Denote $\vec{d^k} \in \mathcal{R}^N$, α^k as the direction and step size employed in the k_{th} iteration, then $\vec{d^k}$, α^k and $\vec{P^{k+1}}$ can be determined as follows.

Denote $g_P(\vec{P})$ as the item to be maximized in (26), then the gradient of $g_P(\vec{P})$ at $\vec{P^k}$, denoted with ∇g_P^k, writes

$$\nabla g_P^k = \left(\frac{\partial g_P^k}{\partial P_1},\dots,\frac{\partial g_P^k}{\partial P_N}\right)^{\mathrm{T}}, \tag{A.1}$$

where

$$\frac{\partial g_P^k}{\partial P_n} = \frac{B\lambda_n}{\left(\sigma_{\mathrm{n},n}^2 + P_n\right)\ln 2}. \tag{A.2}$$

If $\vec{P^k}$ is strictly feasible, that is,

$$\begin{gathered}\sum_{n=1}^{N} P_n < P_{\mathrm{tot}} \\ P_n < P_{n,\mathrm{upp}}, \quad n \in \{1,\dots,N\}\end{gathered} \tag{A.3}$$

then set

$$\vec{d^k} = \nabla g_P^k. \tag{A.4}$$

Otherwise, denote $\mathcal{I}(\vec{P^k})$ as set of indexes of active constraints, for example, if $P_n = P_{n,\mathrm{upp}}, 1 \le n \le N$, then $n \in \mathcal{I}(\vec{P^k})$. $0 \in \mathcal{I}(\vec{P^k})$ refers to $\sum_{n=1}^{N} P_n = P_{\mathrm{tot}}$. Then $\vec{d^k}$ can be obtained by solving following maximization through linear programming,

$$\begin{aligned} &\max \left(\nabla g_P^k\right)^{\mathrm{T}} \vec{d^k} \\ &\text{s.t. } d_n \le 0, \forall n \in \mathcal{I}\left(\vec{P^k}\right), \\ &\quad \sum_{n=1}^{N} d_n \le 0, \text{ if } 0 \in \mathcal{I}\left(\vec{P^k}\right) \\ &\quad -1 \le d_n \le 1,\ n \in \{1,\dots,N\}. \end{aligned} \tag{A.5}$$

If $(\nabla g_P^k)^{\mathrm{T}} \vec{d^k} = 0$, then $\vec{P^k}$ is optimal. Otherwise, compute α^k by solving following one-dimension maximization,

$$\begin{aligned} &\max \phi\left(\alpha^k\right) = g_P\left(\vec{P^k} + \alpha^k \vec{d^k}\right) \\ &\text{s.t. } 0 \le \alpha^k \le \alpha_{\max}, \end{aligned} \tag{A.6}$$

where

$$\alpha_{\max} = \begin{cases} +\infty, & \text{if } \sum_{n=1}^{N} d_n \le 0,\ d_n^k \le 0,\ \forall n, \\ \min\left\{\dfrac{P_{\mathrm{tot}} - \sum_{m=1}^{N} P_m^k}{\sum_{m=1}^{N} d_m}, \dfrac{P_{n,\mathrm{upp}} - P_n^k}{d_n^k}\right\}, & \text{if } \quad 0, n \notin \mathcal{I}\left(\vec{P^k}\right), \\ \min\left\{\dfrac{P_{n,\mathrm{upp}} - P_n^k}{d_n^k}\right\}, & \text{if } \quad 0 \in \mathcal{I}\left(\vec{P^k}\right), \quad n \notin \mathcal{I}\left(\vec{P^k}\right). \end{cases} \tag{A.7}$$

Given $\vec{d^k}$ and α^k, $\vec{P^{k+1}}$ can be set as

$$\vec{P^{k+1}} = \vec{P^k} + \alpha^k \vec{d^k}. \tag{A.8}$$

Then the feasible increasing direction method can be summarized in Algorithm 1.

Acknowledgment

This work was supported by NSFC (No. 60672036, No. 60832008) and Key Project of Provincial Scientific Foundation of Shandong (No. Z2008G01).

References

[1] H. Park and M. van der Schaar, “Bargaining strategies for networked multimedia resource management,” *IEEE Transactions on Signal Processing*, vol. 55, no. 7, part 1, pp. 3496–3511, 2007.

[2] C. Shen and M. van der Schaar, “Optimal resource allocation in wireless multiaccess video transmissions,” in *Proceedings of the IEEE International Conference on Communications (ICC ’07)*, pp. 4581–4586, Glasgow, UK, June 2007.

[3] Y. Su and M. van der Schaar, “Multiuser multimedia resource allocation over multicarrier wireless networks,” *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 2102–2116, 2008.

[4] L.-U. Choi, W. Kellerer, and E. Steinbach, “On cross-layer design for streaming video delivery in multiuser wireless environments,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2006, Article ID 60349, 10 pages, 2006.

[5] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, and W. Kellerer, “Application-driven cross-layer optimization for video streaming over wireless networks,” *IEEE Communications Magazine*, vol. 44, no. 1, pp. 122–130, 2006.

[6] J. Brehmer and W. Utschick, “A decomposition of the downlink utility maximization problem,” in *Proceedings of the 41st Annual Conference on Information Sciences and Systems (CISS ’07)*, pp. 437–441, Baltimore, Md, USA, March 2007.

[7] J. Brehmer and W. Utschick, “Utility maximization strategies in the multi-user MIMO downlink,” in *Proceedings of the 1st International Workshop on Cross Layer Design (IWCLD ’07)*, pp. 86–90, Jinan, China, September 2007.

[8] R. Böhnke and K.-D. Kammeyer, “Weighted sum rate maximization for the MIMO-downlink using a projected conjugate gradient algorithm,” in *Proceedings of the 1st International Workshop on Cross Layer Design (IWCLD ’07)*, pp. 82–85, Jinan, China, September 2007.

[9] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala, “RSVP: a new resource reservation protocal,” *IEEE Network*, vol. 7, no. 5, pp. 8–18, 1993.

[10] M. van der Schaar and S. Shankar N, “Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms,” *IEEE Wireless Communications*, vol. 12, no. 4, pp. 50–58, 2005.

[11] F. Kelly, “Charging and rate control for elastic traffic,” *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[12] F. P. Kelly, A. K. Maulloo, and D. Tan, “Rate control for communication networks: shadow prices, proportional fairness and stability,” *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[13] A. M. Brandenburger and B. J. Nalebuff, *Co-Opetition: A Revolution Mindset That Combines Competition and Cooperation: The Game Theory Strategy That’s Changing the Game of Business*, Currency Doubleday, New York, NY, USA, 1997.

[14] A. Larcher, H. Sun, M. van der Shaar, Z. Ding, et al., “Decentralized transmission strategy for delay-sensitive applications over spectrum agile network,” in *Proceedings of International Packet Video Workshop*, Irvine, Calif, USA, December 2004.

[15] Z. Guan, D. Yuan, and H. Zhang, “Novel coopetition paradigm based on bargaining theory or collaborative multimedia resource management,” in *Proceedings of the 9th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC ’08)*, pp. 1–5, Cannes, France, September 2008.

[16] K. Stuhlmüller, N. Färber, M. Link, and B. Girod, “Analysis of video transmission over lossy channels,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1012–1032, 2000.

[17] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, “Layering as optimization decomposition: a mathermatical theory of netowrk architectures,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, 2007.

[18] S. Shi, M. Schubert, and H. Boche, “Weighted sum-rate optimization for multiuser MIMO systems,” in *Proceedings of the 41st Annual Conference on Information Sciences and Systems (CISS ’07)*, pp. 425–430, Baltimore, Md, USA, March 2007.

[19] J.-L. Goffin and J.-P. Vial, “Convex nondifferentiable optimization: a survey focussed on the analytic center cutting plane method,” Tech. Rep. 99.02, Logilab, University of Geneva, Geneva, Switzerland, 1999.

[20] J. A. Nossek, “CLARA: cross layer assisted resource allocation—theory and applications,” in *Proceedings of the 1st IEEE International Workshop on Cross Layer Design (IWCLD ’07)*, Jinan, China, September 2007.

[21] R. D. Yates, “A framework for uplink power control in cellular radio systems,” *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341–1347, 1995.

[22] Y. Andreopoulos, A. Munteanu, J. Barbarien, M. van der Schaar, J. Cornelis, and P. Schelkens, “In-band motion compensated temporal filtering,” *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 653–673, 2004.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 163826, 6 pages
doi:10.1155/2009/163826

Research Article

Performance Analysis of SNR-Based Scheduling Policies in Asymmetric Broadcast Ergodic Fading Channel

Jesús Pérez, Jesús Ibáñez, Javier Vía (EURASIP Member), and Ignacio Santamaría (EURASIP Member)

Department of Communications Engineering, University of Cantabria, Avda. de los Castros s/n, 39005 Santander, Spain

Correspondence should be addressed to Jesús Pérez, jperez@gtas.dicom.unican.es

Received 1 July 2008; Accepted 26 January 2009

Recommended by Kwang-Cheng Chen

We analyze the performance of SNR-based scheduling algorithms in broadcast ergodic fading channels where multiuser selection diversity is exploited. At each channel state, the user with the highest weighted signal-to-noise ratio is selected to be transmitted. The use of weights associated to the users allows us to control the degree of fairness among users and to arrange them according to a prescribed quality of service. These weights parametrize the scheduling algorithms so each set of weights corresponds to a specific scheduling algorithm. Assuming Rayleigh fading broadcast channel, we derive a closed-form expression for the achievable user's rates as a function of the scheduling algorithm, the channel fading statistics of each user, and the transmit power. With the help of this expression, we solve some interesting inverse problems. For example, for a given arbitrary channel statistics we obtain the optimum scheduling algorithm to achieve a prescribed set of users' rates with minimum transmit power.

1. Introduction

It is well known that the capacity region of broadcast ergodic fading channels is achieved by superposition coding at the transmitter and successive interference cancellation at the receivers (SC-SIC). Using SC-SIC, the transmitter transmits simultaneously to all users using multiresolution coding, and the receivers perform successive decoding. Although optimum in terms of capacity, SC-SIC is complex, and it is not necessarily the best method to use in practical systems because decoding and channel estimation errors can degrade its performance significantly [1].

More feasible are the orthogonal TDMA strategies based on users opportunistic scheduling, where a single user is selected to be transmitted at each fading state. Once a user is selected, the transmitter allocates all the available resources to him (bandwidth and power) utilizing a code adapted to the channel state. Since the channels between the base station and the users usually fade independently, this scheme effectively exploits the multiuser diversity inherent to the broadcast (BC) channel (see, e.g., [2] and references therein). Opportunistic scheduling is commonly used in modern wireless standards as IS-856 (also called CDMA 2000 1xEV-DO), mobile WIMAX, and HSPA [3–5].

In multiuser diversity, the resulting long-term users' rates are determined by the specific scheduling policy. Many criteria have been proposed to schedule the users. Among them, we focus on the so-called SNR-based scheduling policies where the user with the highest weighted signal-to-noise ratio (SNR) is selected to be transmitted. A particular case is the so-called "absolute SNR-based scheduling" (ASS) [6], where the user with the highest channel gain at each channel state is selected. It is well known that ASS maximizes the overall throughput (sum-rate) [2]. Although ASS achieves the sum-rate, it favors users who have good average channel conditions producing quite different individual users' rates in asymmetric broadcast channels. On the other hand, the "normalized SNR-based scheduling" (NSS) schedules the users according to the instantaneous channel gain normalized by its own average [6, 7]. NSS strategy favors users with poor average channel conditions and penalizes advantaged users producing similar users' rates but at expense of a lower overall throughput. In fact, there is a tradeoff between maximizing the overall throughput and achieving throughput fairness. Other common scheduling criterion is based on the instantaneous achievable rates instead on the SNR. In this case the base station transmits to the user with the highest normalized achievable rate [2, 8, 9]. Since the achievable rate is monotonically increasing with the SNR, both scheduling criteria are interchangeable. Further, in BC channels the power constraint at the base station is usually based on the maximum power rather than the long-term average power. Therefore, we assume that the transmit power is constant.

Some performance analyses of opportunistic scheduling can be found in the technical literature. In [6, 7] closed-form expressions for the achievable rates using ASS and NSS are derived. In [10] analytical expressions for the sum-rate of BC channel are derived using ASS and considering different adaptive power allocation strategies. All these performance analyses are restricted to specific scheduling algorithms.

In this work we derive a general closed-form expression for the rates achievable by any SNR-based scheduling algorithm. It generalizes other expressions proposed in the technical literature that are restricted to a single specific scheduling strategy (e.g., ASS and NSS). Each scheduling algorithm is parameterized by a set of weights assigned to the users, so the user with the best weighted channel is selected at each channel fading state. There is a point-to-point correspondence between the scheduling weights and the boundary points of the achievable rates region. The derived expression explicitly describes this relationship. The expression is a simple function of the channel fading parameters, the transmitted power, and the scheduling weights. With the help of this function we solve some interesting inverse problems. For example, the computation of the minimum required transmit power and the optimum scheduling strategy to achieve a given users' rates. Other problem considered is the computation of the optimum scheduling preserving a given relationship among the users' rates for a given transmit power. These inverse problems are formulated as systems of nonlinear equations involving the derived expression.

The rest of the paper is organized as follows. Section 2 shows the BC ergodic channel model. Section 3 presents the parametrization of the SNR-based scheduling policies, where the ASS and the NSS are particular cases. In Section 4 we derive the closed-form expression for the achievable users' rates as a function of the channel fading statistics, the transmit power, and the scheduling algorithm. In Section 5 we pose the inverse problems as set of nonlinear equations involving the derived expression. Simulation results are presented in Section 6. Finally, conclusions are drawn in Section 7.

2. Channel Model

A narrowband broadcast channel with K users is considered. We assume that the transmitter and receivers have a single antenna. The transmitter is subject to an average power constraint denoted by P. We assume independent and identically distributed (i.i.d.) AWGN noise at the Rx antennas, with single-sided power spectral density denoted by N_0 for all users. The receivers' bandwidth is denoted by B, so the noise power at the receivers is BN_0. The baseband-equivalent channel response between the transmitter and the kth user is denoted by h_k, $k = 1,\dots,K$. We assume that the h_k are independent and differently distributed (i.d.d.) zero-mean circularly symmetric complex Gaussian (ZMCSCG) random variables. Then, the channel power gains $g_k = |h_k|^2$ will be exponentially distributed with cumulative distribution functions (c.d.f.) given by

$$F_k(x) = 1 - \exp\left(-\frac{x}{\overline{g}_k}\right), \quad x \geq 0, \tag{1}$$

where $\overline{g}_k$ denotes the average power gain for the kth user channel: $\overline{g}_k = E\{g_k\}$. The probability density functions (p.d.f.) will be

$$f_k(x) = \frac{\exp\left(-x/\overline{g}_k\right)}{\overline{g}_k}, \quad x \geq 0. \tag{2}$$

We assume, without loss of generality, that the channel is normalized so $\overline{\mathbf{g}}^T\mathbf{1} = K$, where $\overline{\mathbf{g}} = [\overline{g}_1, \overline{g}_2, \dots, \overline{g}_K]^T$ and $\mathbf{1}$ is the all-ones vector of size K. Under this normalization, the SNR averaged for all users and fading states will be $\rho = P/BN_0$. Note that the average SNR and the transmit power are interchangeable.

3. SNR-Based Scheduling

The SNR-based scheduling strategies can be parameterized by a set of normalized weights associated with the users, so the system selects the user with the highest weighted channel response.

The set of all possible weight vectors is the subset in $\mathfrak{R}^K$ given by

$$S_{\mathbf{w}} = \{\mathbf{w} = [w_1 w_2 \cdots w_K]^T \mid w_s > 0, \mathbf{w}^T\mathbf{1} = K\}. \tag{3}$$

Then, at each channel state, the system selects the user according to $\arg\max_s\{\eta_s\}$, where $\eta_s = w_s g_s$.

In particular, the ASS and the NSS algorithms correspond to $\mathbf{w} = \mathbf{1}$ and $\mathbf{w} = a\mathbf{1} \cdot / \overline{\mathbf{g}}$, respectively, where a is a normalization factor to fulfill the constrain of (3), and $\cdot/$ denotes elementwise division.

Different scheduling weights lead to different achievable users' rates. Therefore, there is a one-to-one correspondence between all the possible weight vectors and the points on the boundary of the rates region. The achievable rates using ASS and NSS are two of such points.

4. Achievable Rates

Let us define the following effective channel gain for the sth user:

$$g_s^* = \begin{cases} 0, & \eta_s < \eta_{-s} \\ g_s, & \eta_s > \eta_{-s}, \end{cases} \tag{4}$$

where $\eta_{-s} = \max_{k\neq s}\{w_k g_k\}$. The p.d.f. of g_s^* can be expressed as follows:

$$f_s^*(x) = \text{Prob}\{\eta_s < \eta_{-s}\}\delta(x) + f_s(x)\widetilde{F}_{-s}(xw_s), \qquad (5)$$

where $\delta(x)$ is the Dirac delta function, $f_s(x)$ is given by (2), and $\widetilde{F}_{-s}(x)$ is the c.d.f. of η_{-s} given by

$$\begin{aligned}\widetilde{F}_{-s}(x) &= \prod_{k\neq s}^{K} F_k\left(\frac{x}{w_k}\right)\\ &= \prod_{k\neq s}^{K}\left[1-\exp\left(-\frac{x}{(\bar{g}_k w_k)}\right)\right].\end{aligned} \qquad (6)$$

This expression can be expressed as follows:

$$\widetilde{F}_{-s}(x) = \sum_{\mathbf{i}\in S} c_{\mathbf{i}}(1-i_s)\exp\left(-x\mathbf{q}^T\mathbf{i}\right), \qquad (7)$$

where S is the set of binary words of length K, $c_{\mathbf{i}} = (-1)^{\mathbf{i}^T\mathbf{1}}$, i_s denotes the sth component of the vector $\mathbf{i}$, and $\mathbf{q} = [(\bar{g}_1 w_1)^{-1}(\bar{g}_2 w_2)^{-1}\cdots(\bar{g}_K w_K)^{-1}]^T$. From (7) and (2), the second term of (5) reduces to

$$f_s(x)\widetilde{F}_{-s}(xw_s) = -\frac{\sum_{\mathbf{i}\in S} c_{\mathbf{i}} i_s \exp\left(-xw_s\mathbf{q}^T\mathbf{i}\right)}{\bar{g}_s}. \qquad (8)$$

The rate for the sth user will be the rate of the effective point-to-point channel with channel gain g_s^*. Then, for a given channel distribution $\bar{\mathbf{g}}$, scheduling vector $\mathbf{w}$ and average SNR ρ, the achievable rate by the sth, user will be

$$\begin{aligned}R_s(\bar{\mathbf{g}},\mathbf{w},\rho) &= \int_0^\infty \log_2(1+\rho x) f_s^*(x)\,dx\\ &= \int_0^\infty \log_2(1+\rho x) f_s(x)\widetilde{F}_{-s}(xw_s)\,dx.\end{aligned} \qquad (9)$$

Substituting (8) in (9), this can be expressed as follows:

$$R_s(\bar{\mathbf{g}},\mathbf{w},\rho) = -\sum_{\mathbf{i}\in S,\,\mathbf{i}\neq\mathbf{0}} c_{\mathbf{i}}\frac{i_s\exp\left(w_s\mathbf{q}^T\mathbf{i}/\rho\right)}{w_s\bar{g}_s(\mathbf{q}^T\mathbf{i})\ln 2}E_1\left(\frac{w_s\mathbf{q}^T\mathbf{i}}{\rho}\right), \qquad (10)$$

where $E_1(\cdot)$ denotes the exponential-integral function of the first order [11]. Equation (10) explicitly provides the coordinates of the boundary point of the rates region relative to the scheduling vector $\mathbf{w}$, for a given channel distribution $\bar{\mathbf{g}}$ and average SNR ρ. It has some interesting properties.

(i) $R_s(\bar{\mathbf{g}},\mathbf{w},\rho)$ is always a continuous strictly increasing function of ρ, for any $\bar{\mathbf{g}}$ and $\mathbf{w}$. It is demonstrated from (9) that the log function is continuous strictly increasing and that $f_s^*(x)$ is positive and continuous. Therefore, for a given channel distribution $\bar{\mathbf{g}}$, the boundaries of the rates region for different values of ρ never overlap.

(ii) For any $\bar{\mathbf{g}}$ and ρ the rates region is convex.

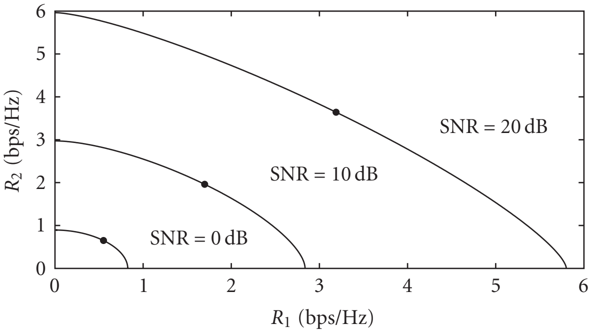


Figure 1: Ergodic rates' regions for a two-users channel when $\bar{g}_1/\bar{g}_2 = 3$ dB.

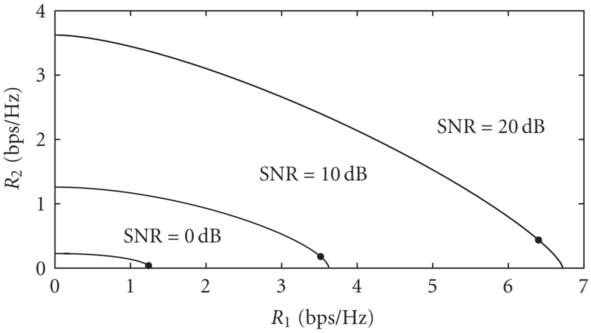


Figure 2: Ergodic rates' regions for a two-users channel when $\bar{g}_1/\bar{g}_2 = 10$ dB.

(iii) $R_s(\bar{\mathbf{g}},\mathbf{w},\rho)$ is continuously differentiable in the convex region $S_{\mathbf{w}}$. The derivatives with respect to $\mathbf{w}$ are

$$\begin{aligned}\frac{\partial R_s}{\partial w_k} &= \sum_{\mathbf{i}\in S,\,\mathbf{i}\neq\mathbf{0}}\frac{c_{\mathbf{i}} i_s i_k w_s}{w_k^2\bar{g}_s\bar{g}_k\ln 2}\left[\frac{e^x E_1(x)(x-1)-1}{x^2}\right],\\ \frac{\partial R_s}{\partial w_s} &= \sum_{\mathbf{i}\in S,\,\mathbf{i}\neq\mathbf{0}}\frac{c_{\mathbf{i}} i_s}{\bar{g}_s\ln 2}\left(\frac{i_s}{\bar{g}_s w_s}-\mathbf{q}^T\mathbf{i}\right)\left[\frac{e^x E_1(x)(x-1)-1}{x^2}\right],\end{aligned} \qquad (11)$$

where $x = w_s\mathbf{q}^T\mathbf{i}/\rho$.

5. Inverse Problems

With the help of expression (10), it is easy to solve some interesting inverse problems.

Problem 1. Given a channel distribution $\bar{\mathbf{g}}^o$ objective rates vector $\mathbf{R}^o = [R_1^o R_2^o\cdots R_K^o]^T$, to find the minimum required average SNR (or transmit power) and the scheduling vector to achieve such rates, this problem can be formulated as follows:

$$\mathbf{R}(\bar{\mathbf{g}}^o,\mathbf{w},\rho) - \mathbf{R}^o = \mathbf{0}, \quad \text{s.t. } \mathbf{w}\in S_{\mathbf{w}},\ \rho > 0, \qquad (12)$$

where

$$\mathbf{R}(\bar{\mathbf{g}}^o,\mathbf{w},\rho) = [R_1(\bar{\mathbf{g}}^o,\mathbf{w},\rho)\cdots R_K(\bar{\mathbf{g}}^o,\mathbf{w},\rho)]. \qquad (13)$$

Considering the constrain $\mathbf{w}^T\mathbf{1} = K$, the expression (12) is a system of K nonlinear equations with K unknowns. Since $\mathbf{R}(\overline{\mathbf{g}}^o, \mathbf{w}, \rho)$ is one-to-one and continuous, there will be a unique solution.

Problem 2. Given a channel distribution $\overline{\mathbf{g}}^o$ and an average SNR ρ^o, to find the maximum achievable rates preserving a given relationship among the users' rates as well as the scheduling vector to achieve such rates, this problem can be formulated as follows:

$$\mathbf{R}(\overline{\mathbf{g}}^o, \mathbf{w}, \rho^o) - a\mathbf{r}^o = \mathbf{0}, \quad \text{s.t. } \mathbf{w} \in S_{\mathbf{w}}, \ a > 0, \qquad (14)$$

where a is a scale factor to be determined, and $\mathbf{r}^0$ is any vector fulfilling the desired relationship among the users' rates. Considering the constrain $\mathbf{w}^T\mathbf{1} = K$, expression (14) is a system of K nonlinear equations with K unknowns including a. Again, it has a unique solution $(\mathbf{w}^*, a^*)$ which provides the required scheduling strategy and the maximum achievable rates $\mathbf{R} = a^*\mathbf{r}^o$.

Other similar problems can be formulated. Due to the properties of $\mathbf{R}(\overline{\mathbf{g}}, \mathbf{w}, \rho)$ (see Section 3), all these problems are well suited to be solved by using conventional gradient-based iterative algorithms. For each problem, the Jacobian matrix can be easily obtained from (11).

6. Numerical Results

Expression (10) gives the achievable users' rates for a given broadcast channel distribution, defined by $\overline{\mathbf{g}}$, for a given weight vector $\mathbf{w}$ and for a given average SNR ρ. By varying $\mathbf{w}$ in (10), we obtain the boundary points of the rates region. As examples, Figures 1 and 2 show the rates regions for a two-users broadcast channel where $\overline{g}_1/\overline{g}_2 = 2$ and $\overline{g}_1/\overline{g}_2 = 10$, respectively. The different curves correspond to different values of average SNR, or equivalently to different transmit powers. The figures also show the points that give the maximum sum-rate, which is achieved using ASS.

Figures 3 and 4 show the individual users rates, as a function of the average SNR, for a 10-users channel using NSS and ASS, respectively. The average channel gains are linearly distributed according to $\overline{g}_k = ak$, $k = 1, \ldots, K$, where $a = 2/(K+1)$ is a constant determined by the channel normalization and $K = 10$.

Figure 3 shows that the NSS algorithm is not totally fair in terms of rates (it is strictly fair in terms of channel access time). The fair scheduling vector can be obtained solving Problem 2 for $\mathbf{r}^o = \mathbf{1}$. Figure 5 shows the optimum weights and the resulting individual rate for different values of average SNR. The optimum scheduling vector changes slowly with the average SNR, especially in the high-SNR regime. We have used a conventional iterative Gauss-Newton method to solve (14). Figure 6 shows the convergence of the users' weights for $\rho = 10$ dB. Starting at $\mathbf{w}_0 = \mathbf{1}$, the algorithm finds the solution after only 4 iterations. To reduce the number of iterations, the starting weights can be heuristically chosen as a function of the average channel gains by assigning higher weights to the worse users' channels. For example, $\mathbf{w}_0 = 1 \cdot / \overline{\mathbf{g}}$ would be a better starting point.

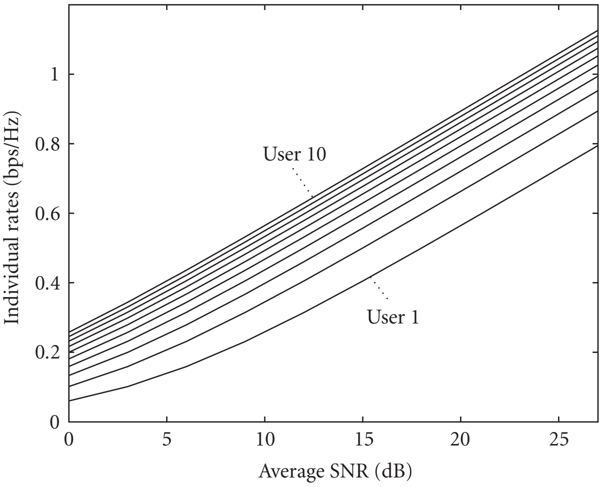


Figure 3: Individual rates for the 10-users channel using NSS.

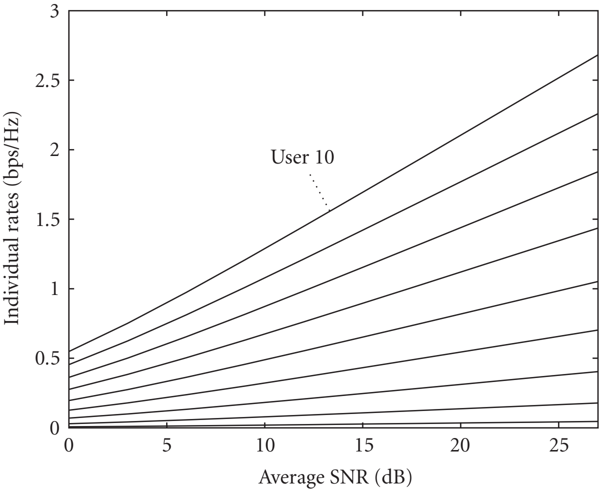


Figure 4: Individual rates for the 10-users channel using ASS.

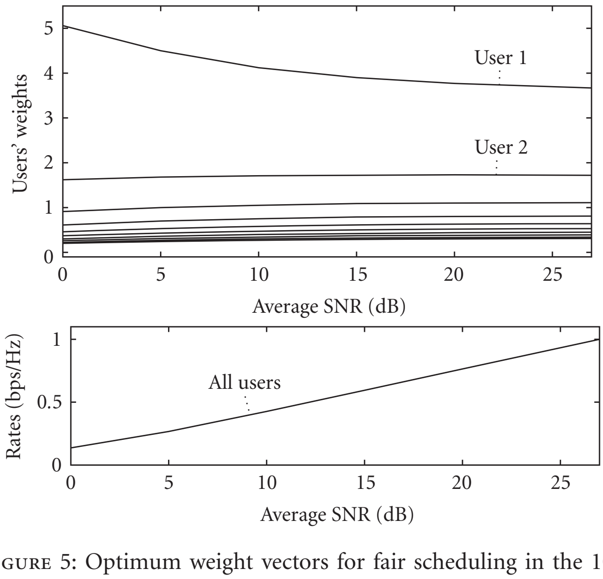


Figure 5: Optimum weight vectors for fair scheduling in the 10-users channel and individual rate.

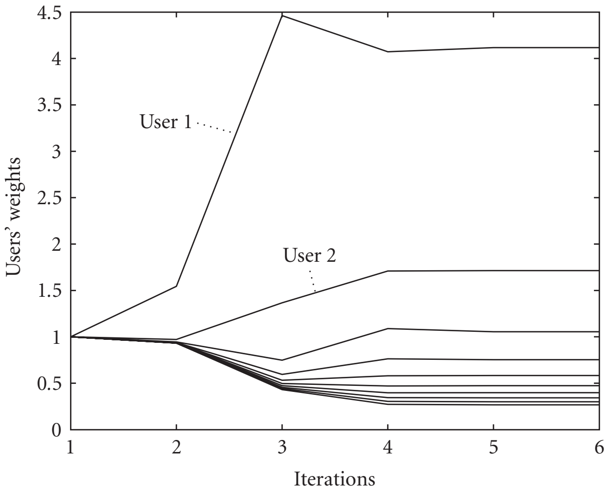


Figure 6: Convergence of the weight vectors for fair scheduling in the 10-users channel using a conventional Gauss-Newton method. The average SNR is $\rho = 10\,\text{dB}$.

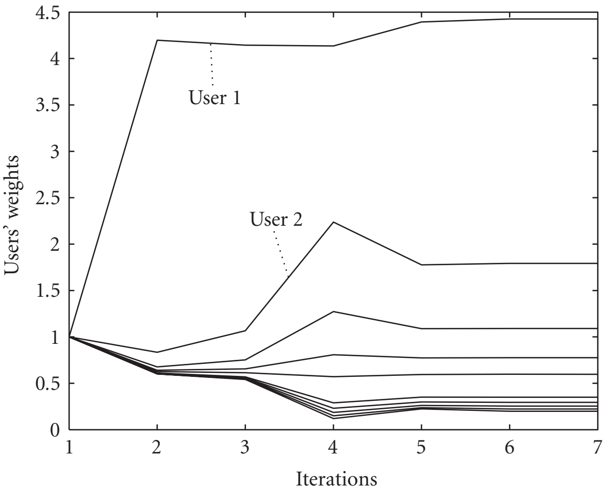


Figure 8: Convergence of the weight vectors in the 10-users channel using a conventional Gauss-Newton method. The average SNR is $\rho = 10\,\text{dB}$.

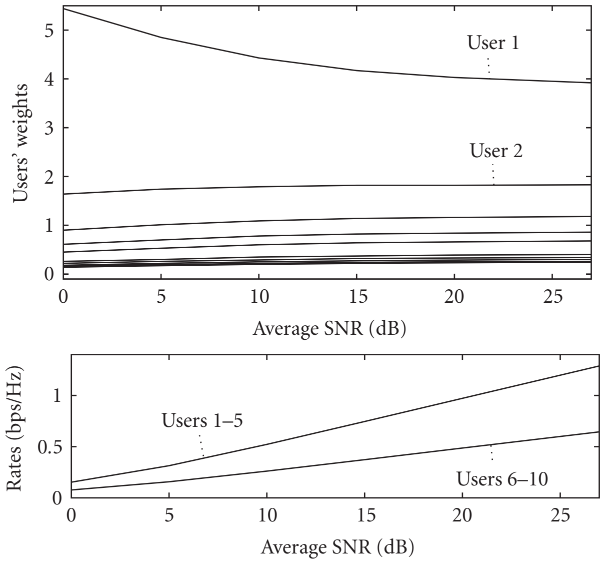


Figure 7: Optimum weights and achievable rates for the two groups of users. The first five curves correspond to the first five users.

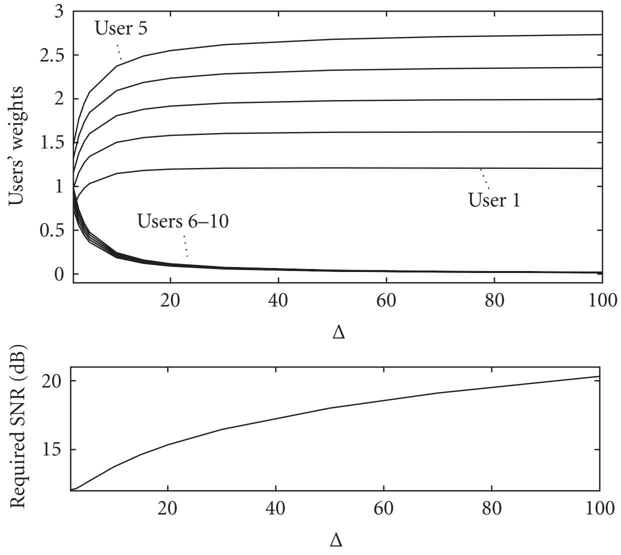


Figure 9: Mininum required average SNR and optimum scheduling weights to achieve the objective rates in different channels determined by the parameters Δ.

Now, assume that we are interested in achieveing different users' rates in the same asymmetric channel. The users are divided in two groups; so the objective rates for the first group double the rates for the second. The first group comprises the users from one to five and the second group from six to ten. To obtain the required scheduling vectors, we solve (14) for $r_k^o = 2$, $k = 1,\ldots,5$ and $r_k^o = 1$, $k = 6,\ldots,10$. Figure 7 shows the achievable individual rates and the scheduling weights to obtain such rates relationship. The convergence to the optimum weights, using a conventional Gauss-Newton algorithm, is depicted in Figure 8 when the average SNR is $\rho = 10\,\text{dB}$. After 5 iterations, the algorithm finds the optimum weights.

As example of Problem 1, we compute the minimum average SNR to achieve the following set of rates $R_k^o = k/K$, $k = 1,\ldots,K$. Again, we consider a 10-users channel but now the average channel gains are given by $\overline{g}_k = a$ for $k = 1,\ldots,4$ and by $\overline{g}_k = a\Delta$, for $k = 5,\ldots,8$, where $a = 2/(\Delta + 1)$. Note that the users are grouped in two sets. In each set the channels are identically distributed. The ratio between the average channel gains of the two sets is determined by the parameter Δ. Figure 9 shows the required average SNR to achieve the objectives rates $R_k^o = k/\text{Kbps/Hz}$ and the optimum scheduling weights, as a function of Δ. Note that as the average channel gains diverge (Δ increases), the required SNR increases.

7. Conclusions

In this paper we studied the performance of the multiuser selection diversity, in broadcast ergodic fading channels, under different SNR-based scheduling schemes. At each fading state, the base station transmits to the user with the highest weighted SNR. By assigning the weights to the users, the base station can arrange the users according to a prescribed quality of service or degree of fairness. Each set of weights corresponds to a specific scheduling policy. We have derived a closed-form expression for the achievable users' rates as a function of the scheduling weights, the transmit power, and the channel fading statistics. With the help of this expressions, we show how to obtain the optimum (in terms of transmit power) scheduling policy to achieve a prescribed set of users' rates. Also, given a transmit power, we obtain the scheduling policy that maximizes the overall throughput preserving a given relationship among the users' rates.

Acknowledgment

This work has been supported by the Spanish Government under projects TEC2007-68020-C04-02/TCM and CSD2008-00010(COMONSENS).

References

[1] A. Goldsmith, *Wireless Communications*, Cambridge University Press, Cambridge, UK, 2005.

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

[3] "Ist-856, high rate packet data air interface specification, document c.s0024 v3.0," December 2001, http://www.3gpp2.org.

[4] WiMAX forum, "Mobile WiMAX—part II: a comparative analysis," http:/www.wimaxforum.org.

[5] A. Farrokh and V. Krishnamurthy, "Opportunistic scheduling for streaming users in high-speed downlink packet access (HSDPA)," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '04)*, vol. 6, pp. 4043–4047, Dallas, Tex, USA, November-December 2004.

[6] L. Yang and M.-S. Alouini, "Performance analysis of multiuser selection diversity," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 6, pp. 1848–1861, 2006.

[7] F. Berggren and R. Jantti, "Asymptotically fair transmission scheduling over fading channels," *IEEE Transactions on Wireless Communications*, vol. 3, no. 1, pp. 326–336, 2004.

[8] D. N. C. Tse and P. Viswanath, *Fundamentals of Wireless Communications*, Cambridge University Press, Cambridge, UK, 2005.

[9] J. M. Holtzman, "Asymptotic analysis of proportional fair algorithm," in *Proceedings of the 12th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '01)*, vol. 2, pp. F33–F37, San Diego, Calif, USA, September-October 2001.

[10] J. Pérez, J. Ibáñez, and I. Santamaría, "Exact closed-form expressions for the sum capacity and individual users' rates in broadcast ergodic Rayleigh fading channels," in *Proceedings of the 8th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC '07)*, pp. 1–5, Helsinki, Finland, June 2007.

[11] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, NY, USA, 1970.