www.ietdl.org

Published in IET Communications
Received on 11th February 2014
Revised on 25th April 2014
Accepted on 7th May 2014
doi: 10.1049/iet-com.2014.0090

ISSN 1751-8628

Cross-layer capacity optimisation in WiMAX orthogonal frequency division multiple access systems with multi-class quality of services and users queue status

Mustafa M. Matalgah[1], Omar M. Hammouri[2], Bimal Paudel[1]

[1]Department of Electrical and Computer Engineering, University of Mississippi, University, MS 38677, USA
[2]P.O. Box 234171, Encinitas, CA, USA
E-mail: ohammouri@gmail.com

Abstract: Given the broad range of applications supported, high data rate required and low latency promised; dynamic radio resource management is becoming vital for the air interface technologies such as WiMAX and LTE adopted by international standards. This study considers the OFDMA system, which has been implemented in both WiMAX and LTE technologies as their air interface multiple access mechanism. A framework for optimised resource allocation with QoS support that aims to balance between service provider's revenue and subscriber's satisfaction is proposed in this study. A cross-layer optimisation design for subchannel and power allocations with the objective of maximising the capacity (in bits/symbol/Hz) subject to fairness parameters and QoS requirements as constraints is presented. The optimisation does not only consider users channel conditions but also queue status of each user as well as different QoS requirements. The QoS classes adopted by the IEEE 802.16e standard, for WiMAX technology, are utilised in this study. In the proposed framework, the problem of power allocation is solved analytically whereas the subchannel allocation is solved using integer programming exhaustive search. The simulation and numerical results obtained in this study have shown improved system performance as compared to other optimisation schemes known in the literature.

1 Introduction

Air interface technologies such as WiMAX and LTE provide quality of service (QoS) support with scheduling services at the media access control (MAC) layer and adopt orthogonal frequency division multiple access (OFDMA) scheme as their multiple-access mechanism. Multiple access is achieved in OFDMA by assigning subsets of subcarriers and time slots to individual users. The subsets of subcarriers considered in frequency domain are referred to as subchannels. The subchannel hence, allows simultaneous low data rate transmission from several users. Based on feedback information about the channel conditions, adaptive user-to-subcarrier assignment can be achieved. Radio resource management tries to efficiently utilise the network resources and the scarcely available radio spectrum while keeping a good grade of services. A significant improvement in the performance of the wireless network can be realised by wisely adopting the cross-layer design approach for optimising resource allocations [1]. Cross-layer design refers to protocol design done by actively exploiting the dependence between protocol layers to obtain certain performance gains, one way of achieving such design is by allowing direct communication between protocols at non-adjacent layers or sharing variables between layers.

Some of the major non-cross-layer techniques with an approach of maximising the capacity while having constraints on total transmit power are maximum sum rate (MSR) [2], maximum fairness (MF) approach [3], proportional fairness approach [4] and proportional rate constraint (PRC) approach [5]. Cross-layer design, however, has been extensively used these days to achieve multiuser diversity gain. This gain is achieved because of channel-state-dependent scheduling where channel state information at the PHY layer is passed on to the packet scheduler at the MAC layer [4, 6]. A simple illustration on the multiuser diversity gain can be found in [1] and a detailed study on the packet scheduling for QoS support in the IEEE 802.16 broadband wireless access system is presented in [7].

In most recent research studies reported in the literature, authors are more focused on cross-layer design approaches [8–16]. Marques *et al.* [8] and Mokari *et al.* [9] present an idea on cross-layer resource allocation to optimise an utility function using channel-state and queue-state information;

however, none of them consider QoS requirement for users as a constraint. An algorithm to maximise the system throughput as a function of queue length subject to QoS requirements was proposed by Tian *et al.* in [10]. In [11] Liu *et al.* propose a cross-layer scheduling algorithm at the MAC layer with multiple connections requiring diverse QoS requirement and verify that the proposed scheduler meet the scheduler design criteria suggested in [12]. In [13], Wang *et al.* explain a QoS-oriented cross-layer packet scheduling algorithm where QoS simply adds on to the delay or queue consideration and is not considered as a separate fairness factor. Hu *et al.* in [14] introduced a cross-layer strategy in the OFDMA system with hybrid adaptive array and switched beam smart antennas. A constraint on total available system power and the effect of users' queue status on resource allocation are not considered in this study. A radio resource allocation for mixed traffic scenario based on channel distribution information is formulated in [15]; however, in this study fairness among users based on their QoS requirement is not considered. Mohanram and Bhashyam in [16] present a novel subcarrier and power allocation scheme used in the multiuser OFDM system. Constraints on user queue status and QoS are not considered in this study.

Given the literature review herein and to the best of the authors knowledge none of the work reported in the literature addresses the problem of cross-layer optimisation by taking into account the channel conditions, queue status and QoS requirements simultaneously. This paper addresses this issue and presents a resource allocation optimisation scheme that takes into account, both the channel conditions and the queue status of each user as well as different QoS requirements to maximise system capacity, which makes the proposed scheme unique to the state-of-the-art research on cross-layer optimisation. The proposed scheme is termed as cross-layer weighted rate constraint (CLWRC) scheme. The significant improvement in the performance of the system in terms of maximisation of system capacity achieved with the implementation of the proposed CLWRC approach is justified by the extensive simulation results.

The rest of the paper is organised as follows. Section 2 presents the proposed cross-layer OFDMA resource allocation system model considered in this study. Section 3 presents the details on different WiMAX QoS service classes and the QoS parameters associated with these classes, also two new cross-layer QoS parameters: service urgency and service satisfaction are introduced in this section. The details on the proposed CLWRC algorithm are presented in Section 4. The simulation and numerical results using the proposed CLWRC approach of capacity maximisation and comparisons with other techniques known in the literature are provided in Section 5. Finally, some conclusions are drawn in Section 6.

2 Cross-layer OFDMA system model

A multiuser WiMAX downlink OFDMA system is shown in Fig. 1. A total of K users sharing L subchannels are considered in the system and the total available transmit power is P_{tot}. Further, the total available system bandwidth, B, is divided into L subchannels. Hence, the bandwidth of each subchannel is B/L and the time slot duration corresponding to each subchannel is $T_s = (L/B)$. Users can be assigned multiple subchannels at a certain time; however, a subchannel cannot be shared among multiple users. Data from users arrive from the MAC layer and is placed into an infinite buffer. These buffers follow a first in first out strategy. A channel fading that follows Rayleigh distribution with envelope $h_{k,l}$ is assumed to be experienced by a user k over subchannel l. Based on the channel-state-information (CSI) and the information on QoS, the subchannel and power allocation algorithm running in BS optimises the subchannel and power allocation to maximise the error-free Shannon capacity while having a constraint P_{tot}. Moreover, the following assumptions are made: (i) outgoing queues for every users are infinite. (ii) The BS has perfectly received the CSI from all subscriber set (SS). (iii) The subchannel and power allocation information is sent to each user on a separate channel. (iv) Coherent bandwidth of the channel is larger than (B/L). This means the channel response on each subchannel is flat. (v) The channel gain remains fixed during one time slot T_s. (vi) Tthe channel is varying in time slow enough that users can estimate the channel perfectly. (vii) All system parameters and QoS parameters associated with all users are assumed to be made available to the BS during the initial setup (signalling) session before the call takes place.

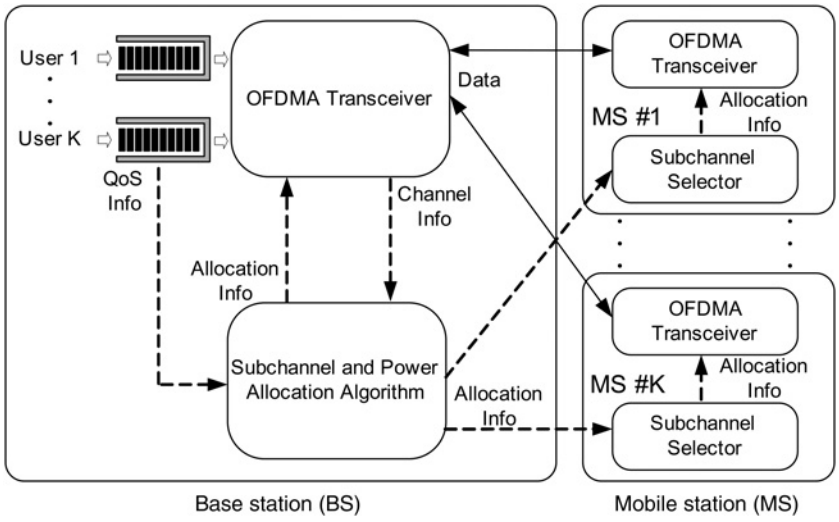

Fig. 1 *Cross-layer WiMAX downlink OFDMA resource-allocation system*

3 WiMAX QoS, service urgency and service satisfaction parameters

In this section, various QoS classes that associate with them a variety of QoS parameters, as defined in standards, are discussed. This section also introduces two cross-layer parameters that can be deemed as QoS parameters. Each of the QoS classes defined in standards and the two newly introduced cross-layer QoS parameters, defined herein, are discussed in detail one by one as follows.

3.1 Standard QoS classes

WiMAX defines five different service classes and associated QoS parameters and different service classes support different applications that have some defined QoS parameters. (i) Unsolicited grant service (UGS) supports real-time service flows that transport fixed-size data packets on a periodic basis, such as voice over IP (VoIP) without silence suppression, (ii) real time polling service (rtPS) supports real-time service flows that randomly transport variable size data packets on a periodic basis, such as moving pictures experts group video, (iii) extended real time polling service (ErtPS) supports real-time service flows that generate packets at variable bit rate with changing bandwidth requirement, such as VoIP with silence suppression, (iv) non-real time polling service (nrtPS) offers unicast polling opportunities on a regular basis, thereby enabling the contention-based polling in the uplink to request bandwidth as in the case of FTP and (v) best effort (BE) supports the application that generates stream of data, such as web browsing with no strict QoS parameter [17, 18].

3.2 Service urgency

Service urgency proposed here is a cross-layer QoS parameter that is dependent on the information about the queues of the services in the data link layer. Let N be the total number of frames considered, then n is defined as the frame number being served such that $n \in \{1, 2, \ldots, N\}$. Now, let $A_k(n)$ be the number of bits arriving at the queue of kth user during frame n, $Q_k(n)$ be the length of queue associated with kth user during frame n and $B_k(n)$ be the number of bits the BS serves from the queue of kth user during frame n. Then the queue length associated with kth user during frame $n+1$ is given by

$$Q_k(n+1) = Q_k(n) + A_k(n) - B_k(n) \tag{1}$$

Every user can be associated with one of the five different service flows. Let SF^x denote an xth service class where x is any element in the set {UGS, ErtPS, rtPS, nrtPS, BE} and $\mathrm{SF}^x(k)$ denote SF^x associated with user k. In other words $\mathrm{SF}^x(k)$ can be defined as

$$\mathrm{SF}^x(k) = \begin{cases} \mathrm{SF}^{\mathrm{UGS}}, & \forall k \text{ with UGS service flow} \\ \mathrm{SF}^{\mathrm{rtPS}}, & \forall k \text{ with rtPS service flow} \\ \mathrm{SF}^{\mathrm{ErtPS}}, & \forall k \text{ with ErtPS service flow} \\ \mathrm{SF}^{\mathrm{nrtPS}}, & \forall k \text{ with nrtPS service flow} \\ \mathrm{SF}^{\mathrm{BE}}, & \forall k \text{ with BE service flow} \end{cases} \tag{2}$$

Also, let Ω^{SF^x} be the set of all users associated with the same SF^x. Then the set Ω^{SF^x} is expressed as

$$\Omega^{\mathrm{SF}^x} = \{1 \le k \le K : \mathrm{SF}^x(k) = \mathrm{SF}^x\} \quad \forall x \in \{\mathrm{UGS}, \mathrm{ErtPS}, \mathrm{rtPS}, \mathrm{nrtPS}, \mathrm{BE}\} \tag{3}$$

and let $Q^{\mathrm{SF}^x}(n)$ be the aggregate queue length corresponding to users associated with the same service class during frame n, then $Q^{\mathrm{SF}^x}(n)$ can be expressed as

$$Q^{\mathrm{SF}^x}(n) = \sum_{k \in \Omega^{\mathrm{SF}^x}} Q_k(n) \tag{4}$$

Finally, the normalised queue length of kth user during frame n, $U_k(n)$, which will be called henceforth the urgency factor, can be defined as

$$U_k(n) = \begin{cases} \dfrac{Q_k(n)}{Q^{\mathrm{SF}^x}(n)}, & \mathrm{SF}^x \in \{\mathrm{rtPS}, \mathrm{nrtPS}, \mathrm{BE}\} \\ 1, & \mathrm{SF}^x \in \{\mathrm{UGS}, \mathrm{ErtPS}\} \end{cases} \tag{5}$$

It should be noted here that the urgency factor $U_k(n)$ is set to 1 for users with a UGS or ErtPS service flow type. It is known from the QoS requirements that the users associated with UGS and ErtPS service classes should be allocated resources on a periodic basis and therefore the concept of urgency does not apply. It should also be noted here that $U_k(n) \in (0, 1]$, UGS and ErtPS service flows are thus assigned the highest urgency factor. However, the urgency factor for rtPS, nrtPS and BE are calculated using (5). It is important to note that the concept of urgency factor does not apply if users do not have any queue. The significance of the urgency factor is two-fold. It gives indication about which user is being under-served relative to other users of the same service flow, and it also conveys information about the queue length of the user to the resource allocation algorithm. The higher the value of $U_k(n)$, the more it is urgent to allocate resources to the user.

3.3 Service satisfaction

Service satisfaction based on different kinds of service flows depends on the information such as data rate, delay satisfaction indicator or flow's coefficient as defined in [11]. Hence, service satisfaction can be deemed as the cross-layer QoS and is considered in this study. Let $\{\gamma_{\mathrm{UGS}}, \gamma_{\mathrm{ErtPS}}, \gamma_{\mathrm{rtPS}}, \gamma_{\mathrm{nrtPS}}, \gamma_{\mathrm{BE}}\}$ be defined as a set of configurable system parameters. Each γ_{SF} denotes a weighting factor that can be used to favour one service class over the other and be configurable depending on the system deployment.

Now let $\mathrm{S}_k(n)$ be the satisfaction factor associated with kth user during frame n. For UGS service flows the satisfaction factor is defined as

$$S_k(n) = \frac{1}{\gamma_{\mathrm{UGS}}} \tag{6}$$

where γ_{UGS} is the UGS class weighting factor. Therefore the satisfaction factor is constant for all the users with UGS service flows and over all frames. As for ErtPS service flows, the satisfaction factor is defined as

$$S_k(n) = \frac{1}{\gamma_{\mathrm{ErtPS}}} \tag{7}$$

doi: 10.1049/iet-com.2014.0090

where γ_{ErtPS} is the ErtPS class weighting factor. Also, the satisfaction factor is constant for all the users with ErtPS service flow and over all frames. For rtPS service flows, if the waiting time of the packet in a queue exceeds a maximum allowed latency or the deadline T_k, then a timeout is defined by the scheduler and hence the satisfaction factor is defined as

$$S_k(n) = \frac{DS_k(n)}{\gamma_{\text{rtPS}}} \tag{8}$$

where γ_{rtPS} is the rtPS class weighting factor, $DS_k(n)$ is the delay satisfaction indicator, which is defined as

$$DS_k(n) = \max\{1,\ T_k - \Delta T - W_k(n) + 1\} \tag{9}$$

where $W_k(n) \in [0,\ T_k]$ is the head of line delay which is defined as the longest waiting time that a packet experiences and $\Delta T \in [0,\ T_k]$ is the guard time region ahead of the deadline T_k, which indicates the time remaining before which the packet should be scheduled to avoid timeout [11]. A lower value of satisfaction factor will require a scheduling algorithm to allocate more resources to the service to meet the delay requirements. The satisfaction factor for users with rtPS service flow has a minimum value of $(1/\gamma_{\text{rtPS}})$. For nrtPS service flows, the satisfaction factor is defined as

$$S_k(n) = \frac{\text{RS}_k(n)}{\gamma_{\text{nrtPS}}} \tag{10}$$

where γ_{nrtPS} is the nrtPS class weighting factor, $\text{RS}_k(n)$ is the rate satisfaction indicator which is defined as

$$\text{RS}_k(n) = \max\{1,\ \hat{\eta}_k(n)/\eta_k\} \tag{11}$$

where η_k is the minimum reserved data rate for kth user, and $\hat{\eta}_k(n)$ is the exponentially weighted average data rate of kth user up to frame n obtained by using the exponentially weighted low-pass filter [19] and can be defined as

$$\hat{\eta}_k(n+1) = \begin{cases} C_k(n), & n = 0 \\ \hat{\eta}_k(n)\left(1 - \dfrac{1}{t_c}\right) + C_k(n)\dfrac{1}{t_c}, & n > 0 \end{cases} \tag{12}$$

where $C_k(n)$ is the user data rate allocated during frame n to kth user. The parameter t_c, window size, controls the latency of the system [18, p. 213]. The satisfaction factor, $S_k(n)$, ensures that the user is receiving an average data rate above the minimum reserved rate, $\hat{\eta}_k(n) \geq \eta_k$. If $\text{RS}_k(n) \geq 1$, then the rate requirement is satisfied, which increases the satisfaction factor. Large values of $\text{RS}_k(n)$, therefore, indicate high degree of satisfaction. The minimum value for $\text{RS}_k(n)$ is 1, which is when the user is underserved and should be allocated more resources to meet the minimum rate requirements. The satisfaction factor for users with nrtPS service flow has a minimum value of $(1/\gamma_{\text{nrtPS}})$. For BE service flows, the satisfaction factor is defined as

$$S_k(n) = \frac{1}{\gamma_{\text{BE}}} \tag{13}$$

where γ_{BE} is the BE class weighting factor. Therefore the satisfaction factor is constant for all the users with BE service flow and over all frames. The reason is that by definition of the QoS requirements, the users with BE service flow should be allocated resources after all other service flows are satisfied, and therefore, the concept of service satisfaction does not apply. The significance of the satisfaction factor is also two-fold. It allows for scalability, as when the system is overloaded, the performance of users with low-priority service classes will be degraded prior to those with high priority service classes, and it also allows users with low-priority service classes to lead when users with higher-priority service classes are well satisfied.

4 Proposed cross-layer algorithm

4.1 Proposed algorithm: optimisation problem formulation

Let $P_{k,l}(n)$ be the power allocated to kth user over subcarrier l during frame n, N_0 be the additive white Gaussian noise (AWGN) power spectral density with zero mean, $h_{k,l}$ be the channel gain for user k over subchannel l and $\rho_{k,l} \in \{0,\ 1\}$ indicates whether or not a subchannel l is used by user k. Then, the spectral efficiency or channel capacity, in bits/symbol/Hz, for a kth user during frame n is expressed as

$$C_k(n) = \sum_{l=1}^{L} \frac{\rho_{k,l}}{L} \log_2\left[1 + P_{k,l}(n) H_{k,l}(n)\right] \text{ bits/symbol/Hz} \tag{14}$$

where

$$H_{k,l}(n) = \frac{h_{k,l}^2}{N_0(B/L)} \tag{15}$$

and the weighted capacity, $R_k(n)$ during frame n, is then expressed as

$$R_k(n) = \frac{U_k(n)}{S_k(n)} C_k(n) \tag{16}$$

The weighted capacity in (16) incorporates both the urgency factor and the satisfaction factor. It is known from the explanation of urgency factor that the services with higher queue lengths have higher urgency factor except for the case of UGS and ertPS service flows where they have highest urgency irrespective of the queue lengths. Hence, the service with highest service urgency requirement needs to be scheduled first and hence the weighted capacity is so defined that it is directly proportional to the urgency factor. However, the satisfaction factor indicates the satisfaction level of the services. If a service meets a specified data requirement, a delay requirement or any other requirements specific to the QoS, then the satisfaction is high. Hence, the service with highest satisfaction can be scheduled later and hence the weighted capacity is defined to be inversely proportional to the service satisfaction. Now the fairness constraint is defined as

$$R_i(n) = R_j(n) = R(n) \quad \forall i, j \in [1, K] \tag{17}$$

Based on the above discussion, the optimisation problem can

be expressed mathematically as

$$\max_{P_{k,l},\rho_{k,l}} C = \sum_{k=1}^{K}\sum_{l=1}^{L}\frac{\rho_{k,l}}{L}\log_2\left(1+P_{k,l}H_{k,l}\right) \quad \text{bits/symbol/Hz} \tag{18}$$

$$\text{subject to } \sum_{k=1}^{K}\sum_{l=1}^{L}P_{k,l} \le P_{\text{tot}} \tag{19}$$

$$P_{k,l} \ge 0 \quad \forall k, l \tag{20}$$

$$\rho_{k,l} = \{0, 1\} \quad \forall k, l \tag{21}$$

$$\sum_{k=1}^{K}\rho_{k,l} = 1 \quad \forall l \tag{22}$$

$$R_i(n) = R_j(n) = R(n) \quad \forall i, j \in [1, K] \tag{23}$$

where P_{tot} is the total available power. The first constraint (19) implies that the total power used by subchannels is not to exceed the total available power. The second constraint (20) states that the power used by all subchannels should be non-negative. In the third constrain (21), $\rho_{k,\ l}$ is only allowed to be 0 or 1 which assures that a lth subchannel is either used or not used by the kth user. Furthermore, no sharing of subchannel is allowed, which is stated by the fourth constraint (22). The last constraint (23) is the fairness constraint presented in (16) and (17) and the introduction of this constraint is what makes the proposed optimisation problem unique as opposed to the one formulated in [5].

4.2 Proposed algorithm: problem solution and implementation

An optimum solution to the optimisation problem is highly computationally complex, so a suboptimal solution is proposed, where subchannel and power allocations are performed separately. An analytical solution to the optimisation problem in (18) can be obtained using the method of Lagrange multipliers as follows. For a given subchannel allocation, Π_k, such that Π_k is the set of subchannels allocated to user k, the capacity of user k during frame n, in bits/symbol/Hz, is expressed as

$$C_k(n) = \sum_{l\in\Pi_k}\frac{1}{L}\log_2\left(1+P_{k,l}H_{k,l}\right) \tag{24}$$

Then the optimisation problem in (18) is reformulated as

$$\max_{P_{k,l}} C = \sum_{k=1}^{K}\sum_{l\in\Pi_k}\frac{1}{L}\log_2\left(1+\frac{P_{k,l}h_{k,l}^2}{N_0(B/L)}\right) \quad \text{bits/symbol/Hz} \tag{25}$$

$$\text{subject to } \sum_{k=1}^{K}\sum_{l\in\Pi_k}P_{k,l} \le P_{\text{tot}} \tag{26}$$

$$P_{k,l} \ge 0 \quad \forall k, l \tag{27}$$

$$\Pi_i \cap \Pi_j = \Phi \quad \forall i \ne j \tag{28}$$

$$\Pi_1 \cup \Pi_2 \cup \cdots \cup \Pi_K \subseteq \{1, 2, \ldots, L\} \tag{29}$$

$$R_i(n) = R_j(n) = R(n) \quad \forall i, j \in [1, 2, \ldots, K] \tag{30}$$

The solution to the optimisation problem in (25) results in

$$P_{k,x} = P_{k,1} + \frac{H_{k,x} - H_{k,1}}{H_{k,x}H_{k,1}} \tag{31}$$

for $k \in \{1, 2, \ldots, K\}$ and $x \in \{1, 2, \ldots, |\Pi_k|\}$. This result is obtained by solving optimisation problem in (25) using the method of Lagrange multipliers as discussed and elaborated in [5] is used. The expression in (31) is the water-filling equation, which means that subchannels with higher SNR receive more power in order to maximise the capacity. A similar equation was obtained in [5] for different constraints (as indicated in Section 4.1).

(1) *Resource allocator:* The proposed resource allocator algorithm based on the two-phase greedy approach is shown in Algorithm 1 (see Fig. 2). The terms used in this algorithm are defined as follows: T_{f} is the frame duration and T is the total traffic duration, such that $T = N \times T_{\text{f}}$, T_{c} and T_{s} are the in-phase and quadrature phase E-field components of Rayleigh fading channel, respectively. $P_{k(\text{tot})}$ is the initial total power allocation to kth user and $\bar{C}$ is the exponentially weighted average capacity.

The working of the resource allocator algorithm depicted in Algorithm 1 (Fig. 2) is described in detail in the following. It first reads the queues lengths, $Q_k(n)$, service flows associated with each user, $\text{SF}^x(k)$, the maximum delay accepted for every rtPS user, T_k, the minimum data rate accepted for every nrtPS, $\eta_k(n)$, from the MAC layer. Likewise, K, L, P_{tot}, B, N_0, γ_{SF^x}, t_c and ΔT are configured by the allocator. With all the information in hand traffic corresponding to different types of QoS classes as discussed in Section 3.1 and based on Table 2 are simulated and queue lengths corresponding to each service flows are calculated as explained by (4). Service urgency and satisfaction parameters are then evaluated using (5)–(13) and a Rayleigh fading channel based on Clarks's model [20, p. 214] is simulated as in [5].

Algorithm 1:

Input: $K,\ L,\ P_{tot},\ B,\ N,\ T_f,\ N_0,\ \gamma_{SF^x},\ \Delta T,\ t_c,\ T_k,\ \eta_k$

Initialise Array: $\bar{C}_k \Leftarrow 0,\ Q_k \Leftarrow 0$

generate $Q_k(n)$ $\quad \forall\ n \in \{1, 2, \ldots, N\}$, $\forall\ k$ with respective SF^x

for $n = 1 \rightarrow N$ **do**

- *find* $Q^{SF^x}(n)$ // *by* (4)
- *find* $U_k(n)$ // *by* (5)
- *find* $S_k(n)$ // *by* (6)-(13)
- *generate* T_c & T_s
- $h_{k,l} \Leftarrow T_c + jT_s$
- **invoke** Subchannel Allocator // *assigns* $\rho_{k,l}$
- **for** $k = 1 \rightarrow K$ **do**
  - $P_{k(tot)} \Leftarrow \frac{P_{tot}}{L}\sum \rho_{k,i}$
  - **invoke** Power Allocator // *assigns* C_k
- **end**
- $C(n) \Leftarrow \sum_{k=1}^{K} C_k$
- **if** $n = 1$ **then**
  - $\bar{C}(n) \Leftarrow C(n)$
- **else**
  - $\bar{C}(n) \Leftarrow \bar{C}(n-1) * (1 - \frac{1}{t_c}) + C(n) * \frac{1}{t_c}$
- **end**

end

Fig. 2 *Resource allocator*

doi: 10.1049/iet-com.2014.0090

The algorithm then proceeds forward by invoking the subchannel allocation, which starts with assigning the subchannel with maximum channel gain for each user in Rayleigh fading environment. The total available system power is equally divided among channels and the weighted data rates for each user dependent on the urgency and satisfaction factor are then calculated. The weighted data rates so generated are then evaluated to allocate the remaining subchannels to the users such that the fairness among users in terms of weighted data rate is maintained. Once subchannel allocation is completed, power allocation (water-filling) is performed based on the derivation result in Section 4.2. Finally, the overall exponentially weighted total average system capacity is evaluated.

5 Simulations and numerical results

In this section, we numerically implement and simulate the solution described in Section 4.2 for the optimisation problem presented in Section 4.1 based on Algorithm 1 (Fig. 2). Table 1 shows the system parameters used for the simulation. In this paper, the same channel model as in [5] is considered. Likewise, five different traffic models (described in Section 3.1) were used to simulate the arrival patterns of the five different service flows. Table 2 summarises the characteristics of the 10 different active users' service flow in the system based on five different service flows.

5.1 Capacity comparison

The results in Fig. 3 show the total average system capacity against average SNR based on the proposed CLWRC algorithm that implements the solution approach presented in Section 4.2 for the optimisation problem formulated in Section 4.1. In this figure, the algorithm is executed for different values of average SNR, where for simplicity, symbol energy is assumed to be 1 J and the system is assumed to be serving 10 users. Hence, for each value of the average SNR in Fig. 3, the corresponding power spectral density of the AWGN channel is evaluated and used in the optimisation problem. Furthermore, the other system parameters needed in the computation are listed in Table 1. For performance comparison purposes Fig. 3 also shows results for PRC algorithm [5] and MF algorithm [3] along with the proposed CLWRC algorithm. A different approach of power distribution among users is used and discussed in [5]. The figure also shows the Shannon theoretical limit. The optimised capacity curves for PRC and MF are based on the solution approach proposed in [5], where MF is explained as the special case of PRC. It can be seen from the figure that the proposed CLWRC algorithm achieves a higher total average system capacity throughout the observed average SNR range (−5 to 30 dB) as compared to PRC and MF algorithms.

Likewise, the results in Fig. 4 depict the total average system capacity against frame number (1–1000). The system parameters listed in Table 1 are used in this case as well. The total average system capacity achieved during each arrival time duration of a frame is depicted in the figure, for an average SNR of 10 dB, using the proposed CLWRC, PRC and MF algorithms. It is obvious from the

Table 1 System simulation parameters

Symbol	Parameter	Value
P_{tot}	total system power	1 W
L	number of subchannels	64
K	number of users	10
N	number of frames	1000
B	total system bandwidth	5 MHz
E_s	symbol energy	1 J
T_f	frame duration	5 ms
Δ_T	guard time ahead of deadline	20 ms
T_c	moving average window size	1000 ms
γ_{UGS}	UGS weighting factor	0.8
γ_{rtPS}	rtPS weighting factor	0.6
γ_{ErtPS}	ErtPS weighting factor	0.4
γ_{nrtPS}	nrtPS weighting factor	0.3
γ_{BE}	BE weighting factor	0.2

Table 2 Traffic simulation parameters

Users (k)	SF	Parameter	Value
user1 (U1)	UGS	CODEC	G.729
		voice processing interval	20 ms
		packet size	66 Bytes
user2 (U2)	UGS	CODEC	G.728
		voice processing interval	30 ms
		packet size	106 Bytes
user3 (U3)	rtPS	Bernoulli trial ($\bar{p}$)	0.4
		mean rate	64 Kbps
		maximum delay	30 ms
user4 (U4)	rtPS	Bernoulli trial ($\bar{p}$)	0.5
		mean rate	16 Kbps
		maximum delay	50 ms
user5 (U5)	ErtPS	CODEC	G.723.1
		voice processing interval	30 ms
		packet size	66 Bytes
		mean ON period	1.2 s
		mean OFF period	1.8 s
user6 (U6)	ErtPS	CODEC	G.711
		voice processing interval	20 ms
		packet size	206 Bytes
		mean ON period	1.2 s
		mean OFF period	1.8 s
user7 (U7)	nrtPS	mean rate	512 Kbps
		minimum rate	128 Kbps
user8 (U8)	nrtPS	mean rate	1 Mbps
		minimum rate	1 Mbps
user9 (U9)	BE	pareto mean rate	10 558 bps
		lognormal mean rate	724 bps
user10 (U10)	BE	pareto mean rate	10 558 bps
		lognormal mean rate	7247 bps

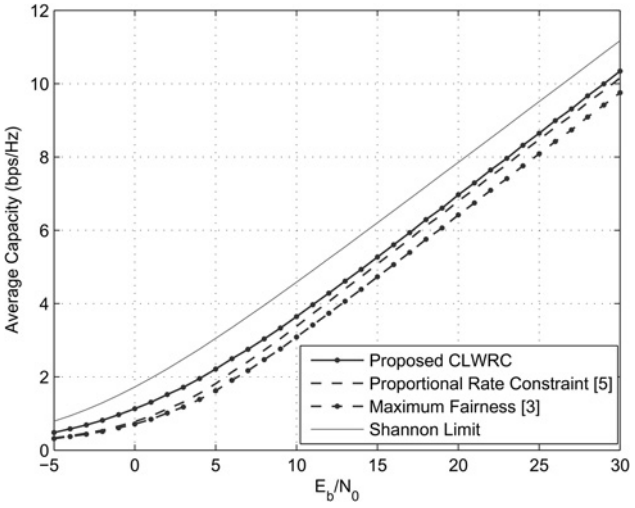

Fig. 3 *Average capacity (bps/Hz) against average SNR per symbol (based on simulation parameters in Table 1)*

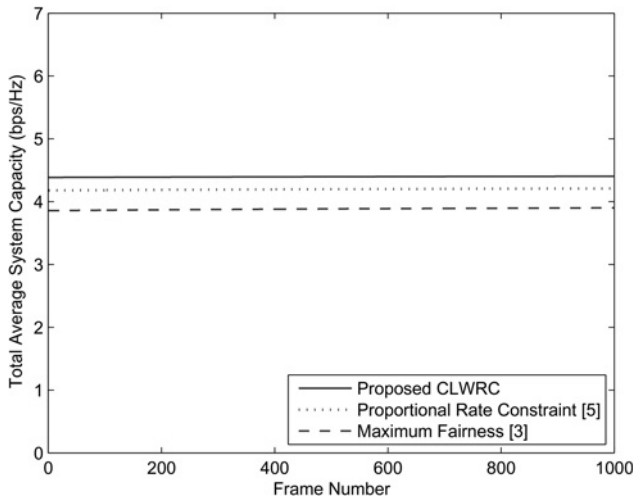

Fig. 4 *Total average system capacity (bps/Hz) against frame number (based on simulation parameters in Table 1)*

figure that the proposed CLWRC algorithm remains superior as compared to the systems implementing PRC and MF algorithms.

The results in Fig. 5 depict the total average system capacity against number of users. For each case of number of users in the system users are equally assigned over the different service classes with defined weighting factor $\gamma_{SF^x}(k)$ (e.g., if $K = 30$ users, 6 out of these 30 users are assigned to each of the 5 different service classes: that is, $SF^x(k) = SF^{UGS}$ for $k = 1, 2, \ldots, 6$, $SF^x(k) = SF^{ertPS}$ for $k = 7, 8, \ldots, 12, \ldots$, $SF^x(k) = SF^{BE}$ for $k = 25, 26, \ldots, 30$). The values of γ_{SF^x} are taken from Table 1, and a system with an average SNR value of 10 dB is assumed. It can be seen from the graph that the proposed algorithm maintains its performance superiority to maximum optimum level as compared to other algorithms for all the range of the number of users. Furthermore, it is also observed in the case of the proposed algorithm, the system capacity increases with the increase in number of users. This behaviour confirms the multiuser diversity advantage in the case of the proposed CLWRC algorithm and is another powerful aspect as compared to PRC and MF algorithms.

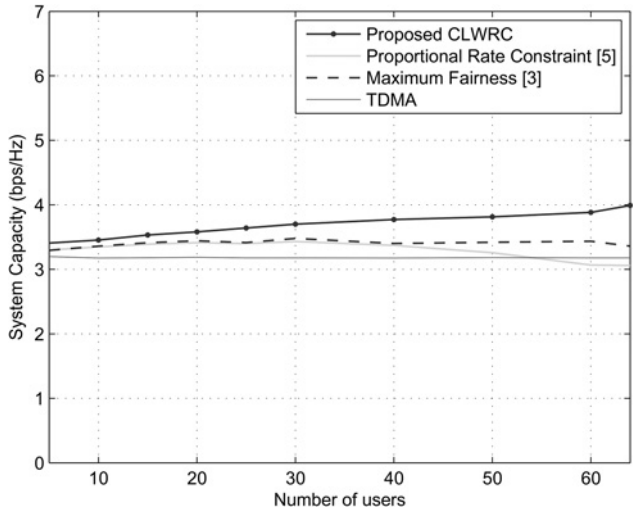

Fig. 5 *Total system capacity (bps/Hz) against no. of users (based on simulation parameters in Table 2 and for average SNR of 10 dB)*

The reason behind this increase in system capacity is described as follows. The proposed algorithm maximises the total system capacity while having constraint on the weighted capacity as governed by (16). For the user with high QoS requirement, the urgency factor is higher whereas the satisfaction factor is lower which increases the weighted capacity of the system that the algorithm tries to maintain. Hence, in our optimisation process, since the different QoS classes are equally assigned to the users, as the number of users in the system increases the users that belong to QoS class with higher QoS service requirement will contribute in increased total average system capacity, as shown in Fig. 5. In contrast, TDMA algorithm compared with here, does not consider urgency and satisfaction factors and hence the capacity is independent of the number of users as is clear in the figure. In the case of MF, the algorithm maximises the system capacity while having constraint on the transmission rate itself and only a slight variation is observed; however, the PRC algorithm considers the fairness parameter and hence the capacity of the system depends on the proportional factor assigned to each user in the system. The PRC algorithm tries to maximise the total system capacity while having constraint on the proportional rate. Therefore in PRC, the proportional rate of the user with lower rate is boosted whereas the one with higher rate is decreased, such that proportionality is maintained among users. Hence, for higher numbers of users in the system, the users with higher proportional factor will cause the system capacity to decrease and the same is reflected in the figure.

The results in Fig. 6 depict a comparison between average system capacity for different number of users pertaining to the proposed CLWRC scheme. A scenario of varying number of users in the system as discussed for the results in Fig. 5 is considered and the optimised average system capacity curves using the proposed CLWRC algorithm for the system serving 10, 30 and 60 users are plotted in Fig. 6. It is evident from the figure that as the number of users in the system increases, there is an improvement in the system capacity. These results confirm the multiuser diversity advantage in the proposed CLWRC scheme.

To demonstrate the starvation effectiveness of our proposed algorithm, we use a users-based stacked-bar plot as depicted in Fig. 7. In Fig. 7 all the attainable user

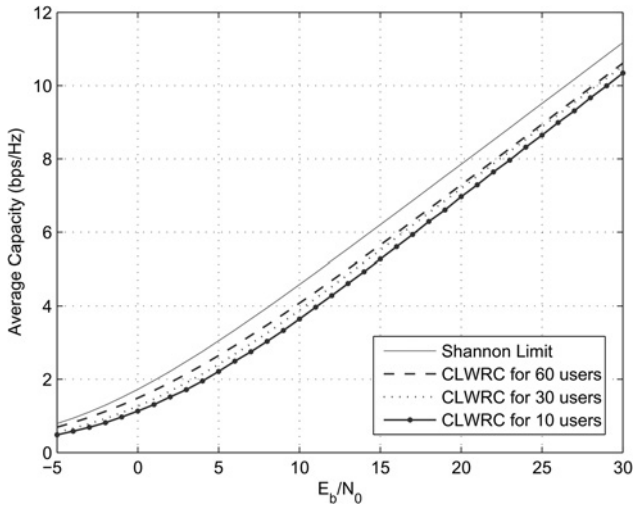

Fig. 6 *Average capacity (bps/Hz) against average SNR per symbol (system serving 10, 30 and 60 users and implementing the proposed CLWRC algorithm)*

doi: 10.1049/iet-com.2014.0090

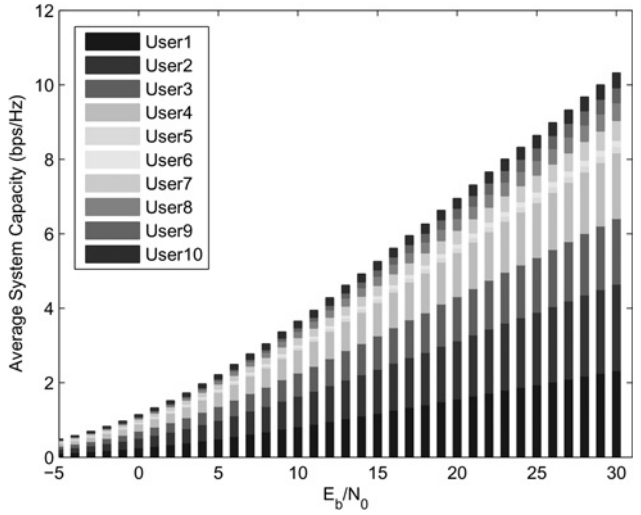

Fig. 7 *System users-based stacked-bar plot to demonstrate the starvation effectiveness in terms of the average capacity (bps/Hz) against average SNR per symbol*

capacities for all individual users being served by the system are stacked in a column over the observed SNR in the range of −5 to 30 dB. A system serving 10 users is considered and the traffic corresponding to each users is depicted in Table 2. It can be observed from the figure that for each SNR value in the entire considered SNR range, every user has a non-zero attainable capacity which indicates that some resources have been allocated to every user according to its QoS requirement and queue status and hence none of them is deprived of its resources. The lower values of the attainable capacity for the users with rtPS, nrtPS and BE traffic as compared to the users with UGS and ertPS traffic do not indicate the deficiency or the starvation of resources, rather indicate a fair distribution of resources among them. Hence, UGS and ErtPS users with high priority and lower satisfaction factors are allocated more resources as compared to rtPS, nrtPS and BE users.

The fairness of the proposed algorithm among the different served users is shown in Fig. 8 and is compared with the MSR algorithm based on the attainable capacity

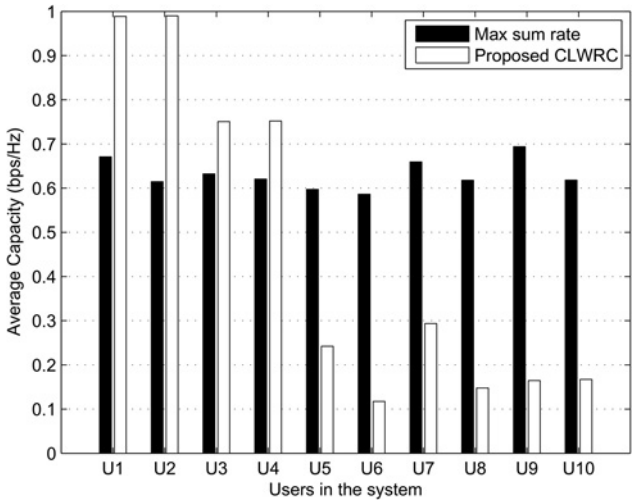

Fig. 8 *Average capacity (bps/Hz) against classified users to demonstrate the fairness of the CLWRC among the classified users served by the system*

Table 3 Execution time (s) of different algorithms for different number of frames

Algorithms	No. of frames		
	1000	100	10
CLWRC algorithm	107.2528	10.7973	1.1079
PRC algorithm	132.6554	13.0038	1.4211
MF algorithm	136.1923	13.5061	1.4498

corresponding to each individual user, with a QoS class and queue status, being served by the system. A system serving 10 users is considered and the traffic corresponding to each users is as depicted in Table 2. The average system SNR is assumed to be 10 dB. It can be seen from the figure that in MSR algorithm the resources are divided among different users irrespective of the QoS demand of that user to maximise the overall system capacity hence almost equal attainable capacity is observed. However, in the proposed CLWRC algorithm a fair, proportional, distribution of resources among the users according to their QoS and queue status is made. This is why an increase in the attainable capacity for the users with higher weighting factors and urgency factors as well as lower satisfaction factors (UGS and ErtPS users) is observed as compared to MSR algorithm, whereas a decrease in the attainable capacity is observed for other users. The decrease in capacity for some users does not mean the loss of capacity, it implies that the capacity is being transferred from one user to the other. This transfer of the user capacity among each other is the crux of fairness. It is worth mentioning that fairness does not imply equality.

5.2 Complexity comparison

As a case study, time complexity is considered here for comparison between the proposed CLWRC algorithm with PRC and MF algorithms. Table 3 shows a comparison between execution times (in seconds) of the proposed CLWRC with the other known PRC and MF algorithms. The algorithm was executed on Intel(R) Core(TM) i5-2430M CPU at 2.40 GHz 2.40 GHz processor for 10 MonteCarlo runs. It can be observed from the table that the proposed CLWRC algorithm has a faster execution time as compared to PRC and MF algorithms. This is because of the fact that in both PRC and MF algorithms, a complex initial power distribution among users, as in [5], is implemented whereas in the CLWRC algorithm an equal power distribution among users is considered.

6 Conclusion

In this paper, a novel resource allocation optimisation scheme for the OFDMA system with multi-class QoS and user queue status is presented. From the numerical results, it has been observed that the proposed scheme results in total average system capacity that is closer to the Shannon limit than other known resource allocation schemes. On the other hand, unlike other known techniques, the proposed algorithm not only maintains the optimum system capacity for different number of users in the system but also increases as the number of users increases. In particular, the proposed cross-layer resource allocator scheme outperforms other known approaches in three aspects; closeness to

Shannon capacity limit, consistency in terms of maximum optimum capacity throughout the frames considered and consistency in maintaining maximum optimum system capacity for different number of users. Similarly, the results presented show the efficiency of the algorithm to minimise the starvation problem and maintain the fairness among users. There are various areas where this work could be extended. An immediate extension to this work would be to consider an error prone channel where the optimisation would be subject to adaptive modulation and coding for a more practical consideration. Study based on imperfect CSI and finite buffer considerations will also be a valuable extension to this work. While WiMAX QoS classes have been utilised in developing the work of this paper, the work can also be extended to LTE standard QoS classes. A comparison on the system performance while implementing the proposed scheduling scheme between WiMAX and LTE will also be an interesting extension to this work. It would also be interesting to consider the scenario with a majority of users traffic demanding the same QoS class and observe the performance of the algorithm. Moreover, the algorithm can be extended to support the control plane besides the data plane and also can be enhanced by supporting multiple users sharing sub-channels in time, adding another dimension to multiuser diversity.

7 References

1 Shakkottai, S., Rappaport, T., Karlsson, P.: 'Cross-layer design for wireless networks', *IEEE Commun. Mag.*, 2003, **41**, (10), pp. 74–80

2 Zhang, Y., Letaief, K.: 'Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems', *IEEE Trans. Wirel. Commun.*, 2004, **3**, (5), pp. 1566–1575

3 Rhee, W., Cioffi, J.: 'Increase in capacity of multiuser OFDM system using dynamic subchannel allocation'. Proc. IEEE Vehicular Technology Conf., Tokyo, Japan, May 2000, vol. 2, pp. 1085–1089

4 Tse, D.: 'Forward link multiuser diversity through proportional fair scheduling'. Pmtresentation at Bell Labs, August 1999

5 Shen, Z., Andrews, J., Evans, B.: 'Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints', *IEEE Trans. Wirel. Commun.*, 2005, **4**, (6), pp. 2726–2737

6 Knopp, R., Humblet, P.: 'Information capacity and power control in single cell multiuser communications'. Proc. IEEE Int. Conf. Communications, Seattle, Washington, June 1995, vol. 1, pp. 321–335

7 Wongthavarawat, K., Ganz, A.: 'Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems', *Int. J. Commun. Syst.*, 2003, **16**, (1), pp. 81–96

8 Marques, A.G., Lopez-Ramos, L.M., Giannakis, G.B., Ramos, J., Caamano, A.J.: 'Optimal cross-layer resource allocation in cellular networks using channel- and queue-state information', *IEEE Trans. Veh. Technol.*, 2012, **61**, (6), pp. 2789–2807

9 Mokari, N., Javan, M.R., Navaie, K.: 'Cross-layer resource allocation in OFDMA systems for heterogenous traffic with imperfect CSI', *IEEE Trans. Veh. Technol.*, 2010, **59**, (2), pp. 1011–1017

10 Tian, H., Xu, H., Gao, Y., Wang, S., Zhang, P.: 'QoS-oriented cross-layer resource allocation with finite queue in OFDMA systems'. Proc. IEEE Wireless Communications and Networking Conf., Las Vegas, Nevada, March–April 2008, pp. 1519–1524

11 Liu, Q., Wang, X., Giannakis, G.B.: 'A cross-layer scheduling algorithm with QoS support in wireless networks', *IEEE Trans. Veh. Technol.*, 2006, **55**, (3), pp. 839–847

12 Fattah, H., Leung, C.: 'An overview of scheduling algorithms in wireless multimedia networks', *IEEE Wirel. Commun.*, 2002, **9**, (5), pp. 76–83

13 Wang, Y., Xu, S., Liu, F., Wang, X., Qian, Y., Wang, Y.: 'A QoS-oriented cross-layer packet scheduling algorithm for a downlink wireless OFDMA system'. IET Conf Wireless, Mobile and Sensor Networks, 2007. (CCWMSN07), 12–14 December 2007, pp. 793–796

14 Hu, H., Chen, H.H., Guo, K., Weckerle, M.: 'Cross-layer adaptive resource allocation for OFDM systems with hybrid smart antennas', *IET Commun.*, 2007, **1**, (5), pp. 831–837

15 Navaie, K.: 'Cross-layer resource allocation in orthogonal frequency multiple access systems based on channel distribution information', *IET Commun.*, 2013, **7**, (5), pp. 439–447

16 Mohanram, C., Bhashyam, S.: 'Joint subcarrier and power allocation in channel-aware queue-aware scheduling for multiuser OFDM', *IEEE Trans. Wirel. Commun.*, 2007, **6**, (9), pp. 3208–3213

17 'IEEE Standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1', IEEE Std. 802.16e-2005 and IEEE Std 802.16-2004/Cor1-2005, February 2005

18 Andrews, J., Ghosh, A., Muhamed, R.: 'Fundamentals of wimax: understanding broadband wireless networking' (Prentice-Hall PTR, Upper Saddle River, NJ, USA, 2007)

19 Vuswabatg, P., Tse, D., Laroia, R.: 'Opportunistic beamforming using dumb antennas', *IEEE Trans. Inf. Theory*, 2002, **48**, (6), pp. 1277–1294

20 Rappaport, T.S.: 'Wireless communications: principles and practice' (Prentice-Hall, Upper Saddle River, NJ, USA, 2002)

doi: 10.1049/iet-com.2014.0090