Using Traffic Asymmetry to Enhance TCP Performance

Toktam Mahmoodi, Vasilis Friderikos, Hamid Aghvami

Centre for Telecommunications Research, King's College London
London WC2R 2LS, UK

Abstract

A wealth of recent work has gone into optimizing the performance of Transmission Control Protocol (TCP) on the downlink channel of wireless networks such as for example, honing its congestion awareness mechanism so that it is minimally affected by random wireless losses, and optimizing achieved fairness of the end-to-end TCP rates. Other work has gone into balancing the allocation of a shared resource between the downlink and uplink in order to optimize TCP performance. We build on such previous research by proposing a cross-layer algorithm for resource allocation in OFDMA systems aiming not only to achieve optimal throughput for competing TCP flows but also to allocate resources appropriately between the downlink and uplink. This is important due to the increasing number of Internet applications where the mobile terminal is the TCP sender (social networking, peer-to-peer, etc.). Therefore, our scheme makes use of the asymmetry in the traffic and by defining the boundary between downlink and uplink capacity dynamically, enhance the TCP performance. Through numerical investigations we show the performance of the proposed scheme in terms of achieved fairness to the receivers and efficient allocation of downlink to uplink ratios based on the TCP traffic.

Keywords: Transmission Control Protocol, Wireless Resource Management, Asymmetric Links.

1. Introduction

The Transmission Control Protocol (TCP) [1] is the prominently used transport layer protocol to achieve reliable end-to-end data transfer in IP based networks. TCP connections as being inherently bi-directional, require ACKnowledgements (ACKs) from the receiver in order to achieve reliable communication. On one hand, the common assumption in data centric networks is that the downlink carries heavier data load, thus less bandwidth is specified for the uplink path. The examples of such mode of operation are available for both wired and wireless networks. In Long Term Evolution (LTE) [2], the allocated bandwidth to uplink is the half of the allocated bandwidth to downlink, while in the case of ADSL up to eight times higher than the uplink bandwidth could be allocated to the downlink. On the other hand, Mobile data usage for 2010 reveals the increase in mobile data traffic [1]. Hence, in many of the applications such as Peer to Peer, and Instant Messaging the sender can be the mobile terminal, which is different from traditional applications such as Web browsing where the sender is a web-server, and therefore uplink capacity plays an important role.

The effect of link asymmetry on the performance of TCP is widely studied in wired networks. Limited available bandwidth and congestion on the revers path breaks down the principle of ACK clocking, and may cause an increase in the Round Trip Time (RTT), which can degrade the TCP throughput on the forward path [3]. Several research works explore these issues and a range of solutions have been proposed. Some of these proposals require explicit support from routers or middle boxes, whereas others are end-to-end schemes. For example, ACK congestion control [4] attempts to reduce the sending rate of ACK traffic, with the assumption that a reduction in ACKs rate may help to reduce the congestion itself. In this research work, we explore the above discussed issue in wireless networks. Our attempt is to bring the requirements of TCP on the reverse path, into the actual radio resource allocation mechanisms. Therefore, the limited capacity on the uplink or congested uplink can be avoided. Unlike the existing end-to-end solutions [4], our proposed solution applies a small modification to the TCP header and remains its state diagram unchanged.

To this end, the contributions of this paper are in two categories. First, with jointly allocation of the downlink and uplink, the optimal proportion between downlink and uplink resources can be achieved. The optimal proportion is selected so as the delivery guarantee of TCP ACK packets is accomplished and also downlink throughput is maximized. It is worthwhile to note that over-allocation or under-allocation to the uplink bandwidth could decrease the performance of the end-to-end connection in the downlink either directly or indirectly. Second, we define the downlink resource allocation problem not only with respect to the channel quality but also the TCP constraints. Thereby, fairness among end-to-end TCP flows is considered in addition to maximizing sum rate.

In this study, the access method is based on the Orthogonal Frequency Division Multiple Access (OFDMA), which is the selected access technique in emerging and future wireless networks [2], [5]. The problem of resource allocation in OFDMA wireless networks is to assign subcarriers and distribute power in order to improve the performance of the system either by maximizing the overall data rate or by considering fairness issues. Although the problem of subcarrier/power allocation in OFDMA-based wireless networks has been widely studied over the past few years, the aspects of fairness as pertain to the end-to-end communications have not been sufficiently addressed. Hence, downlink resource allocation problem in this work is defined so that fairness among end-to-end TCP flows

[1] Allot MobileTrends: Global Mobile Broadband Traffic Report 2010

is satisfied. This is particularly important when the end-to-end TCP flows are based on different TCP versions or their paths have diverse RTTs [6].

Our major contributions in this paper are as follows.

1. To avoid the performance degradation of the end-to-end TCP flows due to the scarce resources in the uplink, a joint uplink-downlink resource allocation problem is proposed. This resource allocation scheme, which is studied for OFDMA-based wireless networks, guarantees to allocate sufficient resources to the uplink of each individual TCP flow with regard to the allocated resources in the downlink. Therefore, the performance of TCP flow is not affected by the asymmetry in traffic but this asymmetry is utilized to enhance TCP performance.
2. We expand the objective of the downlink resource allocation problem in OFDMA to include the theoretical upper bound of TCP throughput; thus fairness among downlink end-to-end TCP flows can be achieved.
3. A new bit in the TCP header is introduced so that TCP congestion control state can be reported. This information is used in computing the achievable throughput by the end-to-end flow. In the slow start phase, TCP throughput simply depends on the Congestion Window (CW) and RTT of the flow, which are both available values in the TCP header [7]. In the congestion avoidance phase, the model in [8] illustrates the steady state throughput given the RTT and PER that each TCP flow may experience in its end-to-end path.
4. The performance of our proposed schemes is thoroughly investigated under various network conditions.

To the best of our knowledge, this is the first research work which propose to adapt the OFDMA resource allocation problem so that TCP requirements in the uplink have been taken into account. In other words, the effect of link asymmetry on the performance of TCP is addressed.

The remainder of this paper is organized as follows. In the next section, an overview of the OFDMA resource allocation techniques, and the state of the art in TCP-aware resource allocation schemes are discussed. Section 3 details the system model and baseline assumptions used in the paper. In Section 4, the proposed joint resource allocation problem is introduced, together with the two variants of the TCP-aware allocation schemes. Section 5 describe the associated techniques to solve corresponding optimization problems. Section 6 presents the numerical investigations of the proposed scheme under various conditions. Finally, this paper concludes in Section 7.

2. Background Study

Much of the OFDMA resource allocation techniques in the literature have concentrated on the allocation of resources in the downlink, e.g., by maximizing the overall data rate subject to power or Bit Error Rate (BER) constraints [9]. Moreover, depending on the duplexing method, the available resources are divided either in time or in frequency between downlink and uplink channels. The proportion of uplink and downlink capacity can be defined as one of the system parameters. For example, in the LTE assumptions [2], the uplink capacity is equal to the half of the downlink capacity.

On the other hand, resource allocation mechanisms can consider the joint downlink-uplink allocation problem. Such allocation techniques can provide a dynamic border between downlink capacity and uplink capacity, in order to allocate the total amount of resources more efficiently. As mentioned earlier, this proportion can be defined dynamically based on various system constraints to guarantee the requested Quality of Service (QoS). In this paper, we constrain the resource allocation problem with the requirements of the bi-directional connection of the transport layer (e.g. TCP) so as we guarantee the delivery of TCP ACK packets for the allocated bandwidth to the downlink of each flow, while the attempt is to maximize this bandwidth.

The downlink resource allocation problem per se, could aim to maximize sum rate [9] subject to power consumption, minimize the overall power consumption while satisfying the minimum rate requirements [10], or consider certain degree of QoS. Furthermore, the downlink resource allocation scheme may consider fairness among users, either by prioritization using the weighted sum rate method [11], or by introducing proportional rate constraints [12]. Another possible approach is presented in [13], in which fairness is considered by maximizing the lowest achieved data rate among the user set. The research presented in [14] addresses the issue of how to provide proportional fairness in OFDMA resource allocations based on the Nash bargaining solution. Although these research works investigate the issues of fairness and QoS with respect to the allocated data rate over the wireless link, aspects as pertain to the end-to-end data transmission perspective have not been sufficiently addressed.

TCP as the prominently used transport layer protocol provides reliable, connection oriented data transfer in IP based networks. However, TCP performance is problematic in wireless networks, which are characterized by random losses and intermittent connectivity. TCP treats wireless random losses as congestion indication, and reduce its congestion window, lowering in that respect unnecessarily its rate [15]. Such factors, in addition to differences in the way different versions of TCP react to random wireless losses mean that –particularly over wireless links– there can be considerable unfairness among TCP flows [6]. Since the majority of applications in the Internet use TCP, TCP's fairness has been well studied in the literature [16]. In this regard, we explore the fairness among the TCP flows as can be accomplished by the wireless resource allocation schemes.

A thorough overview of cross-layer design for resource allocation algorithms in the third generation wireless networks is given by [17], where aspects related to TCP performance over CDMA are also addressed. TCP-aware resource allocation algorithms over a CDMA network are studied in [18], the objective being to maximize throughput. The proposed algorithm in that paper uses information on the TCP state (slow start or congestion avoidance) to allocate the data rate more appropriately at the wireless link. A joint congestion control and power allocation in a CDMA based wireless network is proposed in

[19], in which a generalized network utility maximization frame work is also presented. Furthermore, [20] introduces enhancement in fairness among TCP connections over CDMA network by allowing longer RTT connections to have higher signal-to-interference ratios.

In the context of IEEE 802.16, reference [21] proposes a TCP-aware allocation algorithm which estimates the bandwidth demand based on the long-term data rate, and allocates resources accordingly. Unlike available solutions in the literature, we use the closed form expression of TCP throughput [8] as a mean of TCP-awareness in allocations. Moreover, in contrast with existing TCP-aware resource allocation techniques, we focus on OFDMA-based systems and take both uplink and downlink algorithms into account.

3. System Model

The core of TCP congestion control algorithm are the slow start and the congestion avoidance phases. In the slow start, the achievable throughput by the TCP flow simply depends on the actual value of CW and RTT. On the other hand, throughput in the congestion avoidance phase can be expressed by the TCP steady state throughput [8]. We introduce a new flag in the TCP header, SS/CA flag, that illustrates the actual state of TCP, i.e. SS/CA equal to zero or one represents being in slow start or congestion avoidance successively. [2]

As TCP transits into congestion avoidance phase, the steady state expression of TCP throughput can represent its achievable throughput. Given e_i the probability of a packet being in error (PER) in flow i, the below closed-form function expresses TCP steady state throughput based on the model in [8] that were later revised in [22],

$$B(ss)_i = MSS \cdot \frac{\frac{1-e_i}{e_i} + E\left[W_i\right]}{\overline{RTT}_i\left(\frac{1}{2} \cdot E\left[W_i\right] + 2\right)} \quad bytes/s, \tag{1}$$

where

$$E\left[W_i\right] = -\frac{3b-2}{3b} + \sqrt{\frac{8\left(1-e_i\right)}{3be_i} + \left(\frac{3b-2}{9b^2}\right)^2}. \tag{2}$$

In Equation (1), MSS is the TCP Maximum Segment Size in bytes, b is the number of packets that are acknowledged by receiving an ACK, and e is the probability of a TCP packet in error which can be driven from the BER of wireless link, assuming that the bottleneck is the wireless link. The $\overline{RTT}$ is the average value of RTT, thus the instantaneous variation in the RTT caused by each single loss of acknowledgement does not affect the throughput. The throughput expression here is based on the above TCP version. We should note that even though the above throughput expression is based on the well-used

[2] The SS/CA flag can be implemented in the TCP header, option section with Kind = 30, and Length = 1 [7].

TCP Reno, Equation (1) can be replaced by the throughput expression of other TCP versions.

Therefore, the achievable throughput of a TCP flow i, B_i, can be expressed as,

$$B_i = \begin{cases} MSS \cdot \frac{CW_i}{RTT_i}, & SS/CA = 0, \\ B(ss)_i, & SS/CA = 1. \end{cases} \tag{3}$$

To calculate the TCP throughput in the wireless base station, where the resource allocation takes place, the value of RTT is required. Various methods are presented in the literature to estimate RTT either actively or passively at any router in the middle of the end-to-end path. The passive measurement can be done based on the three-way handshake message [23], or by associating the data segment with the acknowledgement that triggers the packet [24]. TCP timing information can also be included in the Timestamp option of the TCP segment [7]. Experiments show that 90% of the passive measurements are within 10% of the precise RTT value [23]. These methods are not computationally complex and can be easily implemented at the link-layer of the base station.

To this end, we assume n active TCP flows all of which operate in either slow start or congestion avoidance phase. A single cell OFDMA network is assumed with m available subcarriers. Let for flow i the rate on subcarrier j to be r_{ij}. Each user is associated with a single TCP flow, therefore, the achievable rate for user i can be written as follows,

$$R_i = \sum_{j=1}^{m} a_{ij} r_{ij}, \tag{4}$$

where,

$$a_{ij} = \begin{cases} 1 & \text{if subcarrier } j \text{ is assigned to user } i, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The channel gain of user i at subcarrier j is denoted by G_{ij}. With the thermal noise power, σ^2, the ith user's received signal to noise ratio (SNR) on subcarrier j is denoted as,

$$\gamma_{ij} = \frac{p_{ij} G_{ij}}{\sigma^2}. \tag{6}$$

where p_{ij} is the allocated power to flow i on subcarrier j.

Adaptive modulation provides the desired rate in the allocated subcarrier for each individual user. Given $c_1 \approx 0.2$, $c_2 \approx 1.5$, BER is expressed based on the adaptive M-array quadratic modulation (M-QAM) [25].

$$BER_{ij} \approx c_1 \, e^{-c_2 \frac{\gamma_{ij}}{2^{r_{ij}}-1}}. \tag{7}$$

Similar to [14] we assume a fixed and the same BER for all users in all subcarriers i.e. $BER_{ij} = BER \;\; \forall i, j$. Given $c_3 = -ln(BER/c_1)/c_2$, and solving for r_{ij}, the achievable rate for user i on the jth subcarrier can be described as follows,

$$r_{ij} = w_j \log_2 \left(1 + \frac{p_{ij} G_{ij}}{\sigma^2 c_3}\right) \;\; bits/s. \tag{8}$$

In Equation (8), w_j is the bandwidth of subcarrier j which is assumed to be equal for all subcarriers and will be denoted hereafter by w. The wireless channel suffers from slow-fading effect such that the channel is constant within each OFDM frame. The slowly time varying assumption is crucial since it is also assumed that perfect estimation of the subchannels is available for each user. Moreover, mobile users and the base station are synchronized, thus there is no inter-carrier interference.

4. TCP-Aware Resource Allocation Scheme

Despite the fact that TCP has been initially designed for elastic applications it is currently commonly used in various popular streaming applications. It is worthwhile noting that Real Media and Windows Media, the two dominant streaming media applications, both are based on TCP streaming. In that respect, in wireless networks where resources are scarce TCP traffic for such applications should not be treated as best effort but some provision on the data rate have to be considered. In the proposed approach this provision is based on the theoretical average throughput that can be achieved by TCP, based on the specific path characteristics (i.e., RTT, packet error rate).

As mentioned in Section 2, according to the state of the art in OFDMA resource allocation schemes, fairness aspects as pertain to the end-to-end data transmission perspective have not been sufficiently addressed. In this respect, aim of the proposed downlink resource allocation problem is to determine the users'transmission functions $[A]_{ij} = a_{ij}$ and power matrix $[P]_{ij} = p_{ij}$ in order to maximize the overall rate with regard to the power constraints while maintaining fairness among the active end-to-end TCP flows [26][27]. On the other hand, the uplink resource allocation aims to provide appropriate throughput for the uplink. The example detailed here show how either over allocation or under allocation of resources to the uplink can reduce the downlink throughput.

4.1. Motivation

In data centric networks, it is assumed that downlink carries heavier data load, thus less bandwidth is specified for the uplink path. In spite of this fact, with the increasing number of applications uploading data i.e. Emails with large attachments, the uplink resources can become more scarce. TCP connection as a bi-directional connection, requires ACK from the receiver for the transmitted data packets to achieve the reliability. Therefore, the effect of link asymmetry on the performance of TCP, which has been widely studied in the wired networks [28], can also be crucial in wireless networks. To depict a clearer picture, an example is detailed here, in which we assume that a user downloads data over a link with 20 Mbps capacity, while the uplink capacity is limited to the 100

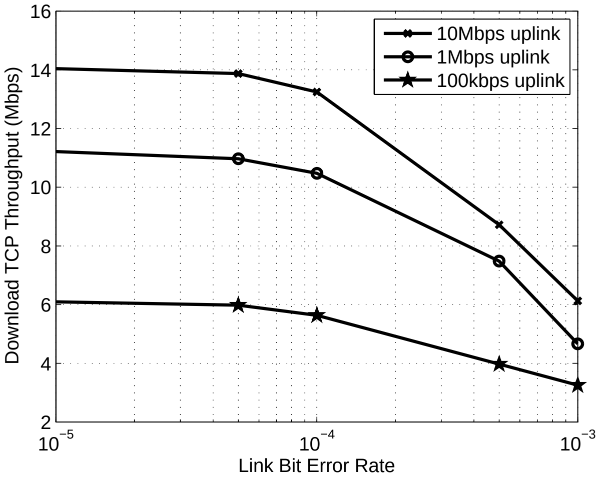

Figure 1: Effect of Link Asymmetry on TCP Throughput: Download Throughput Vs. Link BER, Downlink Capacity is 20 Mbps

kbps. Given the lengths of data packets 1500 B and the lengths of ACK packets 40 B, TCP can only send ACKs for every 5 packets; otherwise the uplink path will be saturated. Therefore, the principle of ACK clocking can break down, thus the sender clocks out new data at a slower rate. In other words, the sender CW grows slower and the TCP flow utilizes the allocated downlink bandwidth inefficiently.

In the above example, if TCP acknowledges every single packet, it can achieve not more than 400 kbps on the downlink. Although by increasing the capacity of uplink, downlink capacity is decreased, download throughput shows an increase due to the successful/on-time delivery of the acknowledgements (can be seen Figure 1). Clearly after some point that depends on the actual uplink traffic, increasing the uplink capacity can results in reducing the download throughput due to the decrease in its capacity. Finding this trade-off depending on the actual traffic is an interesting problem, which is addressed by this paper.

In wireless networks, this phenomena can occur in for example the following scenario where it is assumed that there are 20 mobile users and the capacity of downlink and uplink are 20 Mbps and 10 Mbps consequently. If all the mobile users download data and attain equal share of the bandwidth, each mobile user download 1 Mbps while the total of 500 kbps is the uplink traffic—the same packet size and ACK size to the previous example are assumed. Assuming that 9 of these users upload data with the same rate as download (1 Mbps), capacity of the uplink that can be used for the ACK traffic of the downlink streams is decreased to 1 Mbps. Increasing the number of uploading mobile users to 10, the uplink is congested in a way that the ACK traffic of the downlink streams can not flow. On the other hand, if three more users upload data but with the lower rate than their download data, e.g. the third of downlink data rate, 300 kbps, the uplink capacity for the ACK traffic is decreased to 100 kbps. Further assuming that three more users are uploading data with 30 kbps, the uplink capacity is decreased to 10 kbps. Figure 1 shows the downlink throughput versus the link BER in the described scenario. The presented results in Figure 1 are simulated using OPNET modeler in which the RTT is 60 ms.

4.2. Formal Problem Definition

Considering the above example, in this section we define a joint allocation of downlink and uplink resources, to provide the maximum throughput on the downlink, and also to guarantee the delivery of the downlink packets with the appropriate data rate on the uplink. The objective of downlink allocation scheme is presented in two variants. The first proposed optimization problem aims to maximize sum rate, constrained to the proportional data rate on each individual TCP flow. The proportional rate is weighted with the theoretical TCP throughput, which is the throughput that a TCP flow can achieve, dependent on the end-to-end RTT and the packet error probability of the corresponding flow – this throughput can be considered as the actual capacity of the end-to-end path. In the second proposed TCP-aware resource allocation problem, we investigate the difference between the allocated wireless link rate and the theoretical achievable TCP throughput. This optimization problem aims to maximize the sum rate while minimizing the gap between the allocated data rate to each flow and the theoretical throughput that can be achieved by that TCP flow. These two formulations are detailed in the followings.

4.2.1. Proportional TCP Throughput Constrained (P1)

The formulated optimization problem (P1) aims to maximize downlink sum rate, while TCP fairness is assured by imposing a set of nonlinear constraints into the problem. We propose to constrain the proportional downlink rate among users with respect to the TCP theoretical throughput (the solely downink resource allocation problem is also discussed in [29]). In the uplink, assuming TCP receiver acknowledges every single packet, the minimum required data rate would be a proportion of the downlink data rate for each specific TCP flow depending on the size of the ACK packet, which can be increased for example using the SACK option, $R_{u_i} \geq \rho R_{d_i}$, i.e. $0 < \rho < 1$. Therefore, our optimization problem is also constrained by the minimum achievable uplink rate for each flow. The above described optimization problem can be formulated as follows,

$$\text{(P1)} : \text{Maximize} \sum_{i=1}^{n} \sum_{j=1}^{m} c_j a_{ij} w \log_2 \left(1 + \frac{p_{ij} G_{ij}}{\sigma^2 c_3} \right),$$

$$\text{subject to: } \sum_{i=1}^{n} a_{ij} \leq 1, \qquad \forall j \in \{1, ..., m\} \tag{9}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c_j a_{ij} p_{ij} \leq P_T, \tag{10}$$

$$\sum_{j=1}^{m} (1 - c_j)\, a_{ij} p_{ij} \leq P_t, \qquad \forall i \in \{1, ..., n\} \tag{11}$$

$$\frac{R_{d_i}}{B_i} = \frac{R_{d_1}}{B_1}, \qquad \forall i \in \{2, ..., n\} \tag{12}$$

$$R_{u_i} \geq \rho R d_i, \qquad \forall i \in \{1, ..., n\} \tag{13}$$

$$\sum_{j=1}^{m} c_j \leq m_d, \qquad m_d \in \{1, ..., m\}, \tag{14}$$

$$p_{ij} \geq 0, \qquad \forall i \in \{1, ..., n\}, j \in \{1, ..., m\} \tag{15}$$

$$a_{ij} \in \{0, 1\}, \qquad \forall i \in \{1, ..., n\}, j \in \{1, ..., m\} \tag{16}$$

$$c_j \in \{0, 1\}. \qquad \forall j \in \{1, ..., m\} \tag{17}$$

In this problem, despite the classic approaches in solely maximizing throughput, subcarriers (a_{ij}) and transmission power over each subcarrier (p_{ij}) are allocated such that certain performance metrics of TCP are guaranteed. These performance metrics are provided via the constraints that are detailed below. Constraints (9) ensure that every subcarrier is assigned to only one user. We assume c_j represents the allocation of subcarrier j to downlink (c_j=1) or uplink (c_j=0). Thereby, constraints (10) and (11) restrict the total available power at the base station, P_T, and at each mobile user, P_t. Moreover, constraint (12) provide fairness among TCP flows with maintaining proportional rate with respect to the TCP throughput for each user. Constraint (13) provide the required data rate for uplink, in order to guarantee delivery of the downlink allocated resources, and finally constraint (14) bounds the number of downlink subcarriers to m_d. Note that the optimal value of m_d can be found solving the problem (P1) iteratively for different values of m_d.

4.2.2. Rate Difference from TCP Throughput Constrained (P2)

In this problem, similar to (P1), we attempt to maximize downlink sum rate, but the instantaneous rate allocation is constrained with the TCP theoretical throughput (the solely downlink allocation problem is also discussed in [26]). Given D_i, the difference between allocated data rate to the ith user and the theoretical TCP throughput of flow i, it can be represented as follows,

$$D_i = \mid \alpha B_i - R_{d_i} \mid . \tag{18}$$

In Equation (18), α represents the overhead of the TCP/IP header. The resource allocation problem in this case attempts to minimize D_i while maximizing the sum rate. Therefore, the novel resource allocation problem (P2) can be defined as a multi objective optimization problem. There are various approaches to formulate such a multi objective problem; we use a well-studied approach that combines the multiple objectives into a single objective function whose solution is Pareto optimal.

$$\text{(P2)}: \text{Maximize} \sum_{i=1}^{n}\sum_{j=1}^{m} c_j a_{ij} w \log_2\left(1+\frac{p_{ij}G_{ij}}{\sigma^2 c_3}\right) - \mu \sum_{i=1}^{n} D_i \,.$$

$$\text{subject to: } \sum_{i=1}^{n} a_{ij} \leq 1, \qquad \forall j \in \{1, ..., m\} \tag{19}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c_j a_{ij} p_{ij} \leq P_T, \tag{20}$$

$$\sum_{j=1}^{m} (1-c_j)\, a_{ij} p_{ij} \leq P_t, \qquad \forall i \in \{1, ..., n\} \tag{21}$$

$$R_{u_i} \geq \rho R d_i, \qquad \forall i \in \{1, ..., n\} \tag{22}$$

$$\sum_{j=1}^{m} c_j \leq m_d, \qquad m_d \in \{1, ..., m\}, \tag{23}$$

$$p_{ij} \geq 0, \qquad \forall i \in \{1, ..., n\}, j \in \{1, ..., m\} \tag{24}$$

$$a_{ij} \in \{0,1\}, \qquad \forall i \in \{1, ..., n\}, j \in \{1, ..., m\} \tag{25}$$

$$c_j \in \{0,1\}. \qquad \forall j \in \{1, ..., m\} \tag{26}$$

Constraints (19)-(26) are the same as (9)-(11) and (13)-(17). As mentioned above, problem (P2) has a Pareto optimal solution; thus the solution is not unique and it depends on the value of μ that balances the two objectives. In the above problem, increasing the value of μ shift the allocation balance towards TCP throughput, while decreasing the value of μ shifts the balance towards a data rate maximization problem.

5. Subcarrier Allocation and Power Distribution

5.1. Optimal Solutions

Clearly, subcarrier and power should be assigned jointly to achieve the optimal solution. This joint allocation represents a mixed integer non-linear mathematical programming problem which pose a high computational complexity. Although problems (P1) and (P2) can be solved using well-known optimization techniques, it is prohibitive for the base station to solve these problems in real time due to their complexity.

For real-time implementation and to allow larger instances of the problem to be solved we present a greedy allocation which provides suboptimal but feasible solutions. To this end, we use the method presented in the literature [12] to decouple the problem. The addressed optimization problem can be decoupled to two separate problems; first the allocation of the subcarriers, and second the distribution of the available power into the allocated subcarriers.

5.2. Suboptimal Solutions

We use the approach similar to [12] to decouple the optimization problem. In the subcarrier allocation it is assumed that power is equally distributed in all the subcarriers, therefore the solution is suboptimal. Afterwards, to a certain subcarrier allocation, an optimization problem can be reformulated over the continues variable p_{ij}. Thus, using the water filling approach, power will be distributed optimally.

The principle of the downlink algorithm is to allocate the subcarrier with the highest channel gain available for each user. In addition to that, in the first round of allocation, we let the user with the highest theoretical TCP throughput (largest value of B_i) to first select a subcarrier. Thereafter, in each round of the allocation, in solving problem (P1) users with the lowest proportional rate have priority to select the best available subcarrier. This step performs differently solving problem (P2), i.e. user with the smallest objective function selects the next subcarrier.

Subcarrier allocation to the uplink in both problems is in order to satisfy the uplink minimum rate requirements. We select the initial value of m_d in order to keep the proportion of ρ between number of downlink and uplink subcarriers – $m_d = m \cdot \frac{1}{1+\rho}$. Afterwards, in few iterations, the largest value of m_d which satisfies constraint (12) will be found; clearly this value maximizes the objective function. The above procedure for problem (P1) is detailed in Algorithm 1, in which Ω_i is the set of allocated subcarriers to the user i in the downlink and Ψ_i is the set of allocated subcarriers to this user in the uplink. Algorithm 2 details the subcarrier allocation solving problem (P2) in which steps (a), (b), (d), and (e) are similar to Algorithm 1, and only step (c) is restated.

The problem of power allocation with pre-defined subcarrier allocation, is based on the reformulation of (P1) into a maximization problem over continues variable p_{ij}.

$$(\mathrm{P1}') : \text{ Maximize } \sum_{i=1}^{n} \sum_{j\in\Omega_i} w\log_2\left(1 + \frac{p_{ij}G_{ij}}{\sigma^2 c_3}\right),$$

$$\text{subject to: } \sum_{i=1}^{n} \sum_{j\in\Omega_i} p_{ij} \le P_T, \tag{27}$$

$$\sum_{j\in\Psi_i} p_{ij} \le P_t, \qquad \forall i \in \{1, ..., n\} \tag{28}$$

$$\frac{R_{d_i}}{B_i} = \frac{R_{d_1}}{B_1}, \qquad \forall i \in \{2, ..., n\} \tag{29}$$

$$R_{u_i} \ge \rho R_{d_i}, \qquad \forall i \in \{1, ..., n\} \tag{30}$$

$$p_{ij} \ge 0. \qquad \forall i \in \{1, ..., n\}, j \in \{1, ..., m\} \tag{31}$$

Algorithm 1 Subcarrier Allocation Algorithm for the optimization problem (P1)

```
a) Initialization
    1. $m_{d_1} = \frac{1}{1+\rho} \cdot m$.
    2. Set $R_{d_i}$=0 and $\Omega_i = \phi$ for i=1 to $n$ and $C_d = \{1, ...m_d\}$.
    3. Set $R_{u_i}$=0 and $\Psi_i = \phi$ for i=1 to $n$ and $C_u = \{m_d + 1, ...m\}$.
    4. Sort the users' index in the descending order of $B_i$.
b) for i=1 to $n$
    1. Find the subcarrier $k$ satisfying $|G_{ik} |> |G_{ij} |$ for all $j \in C_d$.
    2. Let $\Omega_i = \Omega_i \cup \{k\}$ and $C_d = C_d - \{k\}$.
    3. Update $R_{d_i}$
c) while $C_d \neq \phi$
    1. Find user $l$ satisfying $R_{d_l}/B_l < R_{d_i}/B_i$ for all $i \in \{1, ..., n\}$.
    2. For user $l$, find the subcarrier $k$ satisfying $|G_{ik}| > |G_{ij}|$ for all $j \in C_d$.
    3. Let $\Omega_l = \Omega_l \cup \{k\}$ and $C_d = C_d - \{k\}$.
    4. Update $R_{d_i}$
d) for i=1 to $n$
    1. Find the subcarrier $k$ satisfying $|G_{ik} |> |G_{ij} |$ for all $j \in C_u$.
    2. Let $\Psi_i = \Psi_i \cup \{k\}$ and $C_u = C_u - \{k\}$.
    3. Update $R_{u_i}$
e) while $C_u \neq \phi$
    1. Find user $l$ satisfying $(R_{u_l} - \rho R_{d_l}) < (R_{u_i} - \rho R_{d_i})$ for all $i \in \{1, ..., n\}$.
    2. For user $l$, find the subcarrier $k$ satisfying $|G_{ik}| > |G_{ij}|$ for all $j \in C_u$.
    3. Let $\Psi_l = \Psi_l \cup \{k\}$ and $C_u = C_u - \{k\}$.
    4. Update $R_{u_i}$
f) If $\sum_{i=1}^{n} R_{u_i} \geq \rho \cdot \sum_{i=1}^{n} R_{d_i}$.
    1. Find the largest $m_d \in \{m_{d_1}, ..., m\}$ such that constraint (13) are satisfied.
    2. Else, find the largest $m_d \in \{1, ..., m_{d_1}\}$ such that constraints (13) are satisfied.
```

Algorithm 2 Subcarrier Allocation Algorithm for the optimization problem (P2)

```
c) while $C_d \neq \phi$
    1. Find user $l$ satisfying $R_{d_l} - \mu D_l < R_{d_i} - \mu D_i$ for all $i \in \{1, ..., n\}$.
    2. For user $l$, find the subcarrier $k$ satisfying $|G_{ik}| > |G_{ij}|$ for all $j \in C_d$.
    3. Let $\Omega_l = \Omega_l \cup \{k\}$ and $C_d = C_d - \{k\}$.
    4. Update $R_{d_i}$
```

In the Algorithms 1 and 2, Ω_{i_1} and Ω_{i_2} are mutually exclusive, if $i_1 \neq i_2$; the same assumption is also valid for Ψ_i. Problem (P2) also can be rewritten as (P2$'$) over the continues variable p_{ij}.

$$(\mathrm{P2}') : \text{ Maximize } \sum_{i=1}^{n} \sum_{j \in \Omega_i} w \log_2 \left(1 + \frac{p_{ij} G_{ij}}{\sigma^2 c_3} \right) - \mu \sum_{i=1}^{n} D_i,$$

$$\text{subject to: } \sum_{i=1}^{n} \sum_{j \in \Omega_i} p_{ij} \leq P_T, \tag{32}$$

$$\sum_{j \in \Psi_i} p_{ij} \leq P_t, \qquad \forall i \in \{1, ..., n\} \tag{33}$$

$$R_{u_i} \geq \rho R_{d_i}. \qquad \forall i \in \{1, ..., n\} \tag{34}$$

$$p_{ij} \geq 0, \qquad \forall i \in \{1, ..., n\}, j \in \{1, ..., m\} \tag{35}$$

Problem (P1$'$) can be solved using the lagrangian dual function.

$$\begin{aligned} L_1 = \; & \sum_{i=1}^{n} \sum_{j \in \Omega_i} w \log_2 \left(1 + \tfrac{p_{ij} G_{ij}}{\sigma^2 c_3} \right) + \\ & \left(\sum_{i=1}^{n} \sum_{j \in \Omega_i} -\nu_{ij} p_{ij} \right) + \lambda \left(\sum_{i=1}^{n} \sum_{j \in \Omega_i} p_{ij} - P_T \right) + \sum_{i=1}^{n} \gamma_i \left(\sum_{j \in \Psi_i} p_{ij} - P_t \right) + \\ & \sum_{i=2}^{n} \eta_i \left(\sum_{j \in \Omega_1} w \log_2 \left(1 + \tfrac{p_{1j} G_{1j}}{\sigma^2 c_3} \right) - \sum_{j \in \Omega_i} \tfrac{B_1}{B_i} w \log_2 \left(1 + \tfrac{p_{ij} G_{ij}}{\sigma^2 c_3} \right) \right) + \\ & \sum_{i=1}^{n} \xi_i \left(\rho \sum_{j \in \Omega_i} w \log_2 \left(1 + \tfrac{p_{ij} G_{ij}}{\sigma^2 c_3} \right) - \sum_{j \in \Psi_i} w \log_2 \left(1 + \tfrac{p_{ij} G_{ij}}{\sigma^2 c_3} \right) \right), \end{aligned} \tag{36}$$

where ν_{ij}, λ, γ_i, η_i, and ξ_i are the lagrangian multipliers. Differentiating the lagrangian dual function with respect to p_{ij} and set the derivatives to zero, power can be distributed similar to [12] based on the water-filling algorithm. The subcarrier set of Ω_i is the complement set of Ψ_i; the source of downlink power is at the base station while the uplink power is provided by the mobile users. Thus, we can allocate the downlink and uplink power independently; in the following, first the downlink power is allocated. Assuming p_{1j}, the allocated downlink power to the first user in each subcarrier j, we first differentiate (36) with respect to p_{1j}, and afterwards with respect to p_{ij}.

$$\begin{aligned}\tfrac{\partial L_1}{\partial p_{1j}} = \; & \tfrac{w}{\ln 2} \cdot \tfrac{G_1j}{\sigma^2 c_3 + p_{1j} G_{1j}} + \\ & (-\nu_{1j} + \lambda + \gamma_1) + \textstyle\sum_{i=2}^{n} \eta_i \; \tfrac{w}{\ln 2} \cdot \tfrac{G_{1j}}{\sigma^2 c_3 + p_{1j} G_{1j}} = 0, \quad \forall j \in \Omega_i.\end{aligned} \tag{37}$$

$$\begin{aligned}\tfrac{\partial L_1}{\partial p_{ij}} = \; & \tfrac{w}{\ln 2} \cdot \tfrac{G_ij}{\sigma^2 c_3 + p_{ij} G_{ij}} + \\ & (-\nu_{ij} + \lambda + \gamma_i) - \eta_i \; \tfrac{B_1}{B_i} \cdot \tfrac{w}{\ln 2} \cdot \tfrac{G_{ij}}{\sigma^2 c_3 + p_{ij} G_{ij}} = 0, \quad \forall i \in \{2, ..., n\}, j \in \Omega_i.\end{aligned} \tag{38}$$

From either (37) or (38), it can be shown that,

$$\tfrac{G_{ik}}{\sigma^2 c_3 + p_{ik} G_{ik}} = \tfrac{G_{il}}{\sigma^2 c_3 + p_{il} G_{il}}, \quad \forall k, l \in \Omega_i, \; \forall i \in \{1, ..., n\}. \tag{39}$$

Let us define M_i as the number of allocated subcarrier to user i—M_i is the number of members in the set Ω_i. Without loss of generality, we can assume that $G_{i1} \leq G_{i2} \leq ... \leq G_{iM_i}$ $\forall i \in \{1, ..., n\}$. Therefore, (39) can be rewritten as,

$$p_{ij} = p_{i1} + (\sigma^2 c_3) \tfrac{G_{ij} - G_{i1}}{G_{ij} G_{i1}}, \quad \forall i \in \{1, ..., n\}, j \in \{1, ..., M_i\}. \tag{40}$$

Equation (40) shows the power allocation to each user, in which more power will be allocated to the subcarrier with higher channel gain—which is based on the water-filling algorithm. The total allocated power to user i, P_{iT}, can be calculated as,

$$P_{iT} = \textstyle\sum_{j=1}^{M_i} p_{ij} = M_i p_{i1} + \sum_{j=2}^{M_i} (\sigma^2 c_3) \tfrac{G_{ij} - G_{i1}}{G_{ij} G_{i1}}, \quad \forall i \in \{1, ..., n\}. \tag{41}$$

Using the power constraint (27), and the rate constraint (29), P_{iT} can be computed for all users, and therefore p_{ij} values for each subcarrier.

Solving the lagrangian L_1 to allocate the uplink power is trivial and similar to the first step of downlink power allocation. The same approach can be used for solving problem (P2′)

6. Numerical Investigations

Numerical investigations are performed in MATLB and the implementation details are discussed in this section. For solving problem (P1), we first assign subcarriers based on Algorithms 1 and assume equal power distribution among the allocated subcarriers. Afterwards, power is distributed optimally using TOMLAB optimization toolbox to solve (P1′). Similar arrangement are made for solving problem (P2), thus subcarriers are allocated based on Algorithm 2 with the assumption of equal power across all subcarriers. Afterwards, the optimal power distribution is calculated solving (P2′).

Our simulation runs over a snapshot of the system, i.e. set of n users are in the system and each within a certain distance from the AP. If number of users covered by the AP changes e.g., new users move to the coverage area of the AP or any of the old users moves out from the AP coverage area, the optimization problems will be re-run. Adding new users to the cell is addressed in the first simulated scenario where users are increased from two to fifteen. Also, if users' mobility results in significant changes in the channel condition and therefore in the PER of TCP, the problem will re-run.

The benchmarks are the sum rate maximization problem, denoted by (BM1), and also the sum rate maximization with an equal rate constraint, denoted by (BM2). Clearly, power and subcarrier constraints of Equations (27), (28) and (31) are also applied to the benchmark problems. The resource allocation problem (BM2) is similar to the proposed problem in [12], using the weighting coefficients equal to one. These two benchmarks represents the two extremes of the resource allocation schemes, (BM1) does not consider fairness in the allocations and aims only to achieve the maximum capacity on the link. On the other hand, (BM2) blindly provides fairness with equal rate allocation to all users, which may affect the overall achievable rate significantly.

6.1. Simulation Parameters and Scenarios

We simulate an OFDMA system with 52 subcarriers (this can easily be increased to larger number of subcarriers but the complexity and thus the simulation time will be increased accordingly). The initial value of m_d is set to 32, and its optimal value is calculated through subcarrier allocation in Algorithm 1, while $\rho = 0.2$. It is further assumed that every single transmitted TCP packet is acknowledged (i.e., $b = 1$), and all TCP flows are long-lived and they are in their congestion avoidance phase, i.e. $SS/CA = 1$.

The rest of simulation parameters are similar to the ones used in [2], which are also summarized in Table 1. The available bandwidth is 5 MHz, maximum available power at the base station is 43 dBm, and at each mobile user is 23 dBm. The thermal noise power, σ^2, is -107 dBm (Johnson-Nyquist noise over 5MHz bandwidth), the target BER is 10^{-4}, and the average SNR of wireless channel is 15 dB. The wireless channel is modeled with ITU pedestrian model ($PL = 40\log_{10} d + 30\log_{10} f + 49$, where d is users' distance from the base station and f is the operating frequency), and frequency selective slow fading. The MSS of each TCP flow is set to the standard maximum transfer unit of an Ethernet network which is 1460 bytes. The presented results in this section are average values taken from 150 Monte Carlo simulations.

In order to investigate the performance of the proposed scheme, we consider a number of different scenarios. The first two scenarios mainly explore the achieved fairness among downlink TCP flows. In these two scenarios, We assume fix and constant value for m_d $(= 32)$ in the two proposed schemes and also in the two benchmark problems. Hence, performance of the four schemes having the same allocation boundary between the uplink and downlink is examined. It is also ensured that the initial value of m_d satisfies constraints (30) and (34).

Unlike the first two scenarios, the third simulation scenario assigns the optimal value of m_d through iterations, thus the effect of uplink capacity on the achieved downlink throughput is also examined. In this scenario, we use step (f) in Algorithm 1 to find the optimal value of m_d. Therefore, in scenario three that

Table 1: Simulation Parameters [2].

Bandwidth	5 MHz
Target BER	10^{-4}
Channel model	ITU Pedestrian
Shadowing standard deviation	8 dB
Total power at the base station	43 dBm
Total power at the mobile user	23 dBm
TCP MSS	1460 B
End-to-End RTT	Random distribution
Number of mobile users	2-15

optimal values of m_d are computed, further to the previously discussed enhancement in the fairness among TCP flows, increase in the aggregated end-to-end throughput can be observed.

- **Scenario one:** Various number of mobile users
  The first simulated scenario performs with the fix value of m_d over various number of mobile users from two to fifteen. The end-to-end RTTs of the TCP flows are uniformly distributed random variables in the range $[10ms, 200ms]$.

- **Scenario two:** Different distributions of RTT
  It is expected that an increase of RTT variations among the end-to-end paths highlights the benefit of the TCP-aware allocation schemes (i.e. P1) in comparison to (BM2). In this respect, the second simulation scenario performs over ten mobile users while the RTTs among the ten TCP flows are normally distributed with the average of 100ms and the standard deviations of 10ms, 20ms, 30ms, and 40ms in consecutive simulation runs.

- **Scenario three:** Dynamic m_d assignment, various number of mobile users
  In the third simulation scenario, we initialize m_d with the previously mentioned value–32 subcarriers for the downlink–, afterwards step (f) in Algorithm 1 finds the optimal value of m_d in few iterations. The two benchmark problems (BM1) and (BM2), operate at the $m_d = 32$; thus the results comparison here mainly show how the end-to-end performance is benefitted from setting the border between downlink and uplink, adaptive to the current load of the system.

Under the conditions where the available resources in the uplink are more than required to be allocated for uplink traffic, these wireless resources can be allocated to downlink and increase the downlink throughput. On the other hand, when the available resources for the uplink channel can not satisfy the data rate requirements of the uplink, increasing the number of uplink subcarriers guarantees the delivery of the ACK packets, and enhances the achievable throughput on the downlink. Considering the above, setting the value of m_d dynamically potentially increases the total throughput, and this issue is further investigated in the third simulation scenario.

6.2. Performance Metrics

In this paper, results are presented in terms of throughput and fairness among TCP flows. To study the level of achieved fairness among TCP flows by the presented resource allocation problems, we first compare the results based on the Jain's fairness index [30], denoted by FI. This index is well-used as a quantitative measure of fairness in both wired and wireless networks. The index FI is 1 when there is a complete fair allocation.

$$FI = \frac{\left(\sum\limits_{i=1}^{n} x_i\right)^2}{n \cdot \sum\limits_{i=1}^{n} {x_i}^2}. \tag{42}$$

Assuming x_i is the data rate of user i, proportional to the optimal rate that can be achieved on the corresponding end-to-end path, then FI as described in (42) can be the measure of fairness among end-to-end flows. The optimal throughput for each TCP flow is the theoretical TCP throughput defined by Equation (1) in Section 3. Thus, in Equation (42), x_i can be replaced by R_i/B_i for each end-to-end flow. The proposed resource allocation problem (P2) provides fair allocation among end-to-end TCP flows using the similar approach to the max-min fairness. Therefore, the minimum achieved throughput in each simulation scenario is used as another measure of fairness among the end-to-end flows.

6.3. Numerical Results

In this section, given fixed and dynamic (optimal) m_d, results are presented for the three discussed scenarios. Figure 2 shows fairness index Equation (42)) as achieved by solving problem (P1), (P2), (BM1), and (BM2) in scenario one. Observed from this figure, the achieved fairness index by resource allocation scheme (P1) is increased approximately 30% in average comparing with the results of allocation scheme (BM1). In addition, Figure 2 shows that, as the number of mobile users competing over the wireless link is increased, distribution of the resources in a fair manner is more challenging. Thus, our proposed scheme can enhance the fairness index more significantly, e.g. this index is increased up to 70% in fifteen-user scenario by problem (P1).

The minimum throughput as achieved by problem (P2) is presented in Figure 3, and compared with the similar results from (P1) and the benchmark problems. It can be seen that the minimum throughput as achieved by the four allocation schemes are equal in the two-user and five-user cases. The difference between these values is increased as the number of mobile users increases, e.g the minimum throughput as achieved by (P2) and (P1) in the fifteen user case are more than five times of the achieved throughput by (BM1). This phenomena is similar to the the observation of larger enhancement in fairness index by increasing the number of mobile users as shown in Figure 2.

Despite the enhancement in the achieved fairness (in terms of minimum throughput and Jain's fairness index) by the proposed resource allocation schemes (P1) and (P2), a degradation in the aggregated throughput is expected. We show that decrease in the sum data rate is not significant comparing with the enhancements in fairness. Figure 4 presents the aggregated data rate as achieved in the four investigated resource allocation problems. Here, achieved data rate by each flow is the minimum of allocated data rate and the end-to-end capacity of that flow (i.e. part of the allocated rate that can be utilized by the flow depending on its end-to-end capacity). Figure 4 shows approximately 5% decrease in the average of sum rate computed based on (P1) comparing with (BM1). Moreover, the degradation in the aggregated data rate is increased in the co-existence of larger number of mobile users, e.g. the sum rate as achieved by (P1) is decreased by 15% in the fifteen-user case. This observation is inline with the more enhancements in fairness index and minimum throughput as shown in Figures 2 and 3.

Further observations from these three figures show that the results of problem (P2) and benchmark problem (BM2) are not significantly different. In Figure 2 the fairness index achieved by (P2) is 12% larger, in Figure 3 the minimum achieved throughput by (P2) and (BM2) are the same, and finally in Figure 4 the aggregated data rate accomplished by (P2) is 4% less than (BM2). Taking into account the characteristics of the end-to-end path that is included in problem (P2), it is expected that increasing the diversity among the end-to-end RTTs results in the more significant differences between the achieved results by (P2) and (BM2), thus we investigate this hypothesis in the next simulation scenario.

Figures 5 and 6 show the results of the second simulated scenario. It can be seen that by increasing the diversity among end-to-end RTT values, the aggregated data rate accomplished by the resource allocation scheme (BM2) is decreased comparing with (P2). Also, the enhancement in the fairness index is more significant when the RTT values are more diverse.

Using the same configuration to the simulation scenario one, and setting m_d dynamically, the results for the third simulation scenario as achieved by (P1) and (P2) are presented in Figures 7-9. Observed from Figure 7, we can see that the level of enhancement in the fairness index is similar to the results of the first simulated scenario. Moreover, Figures 9 and 8 show that dynamic allocation of the border between uplink and downlink improve the total aggregated throughput as well as the aggregated downlink throughput by approximately

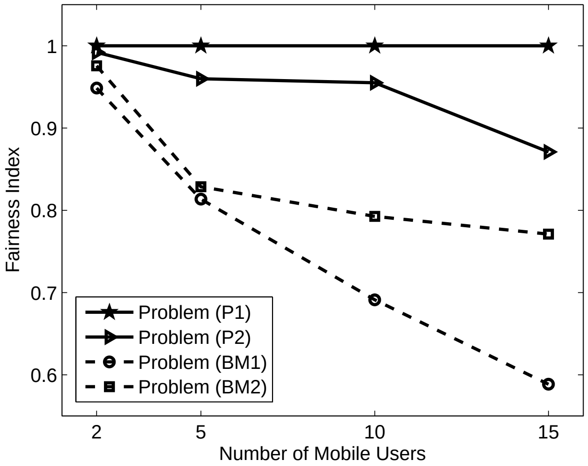

Figure 2: Scenario one – Fix m_d: Jain's Fairness Index as achieved by solving resource allocation problems (P1), (P2), (BM1), and (BM2) Vs. the number of mobile users.

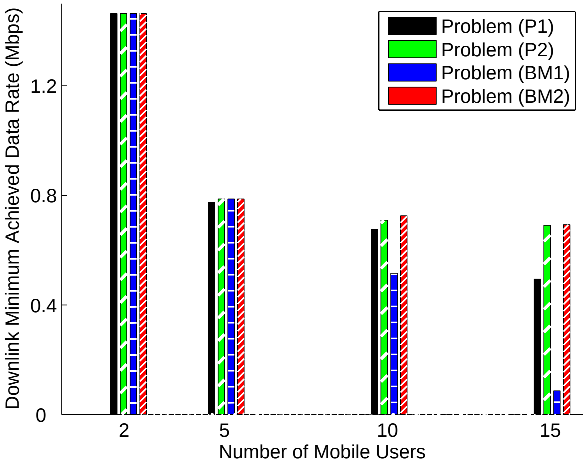

Figure 3: Scenario one – Fix m_d: Downlink minimum data rate (Mbps) as achieved by solving resource allocation problems (P1), (P2), (BM1), and (BM2) Vs. the number of mobile users.

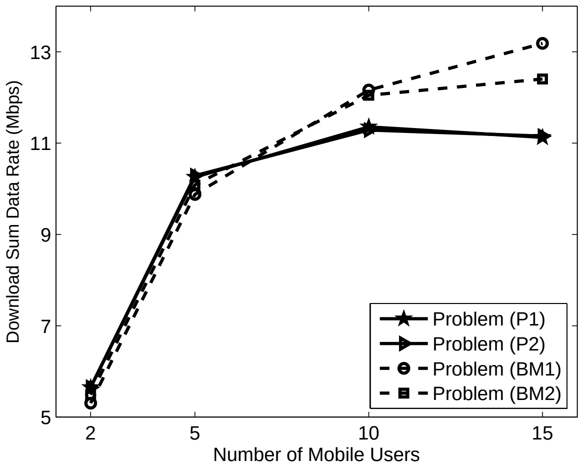

Figure 4: Scenario one – Fix m_d: Downlink sum data rate (Mbps) as achieved by solving resource allocation problems (P1), (P2), (BM1), and (BM2) Vs. the number of mobile users.

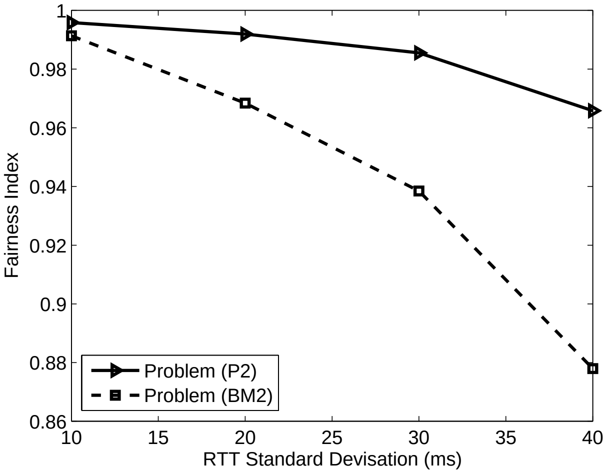

Figure 5: Scenario two—Fixed m_d: Jain's Fairness Index as achieved by solving resource allocation problems (P1), (P2), (BM1), and (BM2) Vs. the standard deviation among end-to-end RTT values.

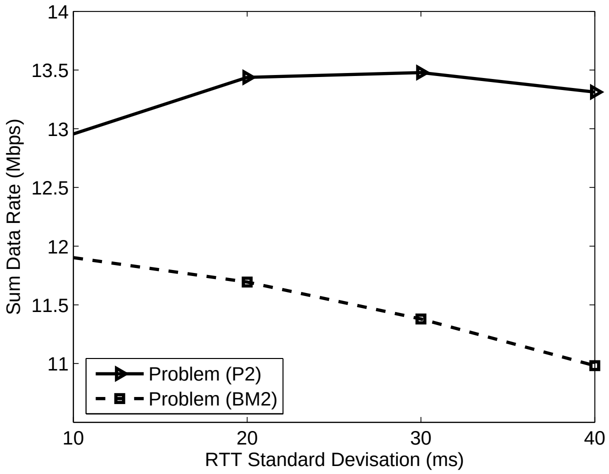

Figure 6: Scenario two—Fix m_d: Downlink sum data rate (Mbps) as achieved by solving resource allocation problems (P2), and (BM2) Vs. the standard deviation among end-to-end RTT values.

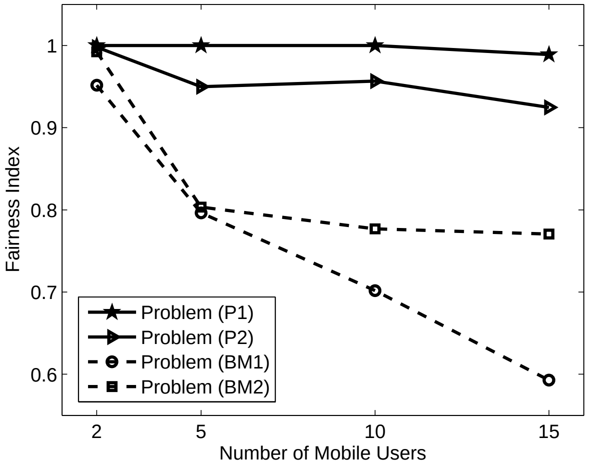

Figure 7: Scenario three – Dynamic m_d: Jain's Fairness Index as achieved by solving resource allocation problems (P1), (P2), (BM1), and (BM2) Vs. the number of mobile users (m_d varies between 52 and 32).

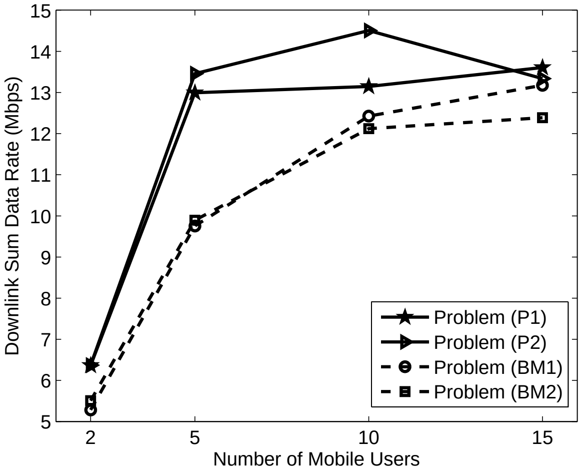

Figure 8: Scenario three – Dynamic m_d: Downlink sum data rate (Mbps) as achieved by solving resource allocation problems (P1), (P2), (BM1), and (BM2) Vs. the number of mobile users (m_d can vary between 50 and 2, having different number of users and traffic load).

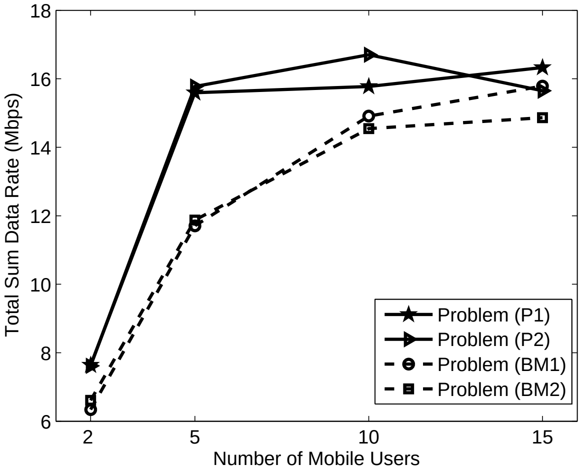

Figure 9: Scenario three – Dynamic m_d: Total (Uplink+Downlink) sum data rate (Mbps) as achieved by solving resource allocation problems (P1), (P2), (BM1), and (BM2) Vs. the number of mobile users (m_d can vary between 50 and 2, having different number of users and traffic load)).

15%. Therefore, setting m_d dynamically and depending on the requirements of TCP connection, not only improve the fairness among end-to-end TCP flows, but also enhance the total achieved throughput.

7. Conclusions

In this paper, the TCP-aware resource allocation algorithm has been proposed. This algorithm contributes to the performance of the end-to-end data transmissions in two folds. First, to provide fairer throughput among TCP flows in the downlink, the theoretical TCP throughputs of all flows have been added to the constraints of the downlink resource allocation problem. Second, to address the problem of asymmetric links and the effect of available uplink resources on the downlink performance, a joint uplink-downlink resource allocation scheme has been proposed. Two different formulations of the TCP-aware downlink resource allocation problem have been discussed. In the first problem, a set of non-linear constraints are added to maintain the proportional downlink rate among users with respect to the TCP theoretical throughput. The second problem attempts to minimize the gap between the allocated data rate and the theoretical TCP throughput. Wide range of simulation scenarios have been carried out to investigate the effect of these resource allocation schemes on the performance of the end-to-end TCP flows. The simulation results have revealed that not only more balanced throughput towards TCP throughput is achieved but also fairness among downlink TCP flows has improved significantly.

The second part of this problem has addressed the issue of scarce availability of resources on the uplink that could result in the degradation of the downlink throughput. This issue and its effect on the performance of TCP has been addressed by proposing a joint uplink-downlink resource allocation scheme that performs in a TCP-aware fashion. Our novel problem has constrained the minimum uplink data rate of each TCP flow based on its allocated data rate in the downlink. The above mentioned constraint is because of the bi-directional nature of TCP connections, which requires sufficient bandwidth in the uplink in order to guarantee the delivery of the downlink packets. The performance of this joint uplink-downlink resource allocation problem has been investigated with a wide range of simulation scenarios. It has been shown that the proposed resource allocation algorithm can enhance the aggregated end-to-end throughput significantly.

8. Acknowledgment

The work reported in this paper has formed part of the Delivery Efficiency Core Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE, www.mobilevce.com. This research has been funded by EPSRC and by the Industrial Companies who are Members of Mobile VCE. Fully detailed technical reports on this research are available to Industrial Members of Mobile VCE.

[1] Information Science Institute, University of Southern California, CA, USA, Transmission Control Protocol, IETF RFC 793 .

[2] TR 25.814, V7.1.0, Physical Layer Aspects for Evloved Universal Radio Access (UTRA), 3rd Generation Partnership Project, Technical Specification Group Radio Access Network .

[3] H. Balakrishnan, V. N. Padmanabhan, G. Fairhurst, M. Sooriyabandara, TCP Performance Implications of Network Path Asymmetry, IETF RFC 3449 .

[4] S. Floys, A. Arcia, J. Iyengar, Adding Acknowledgement Congestion Control to TCP, IETF Internet Draft .

[5] IEEE Std. 802.16e-2005: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, IEEE Computer Society and IEEE Microwave Theory and Techniques Society .

[6] T. Mahmoodi, V. Friderikos, O. Holland, A. H. Aghvami, Cross-layer Optimization to Maximize Fairness among TCP Flows of different TCP Flavors, in: IEEE Global Communications Conference (GLOBECOM '08), 2008.

[7] W. R. Stevens, TCP/IP illustrated, Volume I The protocols, Addison Wesley, 2000.

[8] J. Padhye, V. Firoiu, D. Towsley, J. Kurose, Modeling TCP Throughput: A Simple Model and its Empirical Validation, SIGCOMM Comput. Commun. Rev. 28 (4) (1998) 303–314.

[9] I. C. Wong, B. L. Evans, Optimal Downlink OFDMA Resource Allocation with Linear Complexity to Maximize Ergodic Rates, IEEE Trans. Wireless Commun. 7 (2) (2008) 962–971.

[10] Z. Han, Z. Ji, K. J. R. Liu, Power Minimization for Multi-Cell OFDM Networks Using Distributed Non-cooperative Game Approach, in: IEEE Global Communications Conference (GLOBECOM '04), 2004.

[11] L. Hoo, B. Halder, J. Tellado, J. M. Cioffi, Multiuser Transmit Optimization for Multicarrier Broadcast Channels: Asympotic FDMA capacity region and algorithms, IEEE Trans. Commun. 52 (6) (2004) 922–930.

[12] Z. Shen, J. G. Andrews, B. L. Evans, Adaptive Resource Allocation in Multiuser OFDM Systems With Proportional Rate Constraints, IEEE Trans. Wireless Commun. 4 (6) (2005) 2726–2737.

[13] W. Rhee, J. M. L. Cioffi, Increase in Capacity of Multiuser OFDM System using Dynamic Subchannel Allocation, in: IEEE VTC-Spring, 2000.

[14] Z. Han, Z. Ji, K. J. R. Liu, Fair Multiuser Channel Allocation for OFDMA Networks Using Nash Bargaining Solutions and Coalitions, IEEE Trans. Commun. 35 (8) (2005) 1366–1375.

[15] A. Chockalingam, E. Altman, J. Murthy, R. Kumar, Cross-layer design for optimizing TCP performance, in: IEEE International Conference on Communications (ICC '05), 2005.

[16] S. Pilosof, R. Ramjee, Y. Shavitt, P. Sinha, Understanding TCP fairness over Wireless LAN, in: IEEE Conference on Computer Communications (INFOCOM'03), 2003.

[17] H. Jiang, W. Zhuang, X. Shen, Cross-Layer Design for Resource Allocation in 3G Wireless Networks and Beyond, IEEE Commun. 43 (12) (2005) 120–126.

[18] M. Ghaderi, A. Sridharan, H. Zang, D. Towsley, R. Cruz, TCP-Aware Channel Allocation in CDMA Networks, IEEE Trans. Mobile Comput. 8 (1) (2009) 14–28.

[19] M. Chiang, Balancing Transport and Physical Layers in Wireless Multihop Networks: Jointly Optimal Congestion Control and Power Control, IEEE J Sel. Areas Commun. 23 (1) (2005) 104–116.

[20] L. Galluccio, A. Leonardi, G. Morabito, Tuning Transmission Power for TCP Fairness in Next Generation Wireless Networks: An Analytical Paradigm, Comp. Net. 45 (2) (2004) 207–219.

[21] S. Kim, I. Yeom, TCP-aware Uplink Scheduling for IEEE 802.16, IEEE Commun. Letters 11 (2) (2007) 146–148.

[22] Z. Chen, T. Bu, M. Ammar, D. F. Towsley, Comments on Modeling TCP Reno Performance: A simple model and its Empirical Validation, IEEE/ACM Trans. Net. 14 (2) (2006) 451–453.

[23] H. Jiang, C. Dovrolis, Passive Estimation of TCP Round-Trip Times, ACM Comp. Commun. Review 32 (3) (2002) 75–88.

[24] B. Veal, K. Li, D. Lowenthal, New Methods for Passive Estimation of TCP Round Trip Times, in: the 6th international conference on Passive and Active Network Measurement (PAM'05), Springer-Verlag, 121–134, 2005.

[25] S. T. Chung, A. J. Goldsmith, Degrees of freedom in adaptive modulation: A unified view, IEEE Trans. Commun. 49 (9) (2001) 1561–1571.

[26] T. Mahmoodi, V. Friderikos, O. Holland, A. H. Aghvami, Balancing Sum Rate and TCP Throughput in OFDMA based Wireless Networks, in: IEEE International Conference on Communications (ICC '10), 2010.

[27] T. Mahmoodi, V. Friderikos, H. Aghvami, Allowing Short-Lived TCP Sessions to Ramp-UP in Broadband Wireless Networks, in: IEEE Global Communications Conference (GLOBECOM '09) workshops, 2009.

[28] H. Balakrishnan, H. Katz, V. N. Padmanbhan, The effects of asymmetry on TCP performance, Mobile Networks and Applications 4 (3) (1999) 219–241.

[29] T. Mahmoodi, V. Friderikos, O. Holland, A. H. Aghvami, TCP-aware Resource Allocation Problem in OFDMA based Wireless Networks, in: International Workshop on Cross layer Design (IWCLD), 2009.

[30] R. Jain, D. Chiu, W. Hawe, A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems, Tech. Rep. 301, 1984.