Memis *et al. EURASIP Journal on Wireless Communications and Networking* (2015) 2015:217
DOI 10.1186/s13638-015-0446-7

EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Resource allocation for statistical QoS guarantees in MIMO cellular networks

Mehmet Ozerk Memis[1], Ozgur Ercetin[1], Ozgur Gurbuz[1*] and Seyed Vahid Azhari[2]

Abstract

This work considers the performance of the downlink channel of MIMO cellular networks serving multiple users with different statistical QoS requirements. The paper proposes resource allocation algorithms that aim to optimize the system performance over the sum of the optimal user utility functions by employing the effective capacity theory. Proportionally fair resource allocation among the users is performed via two different approaches and solutions, namely, the Frame Allocation Algorithm (FAA), which involves dynamic time allocation for transmit beamforming, and the Power Allocation Algorithm (PAA), which provides optimal power control for space division multiple access. In FAA, each user is assigned a distinct slot of optimal length, based on the instantaneous channel conditions of the active users in each frame; while in PAA, resource allocation is performed via power assignment by taking into account the long-term averages of the channel conditions across all users. The efficacy of the proposed algorithms are demonstrated via numerical experiments considering realistic channel models and various QoS settings.

Keywords: Effective capacity; MIMO systems; Spatial multiplexing; Multi-user MIMO

1 Introduction

Last-mile connections to end-users are becoming predominantly wireless. In order to deliver the same performance to end-users as if they are connected to a wired network, new techniques to maximize the throughput in all-wireless networks must be developed. One of the most promising approaches in achieving this is the use of multiple-input multiple-output, or MIMO, technology [1].

In MIMO, both the transmitter and receiver are equipped with multiple antenna elements, where each antenna pair provides an independent spatial path between the transmitter and receiver. A MIMO stream is basically a spatial communication channel that is obtained by cooperative coding of multi-antennas on both transmitter and receiver sides, resulting in the spatial degrees-of-freedom (DoF) of the MIMO channel. System capacity scales linearly with the number of antenna elements and characterizing the MIMO channel in terms of its DoF allows better resource allocation in terms of time slot and power to different links. In fact, implementing MIMO technology in wireless networks specifically requires cross layer design, managing the interaction between the physical (PHY) and medium access control (MAC) layers.

There is a plethora of work on cross-layer resource optimization in wireless systems. These works illustrate that significant throughput gain can be obtained by joint optimization of radio resources across PHY and MAC layers, where a typical assumption is that the transmitter has an infinite backlog and the information flow is delay insensitive. However, in practice, it is very important to consider random bursty arrivals and delay performance metrics in addition to the conventional PHY layer performance metrics in cross-layer optimization. In addition, quality of service (QoS) requirements imposed by the higher layers and time-sensitive applications must be taken into account together with resource optimization.

In order to achieve efficient wireless communications while supporting diverse delay QoS requirements, the effective capacity concept can be utilized [2–4]. The effective capacity has been initially defined in [5] to evaluate the capability of a wireless service process in supporting data transmission subject to a statistical delay QoS requirement metric, called QoS exponent and denoted

*Correspondence: ogurbuz@sabanciuniv.edu
[1] Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey
Full list of author information is available at the end of the article

by θ. The higher θ corresponds to the more stringent delay constraint. Also, θ can continuously vary from 0 to ∞, and thus a wide spectrum of QoS constraints can be readily characterized by a general model. However, incorporating the effective capacity model into multi-user communications faces significant challenges, which are not encountered in a single user wireless link [4, 6–10]. Multi-user systems often have to dynamically allocate the wireless resources based on mobile users' channel state information (CSI), and they usually need to balance the performances among all mobile users according to users' diverse QoS requirements. Exploiting the physical characteristics and flexibility of MIMO, while satisfying individual users' QoS requirements remains to be the main challenge of the latest generation wireless networks.

In this paper, we focus on downlink multi-user QoS provisioning via dynamic resource allocation in MIMO cellular networks considering two scenarios considering whether CSI is fully available or not. For these scenarios, we propose effective capacity-based resource allocation schemes, considering MIMO users receiving delay-sensitive data streams from a base station (BS) over time-varying wireless channels. In the first scenario, the BS has access to the instantaneous CSI of users, assuming quasi-static channels. Based on this information, the BS can perform resource allocation per slot, so that an auxiliary optimization problem is solved to obtain the time-shares per user channel, given QoS requirements and CSI, resulting in Frame Allocation Algorithm (FAA). In the FAA scheme, an interference-free time division multiple access (TDMA) based model is assumed, where only one user transmits at a time and the BS acquires the CSI from each user and determines how long each user receives service within a time frame (the shortest duration of time interval in which users' CSI remain constant). For the second scenario, we consider wireless systems where obtaining the instantaneous CSI is challenging or costly, and the BS has to rely only on the average statistical CSI per user channel to perform resource allocation. Hence, in the second resource allocation scheme, namely Power Allocation Algorithm (PAA), we chose to investigate the case of successive interference cancellation assuming a Space Division Multiple Access (SDMA) system, where all users are served simultaneously using superposition precoding. By PAA, the transmission power of each user stream is determined based on the average CSI.

In order to obtain the resource allocation solutions in both approaches, we first model the effective capacity of MIMO links by explicitly considering multi-user scheduling and resource allocation together. Based on the effective capacity model, we formulate each resource allocation problem as a network utility maximization (NUM) problem with each user having potentially a different quality of service requirement. The solutions are obtained numerically under realistic channel models, and the efficacy of the dynamic algorithms are demonstrated by numerical experiments. Summarizing, our main contributions can be listed as follows:

1. Two new effective capacity-based resource allocation schemes are proposed for providing QoS guarantees over MIMO networks.
2. An optimal opportunistic downlink scheduling scheme, namely Frame Allocation Algorithm (FAA) is proposed for TDMA based MIMO systems employing spatial multiplexing. This scheme utilizes the vector of DoFs for all links to properly allocate the right amount of time slots to each user on the downlink, such that a concave utility function of the effective capacity is maximized for all links.
3. A relatively simple formulation for effective capacity is provided, so that the transmitter and receiver only have to keep track of the statistics of the channel in terms of its DoF, as opposed to complex gain matrices.
4. An optimal downlink power allocation scheme, Power Allocation Algorithm (PAA is proposed for multi-user MIMO systems, such that each user is assured its promised level of statistical QoS guarantee, while a certain utility function of all link effective capacities is maximized.
5. Both time-slot and power allocation problems are formulated as convex optimization problems based on effective capacity of the MIMO system, both providing statistical QoS guarantees in terms of the delay violation probability.

The rest of the paper is organized as follows: In Section 2, we present a summary of the related work and background on effective capacity. Section 3 presents our system model, followed by the proposed resource allocation schemes, FAA and PAA in Sections 4 and 5. P Performance Analysis results are presented in Section 6, and our conclusions are summarized in Section 7.

2 Related work and background

Exploiting the characteristics of MIMO technology in the physical layer and translating its performance gains to higher layers has motivated integrated, cross-layer approaches [11]. The main challenge in the integration of the MIMO models to the higher layers is the accurate computation of the MIMO channel capacity, which requires complex matrix operations and methods that do not provide closed-form solutions. In order to overcome this difficulty and enable cross-layer designs, closed-form, simpler, and accurate uncorrelated MIMO channel capacity computation methods have been proposed

in the literature [12]. Capacity computation methods that employ K-state Markov models and Gilbert-Elliot (GE) channel models for correlated MIMO channels are presented in [10, 13]. In both approaches, the MIMO channels are defined via their DoF, i.e., independent signaling dimensions, which facilitate the design of cross-layer resource allocation algorithms.

Cross-layer MIMO resource management has been previously studied for investigating the effect of MIMO operation at PHY, MAC, and higher layers. Works such as [14–16] address admission control, while, e.g., [17] investigates routing along with power control and scheduling. Optimal scheduling policies based on the stream-based structure, considering the trade off between spatial multiplexing and diversity have been proposed in [18]. In [19], a TDMA-based interference aware transmission scheduling [20] is used to resolve contention problems in wireless networks by considering spatial DoF. Opportunistic scheduling policies have also been explored for the channel optimization of MIMO-based wireless networks, where the main goal is to maximize system capacity [21, 22].

Maximizing the system capacity while satisfying QoS requirements of different user applications has been the main challenge of 4G systems. Despite some works, such as [23], which consider different fairness criteria, there are a limited number of studies on scheduling in MIMO networks with QoS constraints. All these approaches handle packet scheduling and resource allocation in the MAC layer, while PHY layer performs beamforming and achieves multiuser diversity gain. Yin and Liu [24] proposes such an approach, with a packet prioritizer followed by a resource allocator, which tries to maximize the throughput for a given packet priority order. In addition, in [9] the notion of effective capacity is used to propose a delay bound estimator for LTE downlink on a 2×2 MIMO system. This estimator is integrated into the LTE radio link controller at the link layer and makes use of buffer status information to provide delay distribution estimations. However, CSI is not considered for scheduling, as we have proposed in this paper.

Much research has been devoted to providing complicated closed-form expressions for the effective capacity of wireless channels [6, 7]. Guo et al. [7] derives the effective capacity for MIMO channels under maximal ratio combining and adaptive modulation, by modeling a MIMO channel in terms of its number of DoFs as a Markov chain and conditioning on the number of DoFs, and [25] uses the effective capacity theory to perform optimal power allocation for a group of independent mobile stations in a virtual MIMO system in the uplink direction. In our problem, we consider multi-user MIMO in the downlink, where total available power is limited and we have a larger number of stations communicating as compared to the limit of two in [25]. Cheng et al. [26] proposes power and spectrum efficiency indexes considering point-to-point MIMO links using effective capacity. Despite performing joint power and spectrum allocation, such that a certain statistical QoS guarantee is provided, this approach does not consider the DoF vector of all links during this allocation, as it is done in our FAA scheme. Moreover, [26] does not consider a multi-user MIMO regime, but considers only point-to-point links. Our approaches on the other hand, assume that many users are scheduled on the downlink, via FAA in the TDMA-based regime, or they are handled using multiuser MIMO with superposition coding in the SDMA-based regime via PAA.

PAA can be considered for possible extensions to popular non-orthogonal multiple access (NOMA) techniques, which are considered to be implemented in 5G technology. The major body of research into NOMA schemes is mainly focused on characterizing its capacity or improving it via different power allocation strategies [27–32]. These works, however, lack proper treatment of QoS, as required for real-time traffic. For example, Hojeijet et al. [33] proposes and compares several power allocation schemes for NOMA systems. Furthermore, Timotheou et al. [34] proposes a downlink power allocation scheme for a NOMA system with the objective of providing max–min fairness among links. In [35], the authors relate QoS to the outage probability experienced by a user. None of these works on NOMA consider delay, which is an important QoS metric for real-time traffic.

There are some works, which address QoS in terms of delay, while employing power allocation. For instance, in [36], the authors formulate joint power allocation and link adaptation for satellite links as an optimization problem with the objective of maximizing the total system effective capacity. Mao et al. [37] models uplink power control as a non-cooperative game, where the objective is to maximize effective capacity. In [38], the authors propose a power allocation scheme that optimally allocates average transmission power to different MIMO streams, such that joint power and spectrum efficiency is achieved, while the statistical QoS requirement captured as the constraint on the link's effective capacity is also satisfied at the same time. Despite using similar approaches in power control, none of these works employ NOMA techniques, such as successive interference cancellation, since user signals are treated as regular interference.

The authors of [39] consider a virtual MIMO uplink system, where power allocation is optimally performed among existing and new users, such that the effective capacity of existing users is satisfied, while the new users get the maximum possible effective capacity. In that work, successive interference cancellation is performed but no specific precoding technique is employed, unlike superposition coding used in our work. More importantly, their scheme is not opportunistic and only two and

three user cases with homogeneous QoS requirements are considered. Last but not least, in [40], the authors compare the performance of TDMA and superposition coding when used for maximizing the effective capacity of the system. The capacity region of each access mechanism is derived and then the optimum resource allocation in terms of time, power, and decoding order is determined, such that effective capacity is maximized. This work does not consider a MIMO channel, unlike ours.

To the best of our knowledge, our work is the first that addresses the multiuser MIMO QoS provisioning problem with the effective capacity approach. By this approach, we formulate this problem under two different cases on the availability of the CSI. We propose a time allocation algorithm (FAA) for the case when the instantaneous CSI is available, and a power allocation algorithm (PAA), when only the average CSI is available at the BS.

2.1 Effective capacity theory

Inspired by the effective bandwidth theory [41], which models the asymptotic stochastic behavior of source traffic to a queueing system, in [5], Wu and Negi have developed a dual effective capacity theory in order to analyze the random and time-varying wireless channel under a probabilistic delay constraint. The effective capacity theory tries to figure out the maximum constant arrival rate that can be served by a stationary channel (service) process at a queue, while satisfying a target delay-QoS requirement, such that the delay does not exceed a given bound $D_{\max}$ with probability, $(1-\epsilon)$.

Let $c(\tau)$ represent the instantaneous channel service in terms of bits that can be served from the queue in a finite length slot of τ seconds, the cumulative channel process, $C(t)$, i.e., the aggregate number of bits that can be served in $[0,t]$ can be found as, $C(t) = \int_{\tau=0}^{t} c(d\tau)d\tau$. Considering stationary ergodic arrival and service processes and an average arrival rate of μ, the probability that the delay, $D(t)$ exceeds $D_{\max}$ is calculated as [5]:

$$\epsilon = \sup_{t} \mathbb{P}\{D(t) > D_{\max}\} = \gamma(\mu).e^{-\theta(\mu)D_{\max}}. \quad (1)$$

Here, θ is the *QoS exponent*, which specifies the decay rate of the tail distribution for the delay process; $\gamma(\mu)$ is the probability of the queue being non-empty, and μ is the arrival rate that satisfies the delay violation probability, i.e., $\epsilon = \gamma(\mu)e^{-\theta(\mu)D_{\max}}$.

The effective capacity is defined as [5]:

$$E_C(\theta) = -\frac{\alpha_C(-\theta)}{\theta}, \quad (2)$$

with $\alpha_C(\theta)$ being the Gartner-Ellis limit of the logarithm of the moment generating function (MGF) of the cumulative channel process, $C(t)$,

$$\alpha_C(\theta) = \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}\left[e^{\theta C(t)}\right]. \quad (3)$$

From effective capacity theory, given a QoS-exponent θ^*, the system can support streams with QoS requirements satisfying $\theta \geq \theta^*$ if the arrival rate μ satisfies $\mu \leq E_C(\theta^*)$. Note that a small θ represents loose delay-QoS requirements, while larger exponent implies more stringent QoS, and as $\theta^* \to 0$, the effective capacity converges to the ergodic capacity, whereas for $\theta^* \to \infty$, it converges to the delay-limited capacity.

3 System model

We consider the downlink of a single cell in a MIMO-based wireless network, where BS is deployed with n_t antennas to communicate with multiple receivers each with n_r antennas. We assume a Gaussian broadcast scenario, in which the base station is sending independent messages to L receivers, and the channel gain matrix observed by each receiver l is denoted by H_l, consisting of entries $H_l(j,k) \sim N\left(0, \frac{1}{\sigma_l^2}\right)$. Also, $||\cdot||$ denotes the matrix norm, where $||H_l||^2 = \lambda_{\max}\left(H_l^* H_l\right)$, i.e., the maximum eigenvalue of the random matrix $H_l^* H_l$. N_o is the variance of the Gaussian noise observed. The total transmit power available at the base station is P Watts.

We envision that data can be transmitted from the BS to multiple users over MIMO links employing either of the two schemes: (1) time division multiple access mode with spatial multiplexing, which is interference-free, and (2) space division multiple access mode with superposition coding employed for interference cancellation (i.e., multi-user MIMO). We propose two scheduling and resource allocation algorithms, namely, *Frame Allocation Algorithm* (FAA) and *Power Allocation Algorithm* (PAA) to work TDMA and SDMA transmission modes, respectively.

The proposed algorithms are designed with the aim of maximizing the effective capacity of the downlink channel serving L active users, each with different QoS exponent, θ^l, $l = 1,\ldots,L$. Users obtain a utility which is a concave function of their effective capacity. In this paper, we assume a logarithmic utility function which is shown to achieve proportional fairness among the users [42].

4 Frame allocation algorithm for time division multiple access

The Frame Allocation Algorithm (FAA) exploits spatial multiplexing of MIMO communication, as the BS communicates with one user at a time over a point-to-point

MIMO link in time division multiple access mode, hence, interference is avoided.

A fraction, Φ^l, of the unit time frame is allocated to each user, i.e., each MIMO link, l, so that $\sum_{l=1}^{L} \Phi^l = 1$. Time durations to be allocated for the active users are variable size slots, which are changed dynamically, frame-by-frame, considering the users' QoS constraints and instantaneous channel conditions in terms of the DoF vector for all MIMO links. FAA is the solution of a network utility maximization problem, with user utilities given as functions of their effective capacities. For this purpose, the effective capacity of a single MIMO link, as well as the total effective capacity of the system with FAA need to be derived.

4.1 Channel model

The point-to-point MIMO channel is modeled as a discrete time Markov chain, where each state $i = 1, \ldots, d; (d = \min\{n_t, n_r\})$ represents the number available degrees of freedom (DoF) occurring with probability π_i [13]. For each link l, the total average signal to noise ratio (SNR) is $\bar{\rho}^l = P.\sigma_l^2/N_o$. Given the average SNR of each link, $\bar{\rho}^l$, the discretized Markov channel model can be obtained by considering sufficiently large number of channel realizations, applying singular value decomposition and water filling algorithm [1] for each channel matrix (H_l), marking the number of values exceeding the water level as the available DoF of the link, and then counting the occurrences of the different DoF to obtain the probability of lth link having i DoF, i.e., π_i^l, for all $i = 1, \ldots, d$.

Active users are served once in each frame, which is of unit length normalized with respect to the channel coherence time. Hence, the available DoF and the total average SNR $\bar{\rho}^l$ per link remain constant throughout a frame. Due to fading, however, the available DoF per link can change independently from one frame to another.

The BS is assumed to have full channel state information (CSI), i.e., the Markov characterization of each MIMO link, and instantaneous CSI, which is the currently available DoF for the given time frame. The capacity of a MIMO link l in state i is approximately given as [43],

$$R_i^l = i \cdot \log_2\left(1 + \frac{\bar{\rho}^l}{i}\right) \tag{4}$$

with units bps/Hz. This is a simple but accurate estimation of the ergodic (optimal) MIMO channel capacity obtained after singular value decomposition and water filling [43].

4.2 Effective capacity formulation

In order to calculate the effective capacity of a single MIMO link, we first determine the moment generating function (MGF) of the channel process. Note that the cumulative channel process of each MIMO link can be described as an uncorrelated homogeneous Markov Modulated Process (MMP). For a general MMP, its MGF is given by $\boldsymbol{\pi}(\Gamma(\theta)\mathbf{Q})^{t-1}\Gamma(\theta)\mathbf{1}^T$, where $\boldsymbol{\pi}$ is the steady-state probability vector, $\Gamma(\theta) = \text{diag}\left(e^{\theta R^0}, \ldots, e^{\theta R^d}\right)$ is the rate matrix, $\mathbf{Q}$ is the state transition matrix and $\mathbf{1}$ is the column vector of ones [44]. It follows that the MGF of the point-to-point MIMO channel process is,

$$M_C\left(\theta^l, t\right) = \mathbb{E}\left[e^{\theta^l C(t)}\right] = \sum_{i=0}^{d} \left(\pi_i^l\right)^t e^{\theta^l R_i^l t}, \tag{5}$$

where R_i^l is the transmission rate of MIMO link l when it has i DoF. Note that (5) reduces to the MGF of the ON–OFF traffic source, when a MIMO link has one antenna at both transmitter and receiver sides [44].

Once the MGF of the service process is determined, the Gartner-Ellis limit of *log-MGF* can be calculated according to (3). However, due to the complexity of obtaining a closed-form expression, we use an upper-bound on the *log-MGF* of the channel service process, so that the given QoS constraint is not violated, i.e.,

$$\begin{aligned} \log\left(M_C\left(\theta^l, t\right)\right) &= \log\left(\sum_{i=0}^{d} e^{t\left(\log \pi_i^l + \theta R_i^l\right)}\right) \\ &\leq \log\left((d+1)e^{t \max_i\left\{\log \pi_i^l + \theta R_i^l\right\}}\right). \end{aligned} \tag{6}$$

Then, substituting (6) into (3), we obtain,

$$\alpha_C\left(\theta^l\right) = \max_i\left\{\log \pi_i^l + \theta R_i^l\right\}. \tag{7}$$

Finally, a lower bound for the effective capacity of a single MIMO link across all its DoF is obtained by substituting (7) in (2),

$$E_C^l\left(\theta^l\right) = \min_{i=1,\ldots,d}\left\{R_i^l - \frac{\log \pi_i^l}{\theta^l}\right\}. \tag{8}$$

Figure 1 illustrates the tightness of this lower bound for various number of antennas and SNR levels. Here, the actual effective capacity values as obtained from (6) is compared against our lower bound estimate, expressed by (8) over a range of QoS indices, θ. The difference always stays within 10 % as shown in the plots and it becomes smaller as the QoS index is increased. This means that for more strict QoS requirements, we are less likely to over provision, which is desirable.

In FAA, for each MIMO link, a fraction of time is reserved at each time frame depending on the instantaneous DoF of all links in the system. Let $\delta^l(t)$ be the available DoF in frame t and $\phi^l(t)$ represent the *aver-*

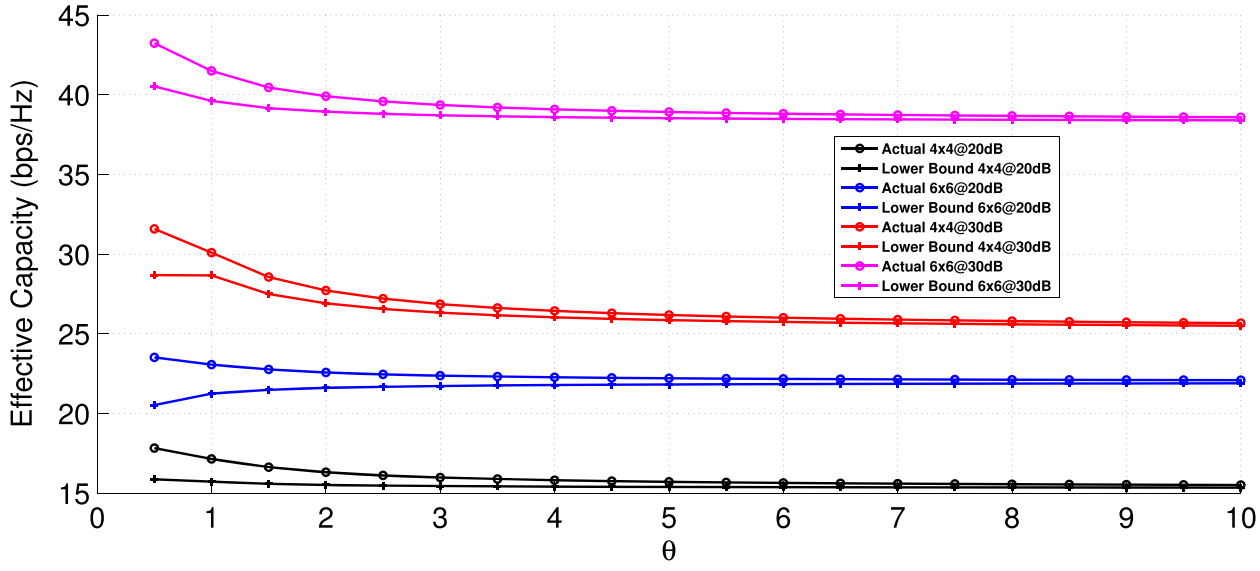

Fig. 1 Comparison of effective capacity lower bound and actual effective capacity

age fraction of frame t reserved for this MIMO link at this DoF. It follows from (4) that the *average* transmission rate of a MIMO link l given $\delta^l(t) = i$ is approximately obtained as

$$R_i^l(t) = i \log_2\left(1+\frac{\bar{\rho}^l}{i}\right)\phi^l(t), \tag{9}$$

where we have reused notation for $R_i^l(t)$. We now calculate ϕ^l as

$$\phi^l = E_{\delta_i^{-l}}\left[\Phi\left(\delta_i^{-l}\right)\right], \tag{10}$$

where $\boldsymbol{\delta}_i^{-l} = (\delta^1,\ldots,\delta^{l-1},i,\delta^{l+1},\ldots,\delta^L)$ is the vector of all link DoFs for which the designated link l has i DOFs. Also, $\Phi(\delta)$ is the frame allocation vector for a specific DoF vector of δ. It follows that ϕ^l can be expressed in terms of channel state probabilities π as

$$\phi^l = \sum_{\forall\delta_i^{-l}}\left(\Phi\left(\delta_i^{-l}\right)\prod_{m=1..L}\pi_{\delta_m}^m\right), \tag{11}$$

where δ_m is the mth element of δ_i^{-l} and the superscript m refers to the link index. Plugging (11) into (9) we obtain

$$R_i^l(t) = i\log_2\left(1+\frac{\bar{\rho}^l}{i}\right)\sum_{\forall\delta_i^{-l}}\left(\Phi\left(\delta_i^{-l}\right)\prod_{m=1..L}\pi_{\delta_m}^m\right). \tag{12}$$

Eventually, by substituting $R_i^l(t)$ in (8) with (12), we arrive at the following expression for the effective capacity of some link l,

$$E_C^l\left(\theta^l\right) = \min_{i=1,\ldots,d}\left\{i\cdot\log\left(1+\frac{\bar{\rho}^l}{i}\right)\sum_{\forall\boldsymbol{\delta}_i^{-l}}\left(\Phi^l\left(\boldsymbol{\delta}_i^{-l}\right)\prod_{m=1..L}\pi_{\delta_m}^m\right)-\frac{\log\pi_i^l}{\theta^l}\right\}. \tag{13}$$

Our objective is to determine $\Phi^l(\boldsymbol{\delta})$ for all $\boldsymbol{\delta}$ such that total system utility is maximized given the channel distributions and user QoS parameters, i.e.,

$$\max_{\Phi^l(\boldsymbol{\delta})}\sum_{l=1}^{L}\log\left[1+E_C^l\left(\theta^l\right)\right]. \tag{14}$$

The optimization problem in (14) is a non-convex optimization problem due to the min operator in the definition of effective capacity. Hence, we modify the problem by adding d additional inequality constraints for each possible DoF for each link arriving at,

$$\max_{\Phi^l(\boldsymbol{\delta})}\sum_{l}\log\left[1+\gamma^l\right] \tag{15}$$

subject to

$$\gamma^l \le i\cdot\log\left(1+\frac{\bar{\rho}^l}{i}\right)\sum_{\boldsymbol{\delta}_i^{-l}}\left(\Phi^l\left(\boldsymbol{\delta}_i^{-l}\right)\prod_{m=1..L}\pi_{\delta_m}^m\right)-\frac{\log\pi_i^l}{\theta^l},\forall l,i \tag{16}$$

$$0\le\Phi^l(\boldsymbol{\delta})\le 1,\forall l \tag{17}$$

$$\sum_{l}\Phi^l(\boldsymbol{\delta})\le 1 \tag{18}$$

where the first set of constraints in (16) are defined $\forall\, l$ and $i = 1, \ldots, d$, and $L{\cdot}d$ constraints are obtained by the decomposition of the effective capacity expressions into their states.

The constraints in (17) and (18) represent the slot durations of the links and the limited resource constraints, defined for all l and $\boldsymbol{\delta}$.

Note that the optimization problem stated in (15)–(18) has $L.d^L$ decision variables embedded in $\Phi^l(\delta)$, and $L(d+1)+1$ constraints. As d or L increases, the number of decision variables grows exponentially, enlarging the search space of the problem. Thus, we introduce a new algorithm, i.e., *dynamic-FAA*, which iteratively solves a simplified version of the *static-FAA* optimization problem (15)–(18) by updating slot allocations and in turn the effective capacity for each link per frame. This significantly reduces the search space of the optimization problem. However, a modified version of the problem is now solved repetitively. Due to the changes in the original problem statement, the slot allocation variables used before become time-dependent. Thus, frame index t is added to the slot allocation variables. Additionally, they are denoted by a tilde ($\tilde{}$) mark to distinguish them from those variables used in static resource allocation. Table 1 provides a complete list of the variables used in dynamic-FAA algorithm.

In dynamic-FAA, we introduce the *instantaneous* frame allocation for link l, i.e., $\tilde{\phi}^l(\tilde{\delta}(t))$, as the new decision variable. We also describe $\tilde{\Phi}^l(\tilde{\delta}(t))$ as the updated slot allocation for link l based on the current DoF vector $\tilde{\delta}(t)$. The natural outcome of iteratively updating frame allocations is the dynamic update of effective capacity for each link. In each frame t with $\tilde{\delta}^l(t) = i_l$, the algorithm computes each utility function by taking the logarithm of an auxiliary function $\upsilon^l = R^l_{i_l} - \frac{log\pi^l_{i_l}}{\theta^l}$, which is obtained by decomposing effective capacity. With this approach, depending on the instantaneous available DoF of a link, the effective capacity of each link is updated per frame, and the number of constraints obtained by the decomposition reduces from d to 1. With these changes, the optimization problem for each frame t reduces to,

$$\max_{\tilde{\phi}^l(\tilde{\delta}(t))} \sum_l \log[\,1+\upsilon^l] \quad \text{subject to} \tag{19}$$

$$\upsilon^l \le i_l \log\left(1+\frac{\bar{\rho}^l}{i_l}\right)\sum_{\boldsymbol{\delta}^{-l}_{i_l}}\left(\left\{\widetilde{\Phi}^l\left(\boldsymbol{\delta}^{-l}_{i_l}\backslash\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right)+\widetilde{\Phi}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right)\right\}\prod_m \pi^m_{\delta_m}\right)-\frac{\log\pi^l_{i_l}}{\theta^l}, \forall l \tag{20}$$

$$\widetilde{\Phi}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right) = \alpha\left[\widetilde{\Phi}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t^-})}\right)\right]+(1-\alpha)\left[\widetilde{\phi}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right)\right], \forall l \tag{21}$$

$$0 \le \tilde{\phi}^l\left(\tilde{\boldsymbol{\delta}}(t)\right) \le 1, \forall l \tag{22}$$

$$\sum_l \tilde{\phi}^l\left(\tilde{\boldsymbol{\delta}}(t)\right) \le 1 \tag{23}$$

The formulation of (19)–(23) has a few subtle differences with that of (15)–(18). In particular, note that $\widetilde{\Phi}^l\left(\boldsymbol{\delta}^{-l}_{i_l}\backslash\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right)$ denotes *average* slot allocations for link l for all DoF vectors $\boldsymbol{\delta}^{-l}_{i_l}$ up to but not including the current frame DoF vector $\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}$ and it is therefore a known value. In addition, $\widetilde{\Phi}^l(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})})$ is the updated average slot allocation for link l for the current DoF vector $\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}$. This is reflected in (21), where t^- is the last frame index at which the current DoF was encountered, and α is a constant between 0 and 1 for implementing the moving average. Note that, $\widetilde{\Phi}^l(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t^-})})$, which is the last updated value of average slot allocation for the current DoF vector is also a known value at frame t. It follows that the optimization problem presented by (19)–(23) has a total of L decision variables $\tilde{\phi}^l\left(\tilde{\boldsymbol{\delta}}(t)\right)$ and $2L+1$ constraints per frame.

Algorithm 1 outlines our proposed dynamic time slot allocation algorithm, (dynamic-FAA). Table 1 displays all the variables used in the algorithm. The algorithm starts with some arbitrary (e.g., equal) time slot allocation $\widetilde{\Phi}^l(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})})$ and iteratively improves total system utility function in a while loop. The while loop iterates over consecutive frames, and for each frame t corresponding to a DoF vector $\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}$, the reduced convex problem presented in (19)–(23) is solved

Table 1 Parameters used in dynamic-FAA

Parameter	Description
L	Total number of MIMO links
l	Link index
i	DoF index
d	DoF of MIMO link, i.e. $d = \min\{n_t, n_r\}$
θ^l	QoS of lth link
$\boldsymbol{\xi^k}$	kth vector of DoFs of L links, i.e., $\boldsymbol{\xi^1} = (1, ..., 1)$ and $\boldsymbol{\xi^{d^L}} = (d, ..., d)$
$\delta^l(t)$	DoF of lth link at time t, i.e., $\delta^l(t) = i \in \{1, ..., d\}$
$\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}$	Vector of DoFs of all links, i.e., $\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})} = (\delta^1(t), \ldots, \delta^L(t)) \in \{\boldsymbol{\xi^1}, ..., \boldsymbol{\xi^{d^L}}\}$
$\boldsymbol{\delta}^{-l}_i$	DoF vector with lth link having i DoF, i.e., $\boldsymbol{\delta}^{-l}_i = (\delta^1, \ldots, \delta^{l-1}, i, \delta^{l+1}, \ldots, \delta^L)$
$\widetilde{\phi}^l(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})})$	Instantaneous slot allocation for lth link
$\widetilde{\Phi}^l(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})})$	Updated slot allocation for lth link
π^l_i	Probability of lth MIMO link having i DoF
$\bar{\rho}^l$	Average transmit SNR of lth MIMO link, i.e., $P\sigma^{l^2}_{ij}/\sigma^2_n$
α	Exponential moving average weight used in time-slot update
υ^l	Auxiliary function $R^l_i - \frac{\log\pi^l_i}{\theta^l}$
$\widetilde{E}^l_C(\theta^l)$	Updated effective capacity for link l
$\Psi(t)$	System utility function, i.e., $\sum_l \log(1+\widetilde{E}^l_C(\theta^l))$
$\mathbf{0}$	Vector of zeros
ε	Halt condition

Algorithm 1: Dynamic Frame Allocation Algorithm (FAA) for TDMA System

1 **Input**: $L, d, \bar{\rho}^l, \sigma_{ij}^{l\,2}, \sigma_n^2, \alpha_{(t)}, \theta^l, \varepsilon$

2 **while** $|\Psi_{(t)} - \Psi_{(t-1)}| > \varepsilon$ **do**

3 $t \leftarrow$ current frame index

4 $\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})} \leftarrow$ current DoF vector

5 $i_l \leftarrow \widetilde{\delta}^l{}_{(t)}$

6 $t^- \leftarrow$ last frame index with same DoF vector $\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}$

7 7.1 **Solve** $\max_{\widetilde{\boldsymbol{\phi}}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right)} \sum_l \mathbf{log}\left[\mathbf{1}+\boldsymbol{\upsilon}^{\boldsymbol{l}}\right]$

7.2 subject to

7.3 $\upsilon^l \le i_l \log\left(1+\frac{\bar{\rho}^l}{i_l}\right) \sum_{\boldsymbol{\delta}_{i_l}^{-l}} \left(\left\{\widetilde{\Phi}^l\left(\boldsymbol{\delta}_{i_l}^{-l} \backslash \widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right)\right.\right.$

8 $\left.\left. +\widetilde{\Phi}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right)\right\} \prod_m \pi_{\delta_m}^m\right) - \frac{\log \pi_{i_l}^l}{\theta^l}$

7.4 $\widetilde{\Phi}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right) \leftarrow \alpha\left[\widetilde{\Phi}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t}^-)}\right)\right] + (1-\alpha)\left[\widetilde{\phi}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right)\right]$

7.5 $0 \le \widetilde{\phi}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right) \le 1$

7.6 $\sum_l \widetilde{\phi}^l\left(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})}\right) \le 1$

9 **for** $l = 1$ **to** L **do**

10 Update $\widetilde{E}_C^l\left(\theta^l\right)$ as in Eq (13)

11 **end**

12 $\Psi(t) \leftarrow \sum_l \log\left(1+\widetilde{E}_C^l\left(\theta^l\right)\right)$

13 $t \leftarrow t+1$

14 **end**

15 obj = $\max \sum_l \log\left(1+\widetilde{E}_C^l\left(\theta^l\right)\right)$

(lines 7.1 through 7.6). The solution provides current time slot allocations, $\widetilde{\phi}^l(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})})$, for that particular frame, which also updates the average time slot allocation $\widetilde{\Phi}^l(\widetilde{\boldsymbol{\delta}}_{(\boldsymbol{t})})$ through an exponential moving average (line 7.4). Therefore, the algorithm provides better time slot allocations as time moves forward.

The iterative algorithm will stop after enough DoF vectors are experienced by the system, so that system utility function converges to the optimal value. This is achieved by monitoring the incremental improvements in total system utility function in every iteration of the reduced optimization problem. From that point on, the vector of time slot allocations for each DoF vector is already computed by the scheduler and only table look-up will be used to allocate time slots to each link based on the DoF configuration of all links.

5 Power allocation algorithm for space division multiple access

In the previous section, we have investigated the dynamic allocation of time slot lengths among the users based on the instantaneous CSI feedback acquired from the users. However, in multi-user systems, it is well known that the acquisition of instantaneous CSI introduces significant overhead to system operations. For example, in code division multiple access (CDMA)/High Data Rate (HDR) system, the SNR of each link is measured, from which a value representing the maximum data rate that can be supported is determined. This information is then sent back to the BS via the reverse link data rate request channel (DRC). According to CDMA/HDR specifications, the channel state information is 4 bits long and it is updated every 1.67 ms. If there are 25 users in a cell, 100 bits of channel information has to be sent every 1.67 ms. This requires 60 kbps of channel rate to be dedicated only for CSI feedback. The overhead of acquiring CSI is twice the minimum data rate and is approximately more than 20 % of the average transmission rate as specified by CDMA/HDR specification. Clearly, in a MIMO system, this overhead is expected to be significantly higher.

In this section, we investigate the case when instantaneous CSI is not available at the BS, so the resource allocation is based only on the average channel distributions. Note that in this case, water-filling cannot be used and we employ an equal power split across MIMO streams. For this purpose, we consider SDMA system with *superposition coding* in order to simultaneously serve multiple users as we investigate the static allocation of power resource among the users based on their channel statistics and QoS requirements.

5.1 Superposition coding and channel model

In the context of MIMO fading channels, superposition coding together with rate and power allocation has been applied to maximize the average transmission rate [45]. In superposition coding, the encoder constructs the signals in a nested fashion in which the code-word that is intended for a certain receiver is a "satellite" of the code-word that is intended for the next more degraded receiver.

Let us first consider the two receiver case, and a scenario, where the signal observed by receiver 2 is more degraded than that observed by receiver 1. The transmitter wishes to communicate two independent messages simultaneously to both receivers. To do so, the transmitter synthesizes the signal, X, by superimposing the signal V, which contains the message intended for receiver 1 on the signal U, which contains the message intended for receiver 2. The signal U is typically visualized as the center of a cluster of code-words and is chosen from a code-book with rate R^2. In each cluster, there are $(2)^{nR^1}$ satellites centered around U, where n is the length of the code-word and R^1 is the rate of the code-book used for receiver 1. For Gaussian channels, when the transmit power budget is P, it was shown that the capacity achieving code-books are independent and Gaussian and that the average powers with which these code-books are transmitted are $(1-\beta).P$ and $\beta.P$, where $\beta \in [0,1]$ is a partition of power among code-books.

The decoding of superposition encoded signals is as follows. The Gaussian signal V contains the message intended for receiver 1. When operating at the boundary of the capacity region, this signal is not decodable by receiver 2, and hence receiver 2 sees it as additive Gaussian noise. Thus from

receiver 2's perspective, the situation resembles an additive white Gaussian noise (AWGN) channel with signal power $\beta||H_2||^2P$ and noise variance $N_0+(1-\beta)||H_2||^2P$. For receiver 2 to decode the signal U, the rate R^2 must satisfy

$$R^2 \leq \log\left(1+\frac{\beta P||H_2||^2}{(1-\beta)P||H_2||^2+N_0}\right). \tag{24}$$

Since receiver 1 observes a channel that is less degraded than the channel observed by receiver 2, it can decode the signal U and subtract it from its received signal. Having done that, receiver 1 has a signal of power $(1-\beta)||H_1||^2P$, and noise variance N_0. Similarly, receiver 1 can correctly decode signal V, if

$$R^1 \leq \log\left(1+\frac{(1-\beta)P||H_1||^2}{N_0}\right). \tag{25}$$

For the BS to send independent messages to $L>2$ receivers, it generates L independent Gaussian code-books, one for each degradation level. The transmitter superimposes L code-words, one from each code-book, to generate the transmitted signal. The transmitted signal can be regarded as a code-word from nested clusters. Each code-book represents a set of cluster centers that are decodable by the receiver at the corresponding degradation level as well as less-degraded receivers. For more-degraded receivers, these cluster centers are observed as undecodable satellites that contribute to the total noise observed by these receivers. Let ψ^l denote the particular degradation level of receiver l. The receivers at degradation levels $k<\psi^l$ are considered as less-degraded receivers.

As code-words are transmitted from the nested clusters, the transmitter partitions its power, and in order to decode superposition-coded messages, each receiver begins by decoding and subtracting the signals intended for more-degraded receivers. Treating the signals intended for less-degraded receivers as additive Gaussian noise, each receiver then proceeds to decode its intended signal.

Given a power partition $\boldsymbol{\beta}=(\beta^1,\ldots,\beta^L)$, and degradation levels ψ^l, for all $l=1,\ldots,L$, the lth receiver is able to decode its intended signal, if the rate of the corresponding code-book satisfies:

$$R^l(\boldsymbol{\beta}) \leq \log\left(1+\frac{\beta^l P||H_l||^2}{\sum_{j=1}^{L}\mathcal{I}_{\psi^j<\psi^l}\beta^j P||H_l||^2+N_0}\right), \tag{26}$$

where β^l is the partition of power allocated for user l, and $\mathcal{I}_{x<y}$ is an indicator function which takes value 1 when $x<y$, and 0 otherwise.

5.2 Effective capacity formulation

In order to determine the effective capacity of the channel process per MIMO link with superposition coding, we first need to calculate the MGF of the rate of each link under a given power partition β^l, $l=1,\ldots,L$. The instantaneous channel bit rate R^l for some link l, which is allocated a fraction, β^l, of the total BS transmit power, P, will be the sum of instantaneous rates of $d=\min\{n_r,n_t\}$ independent parallel MIMO streams among which the transmit power is symmetrically partitioned [46], i.e.,

$$R^l=\sum_{i=1..d}\log_2\left[1+\frac{P\beta^l d^{-1}\lambda_i^l}{\left(\sigma_{ij}^l\right)^2 P\sum_{k=1..L}I_{\psi^k<\psi^l}\beta^k+\sigma_n^2}\right], \tag{27}$$

where $(\sigma_{ij}^l)^2$ is the variance of the MIMO channel gain matrix entries, σ_n^2 represents the Gaussian noise present in the medium, and λ_i^l is the ith eigenvalue of the MIMO channel gain matrix for link l . Moreover, ψ^k is the encoding order of link k for superposition coding and $I_{\psi^k<\psi^l}$ is simply an indicator function that counts those links interfering with l (meaning they are encoded before this link.) For ease of use, we denote the common term in (27) as

$$\zeta^l(\boldsymbol{\beta},\boldsymbol{\Psi})=\frac{P\beta^l d^{-1}}{\left(\sigma_{ij}^l\right)^2 P\sum_{k=1..L}I_{\psi^k<\psi^l}\beta^k+\sigma_n^2}, \tag{28}$$

and rewrite R^l as

$$R^l=\sum_{i=1}^{d}\log_2\left[1+\zeta^l(\boldsymbol{\beta},\boldsymbol{\Psi})\lambda_i^l\right]. \tag{29}$$

For a given QoS parameter θ^l, MGF of rate R^l is expressed by

$$M_C^l(\theta^l,t)=\log\mathbb{E}\left[e^{-\theta^l R^l(\boldsymbol{\beta},\boldsymbol{\Psi})}\right], \tag{30}$$

where the expectation is over random MIMO link realizations.

However, determining the effective capacity for the instantaneous channel rate given by (29) does not result in a closed form solution. Therefore, we use the central limit theorem (CLT) to estimate it. This follows from an approach which is also pursued in [47] and which we have also studied in [48].

The instantaneous channel rate R^l is in fact the sum of d functions of random variables λ_i^l (i.e., channel eigenvalues) with $\mathcal{X}_i^l=\log_2\left[1+\zeta^l(\boldsymbol{\beta},\boldsymbol{\Psi})\lambda_i^l\right]$,

$$R^l=\sum_{i=1}^{d}\mathcal{X}_i^l. \tag{31}$$

We follow the approach of [49] to estimate the effective capacity. The effective capacity under a given power partition $\boldsymbol{\beta}$, for MIMO link $l=1,\ldots,L$, $E_C^l(\theta^l)$ is again determined according to (2). However, in this case, no closed form expression exists for general channel models. Hence, we characterize $\mathcal{X}_i^l$'s in terms of their means and variances.

With no time-correlation among samples, the accumulated channel process for link l, i.e., $C^l(t)$, is simply the addition of t uncorrelated and iid random variables. Expressing the instantaneous channel rate R^l by $c^l(\tau)=\sum_{i=1}^{d}\mathcal{X}_i^l(\tau)$, the cumulative random variable $C^l(t)$ can be expressed by,

$$C^l(t)=\sum_{\tau=0}^{t}c^l(\tau)=\sum_{\tau=0}^{t}\sum_{i=1}^{d}\mathcal{X}_i^l(\tau) \tag{32}$$

As $t\to\infty$, CLT can be applied and $C^l(t)$ can be considered as a Gaussian random variable with mean $\mu_{C^l}=t\mu_{c^l}$ and variance $\sigma_{C^l}^2=t\sigma_{c^l}^2$.

The use of this theorem enables us to express both the mean and the variance of the instantaneous channel rate as $\mu_{c^l} = \sum_{i=1}^{d} \mu_{\mathcal{X}_i^l}$ and $\sigma_{c^l}^2 = \sum_{i=1}^{d} \sigma_{\mathcal{X}_i^l}^2$, and the statistics of the accumulated channel process as $\mu_{C^l} = t\sum_{i=0}^{d} \mu_{\mathcal{X}_i^l}$ and $\sigma_{C^l}^2 = t\sum_{i=0}^{d} \sigma_{\mathcal{X}_i^l}^2$. Note that these statistics are functions of $\zeta^l(\boldsymbol{\beta}, \boldsymbol{\Psi})$, which is itself a function of power allocation vector $\boldsymbol{\beta}$ and encoding order $\boldsymbol{\Psi}$.

Finally, the effective capacity expression [49] for the resulting Gaussian random process $C^l(t)$ is given by

$$E_C^l\left(\theta^l, \zeta^l(\boldsymbol{\beta}, \boldsymbol{\Psi})\right) = \mu_{c^l}\left(\zeta^l(\boldsymbol{\beta}, \boldsymbol{\Psi})\right) - \frac{\theta^l}{2}\sigma_{c^l}^2\left(\zeta^l(\boldsymbol{\beta}, \boldsymbol{\Psi})\right), \tag{33}$$

where the dependence of effective capacity on power allocation ($\boldsymbol{\beta}$) and link encoding order ($\boldsymbol{\Psi}$) is clear.

Our objective is again to maximize total system utility, which is a function of individual effective capacities obtained by optimally partitioning of the transmit power of the base station and at the same time selecting the optimal encoding order among links for performing superposition coding, i.e.,

$$\max_{\boldsymbol{\beta}, \boldsymbol{\Psi}} \quad \sum_{l=1}^{N} \log\left[1 + E_C^l(\theta, \zeta^l(\boldsymbol{\beta}, \boldsymbol{\Psi}))\right] \tag{34}$$

subject to

$$E_C^l\left(\theta^l, \zeta^l(\boldsymbol{\beta}, \boldsymbol{\Psi})\right) = \mu_{c^l}\left(\zeta^l(\boldsymbol{\beta}, \boldsymbol{\Psi})\right) - \tfrac{\theta^l}{2}\sigma_{c^l}^2\left(\zeta^l(\boldsymbol{\beta}, \boldsymbol{\Psi})\right) \tag{35}$$

$$\sum_l \beta^l \leq 1 \tag{36}$$

This is a very difficult problem, since the solution space consists of all the different $L!$ encoding orders. Hence, at this point, we have used a simple heuristic to fix the encoding order and solve for the other remaining variables, in particular, the power allocation fraction ($\boldsymbol{\beta}$). Our heuristic for determining a good encoding order is to first neglect superposition coding and look at a link as if it were allocated the entire power budget. That is, the higher effective capacity a MIMO link has under full power budget and no interference assumption from neighboring MIMO links, the earlier it is encoded by the BS.

As a result, the effective capacity of a link with encoding order o becomes only a function of power allocation vector $\boldsymbol{\beta}$ and it is denoted by $\gamma^o(\boldsymbol{\beta}) = E_C^l(\theta^l, \zeta^l(\boldsymbol{\beta}))$. In this regard, we aim to solve the following optimization problem:

$$\max_{\boldsymbol{\beta}} \ \sum_{l=1}^{N} \log\left[1 + \gamma^o(\boldsymbol{\beta})\right] \tag{37}$$

$$\sum_l \beta^l \leq 1$$

Since a closed form expression for effective capacity does not exist for this case, this problem can be solved numerically for a given channel characterization. The solution algorithm, named as the Power Allocation Algorithm (PAA) is described in Algorithm 2. Table 2 displays all the parameters used in the algorithm.

The first loop in the PAA Algorithm (steps 3–7) applies CLT to estimate the effective capacities without considering interference from superposition coding (i.e., $I_{\psi^k < \psi^l} = 0, \forall k, l$). This is later used to derive the encoding order based on our heuristic in line (8). Using the encoding order, we now reapply CLT to estimate effective capacities with the effect of interference.

Table 2 Parameters used in PAA

Parameter	Description
o	Encoding order index
P	Transmit power budget
$\sigma_{ij}^{l^2}$	Variance of the channel entries of lth link
σ_n^2	Gaussian noise present in the medium
$\boldsymbol{\beta}$	Power allocation vector, i.e. $\boldsymbol{\beta} = (\beta^1, \ldots, \beta^l, \ldots, \beta^L)$,
$\boldsymbol{\Psi}$	Degradation level vector, i.e. $\boldsymbol{\Psi} = (\psi^1, \ldots, \psi^l, \ldots, \psi^L)$
$\mathcal{I}_{x<y}$	Indicator function that takes value 1 when $x < y$
ζ^l	Auxiliary variable $\zeta^l(\boldsymbol{\beta}) = \frac{d^{-1}P\beta^l}{\sigma_{ij}^{l^2}\sum_{k=1}^{L}\mathcal{I}_{\psi^k<\psi^l}\beta^k+\sigma_n^2}$
c^o	oth link capacity process $= \sum_{i=1}^{d} \log_2\left[1 + \zeta^o(\boldsymbol{\beta})\lambda_i^o\right]$
γ^o	Effective capacity of a link with encoding order o

This is performed in the second loop (steps 10–15). Note that we have denoted $\zeta^l(\beta, \Psi)$ from (28) with $\zeta^o(\beta)$ indicating that it represents a given encoding order o and no longer has Ψ as a parameter. We then numerically solve the optimization problem using a conventional solver (steps 16–17).

6 Performance analysis

In this section, we analyze and compare the performance of the two proposed resource allocation methods in numerical experiments. In our experiments, we investigate the behavior of the methods with respect to heterogeneity of users' QoS demands and channel conditions, and we analyze how heterogeneous QoS requirements and channel conditions affect the resource allocation decisions in MIMO cellular networks. Specifically, we consider a small network, since having a large number of users would have obscured the effects of varying channel conditions and QoS requirements on resource allocation.

In our numerical studies, we consider a cellular downlink MIMO network where there are $L = 3$ users receiving service from a base station, as shown in Fig. 2. Both the BS and users have three antenna elements, and thus, the maximum degrees of freedom of MIMO links between the BS and users is $d = 3$. We assume a Gaussian broadcast scenario, in which the BS is sending independent messages to all receivers, and the channel gain matrix observed by each receiver l is denoted by H_l, consisting of entries $H_l(j,k) \sim N\left(0, \frac{1}{\sigma_l^2}\right)$. Total noise normalized transmit power available at the BS is $P = 5$ Watts. The duration of a time slot is one time unit. The users' QoS requirements are indicated by QoS parameter θ^l, $l = 1, 2, 3$.

We performed three experiments for varying channel conditions and QoS parameters. The values of the parameters used in each experiment are included in Table 3. In the first experiment, we consider homogeneous channel conditions and homogeneous user QoS requirements.

Algorithm 2: Power Allocation Algorithm (PAA) in SDMA System

Input: $L, d, P, \sigma_{ij}^{l^2}, \sigma_n^2, CSI, \theta^l$

/*Perform initial estimate of effective capacity ignoring superposition coding order:*/

for $l = 1$ **to** L **do**

  Estimate mean and variance of $\mathcal{X}_i^l$ from channel statistics

  $\mu_{c^l} \leftarrow \sum_i \mu_{\mathcal{X}_i^l}$ and $\sigma_{c^l}^2 \leftarrow \sum_i \sigma_{\mathcal{X}_i^l}^2$

  $E_C^l \leftarrow \left(\mu_{c^l} - \frac{\theta^l}{2}\sigma_{c^l}^2\right)$

end

Form $\boldsymbol{\psi}$ based on the set $\left\{E_C^l\right\}_{l=1}^L$

/*Do in order o of superposition coding:*/

for $o = 1$ **to** L **do**

  $\zeta^o(\boldsymbol{\beta}) \leftarrow \frac{d^{-1}P\beta^o}{\sigma_{ij}^{o^2}P\sum_{k=1}^L \mathcal{I}_{\psi^k<\psi^o}\beta^k+\sigma_n^2}$

  Estimate mean and variance of $\mathcal{X}_o^l$ considering superposition coding order

  $\mu_{c^o}(\zeta^o) \leftarrow \sum_i \mu_{\mathcal{X}_i^o}$ and $\sigma_{c^o}^2(\zeta^o) \leftarrow \sum_i \sigma_{\mathcal{X}_i^o}^2$

  $\gamma^o(\boldsymbol{\beta}) \leftarrow \mu_{c^o}(\zeta^o(\boldsymbol{\beta})) - \frac{\theta^o}{2}\sigma_{c^o}^2(\zeta^o(\boldsymbol{\beta}))$

end

numerically solve:

obj = $\max_{\beta^o} \sum_o \log\left[1+\gamma^o(\boldsymbol{\beta})\right]$

$\sum_o \beta^o \leq 1$

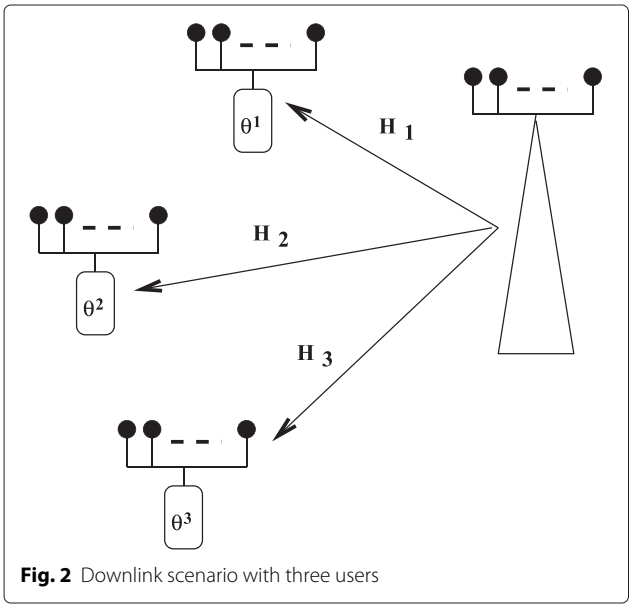

Fig. 2 Downlink scenario with three users

Table 3 The values of parameters used in numerical experiments

Experiment #	Channel gains	QoS guarantees
I	$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.9$	$\theta^1 = \theta^2 = \theta^3 = 0.25$
II	$\sigma_1^2 = 0.3, \sigma_2^2 = 0.6, \sigma_3^2 = 0.9$	$\theta^1 = \theta^2 = \theta^3 = 0.25$
III	$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.6$	$\theta^1 = 0.25, \theta^2 = 0.75, \theta^3 = 1.25$

In the second experiment, we consider heterogeneous channel conditions but homogeneous QoS requirements. Finally, in the third experiment, we consider heterogeneous QoS requirements and homogeneous channel conditions. Hence, with these three experiments, we aim to understand how much effect the channel conditions and QoS requirements have on resource allocation. Note that static and dynamic-FAA are determined as the solution of an optimization problem where a lower bound on the effective capacity is used as the objective function. In the simulations, we provide the *exact* value of the effective capacity of each of the users determined according to the resource allocation decision found as the solution of this optimization problem.

Table 4 summarizes the result of comparing maximum utility gained by solving the original version of the optimal slot allocation problem described by (15) through (18) and referred to as static-FAA with dynamic-FAA, which is outlined in Algorithm 1. The results suggest a performance gap of less than 7 %. The best-case scenario is when channel condition and QoS requirements are homogeneous (experiment I), whereas the worst case in terms of the gap in utility corresponds to the homogeneous channel and non-homogeneous QoS requirement case (experiment III).

Let us now compare dynamic-FAA and PAA. In Fig. 3 for dynamic-FAA, we observe that all users are allocated almost the same slot length in the homogeneous QoS homogeneous channel experiment (No.I), since the channel variance and QoS requirements are the same. Meanwhile, Fig. 4 depicts the PAA results for the same experiment. Under the same conditions, the power levels allocated to each user differ from each other. This is because in superposition coding, each user treats the signals intended for less-degraded receivers as additive Gaussian noise. Even though the transmission powers of users differ significantly, the effective capacities of each

Table 4 Comparison of total utility for static-FAA and dynamic-FAA

Experiment #	Equal time allocation	Static-FAA	Dynamic-FAA	Change %
I	21.0812	21.3954	21.3281	1.49
II	21.0221	21.9185	21.8485	4.26
III	21.0944	22.4313	22.3368	6.33

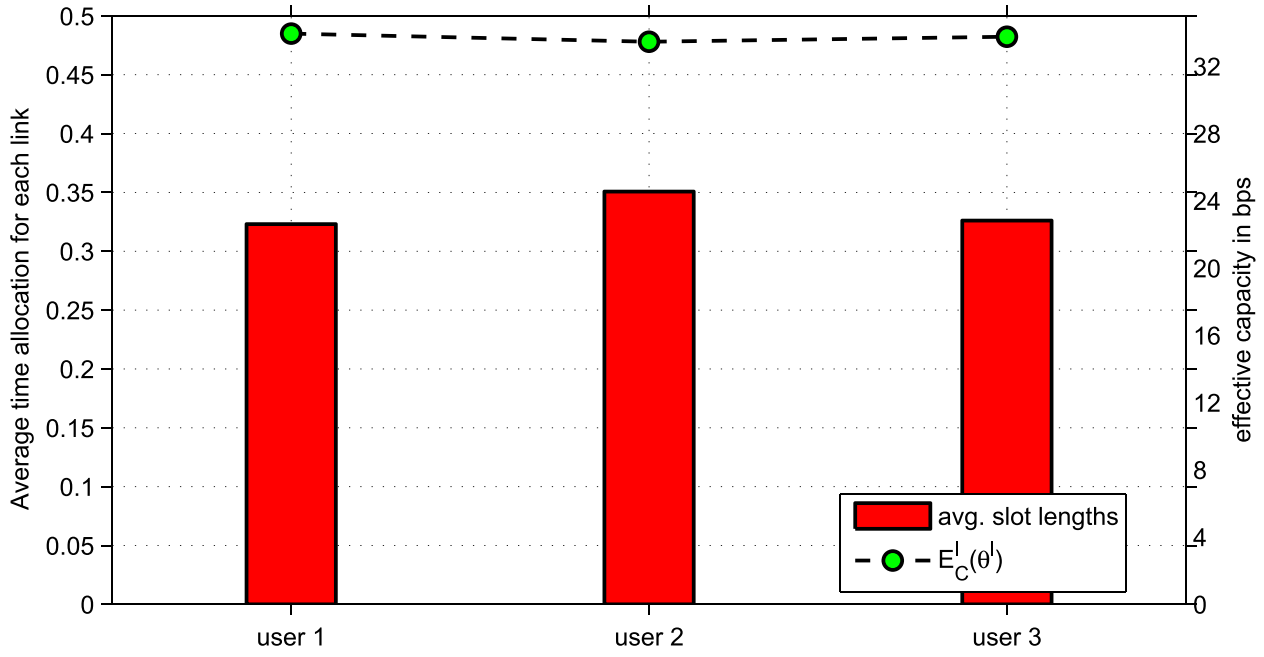

Fig. 3 Resource allocation decisions for experiment I under dynamic-FAA

user is almost the same as expected. Expected total utility under dynamic-FAA and PAA are given in Table 5. In the same table, we also depict the performance of two simple resource allocation policies that do not take into account the channel variance or QoS requirement. The so-called equal time allocation (ETA) algorithm assigns equal slot lengths to each user, and the so-called equal power allocation (EPA) algorithm assigns equal power levels to all users while transmitting signals according to superposition coding. We choose to compare our proposed

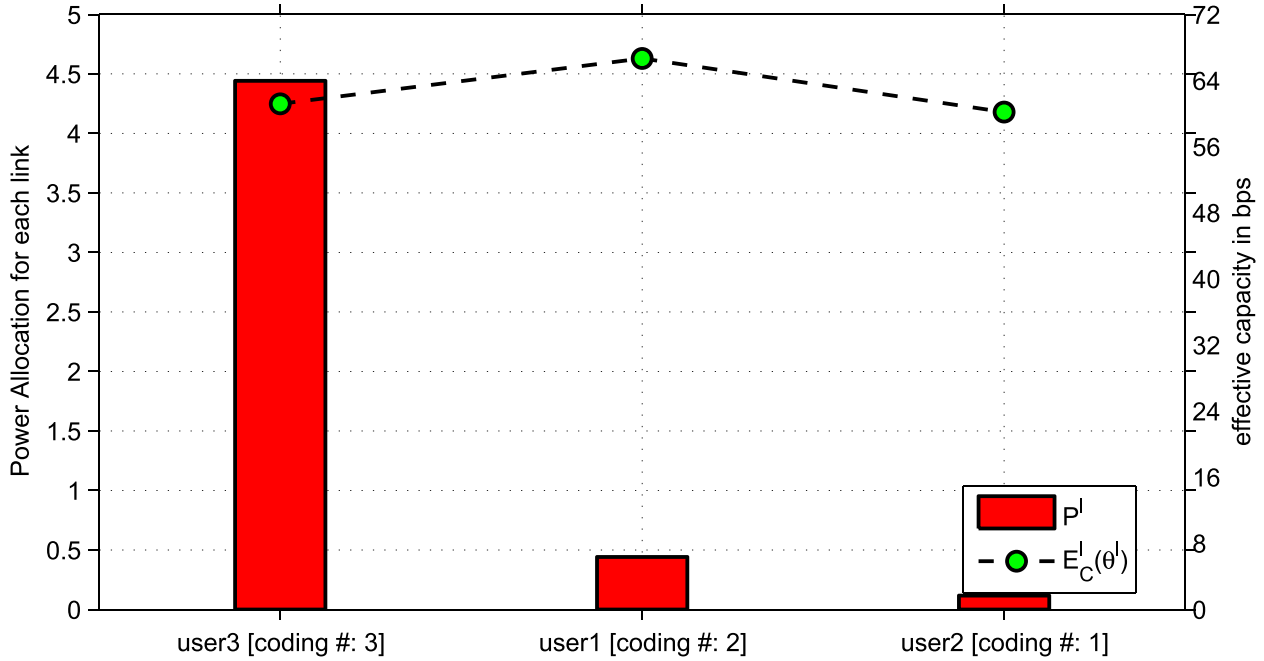

Fig. 4 Resource allocation decisions for experiment I under PAA

Table 5 Total utility and percentages of improvement

Experiment #	ETA	Dynamic-FAA	Change [%]	EPA	PAA	Change [%]
I	9.9013	10.3412	4.44	11.7575	12.4887	6.22
II	10.6575	11.8662	11.34	11.7992	12.4725	5.71
III	8.6986	10.7394	23.46	10.1263	10.9313	7.95

algorithms with these basic but practical algorithms as benchmarks, which are implemented in real systems and hardware. Note that, according to Table 5, the total utility with PAA is higher than that with dynamic-FAA. More importantly, despite the QoS requirement and channel conditions being the same, both dynamic-FAA and PAA performs slightly better than the corresponding equal resource allocation policies ETA, and EPA, respectively.

In Figs. 5 and 6, the performance results of dynamic-FAA and PAA under experiment II (i.e., non-homogeneous channel) are given. For dynamic-FAA, it can be seen that the user with the lowest channel gain is given the shortest slot length. One important observation here is that there is a linear relationship between the allocated slot length and the channel gain for each user. For PAA, we observe that the signal of the user with the highest channel gain is encoded first. As a result, BS allocates the lowest power to the signal of this user. An overall performance gain of roughly 11 and 6 % is achieved for dynamic-FAA and PAA comparing to ETA and EPA, respectively.

In Figs. 7 and 8, the results of experiment III corresponding to homogeneous channel conditions and heterogeneous QoS requirements are given. From Fig. 7, we can observe that the user with the lowest QoS-exponent, i.e., the loosest delay requirement, is assigned the shortest slot length. Despite this allocation, its effective capacity is the largest among all users. This can be explained by the fact that QoS exponent affects the value of the effective capacity more than the channel gains. In Fig. 8, we see that user 3, whose effective capacity is expected to be low due to its strict QoS demand, is encoded first in order to save it from additional utility loss due to noise originating from the signals of the other users.

It is worthwhile to note that both dynamic-FAA and PAA perform better than equal resource allocation especially when QoS requirements are heterogeneous, as opposed to when channel conditions are different. Moreover, under the same conditions, PAA's performance is better than dynamic-FAA's performance. In particular, dynamic-FAA improves utility function by 11.34 % when channel conditions are different, while the improvement is 23.46 % when QoS requirements are heterogeneous. For the case of PAA this becomes, 5.71 and 7.95 % for heterogeneous channel conditions and heterogeneous QoS requirements, respectively. This attribute is to the benefit of the 4G broadband wireless access technologies such as LTE, which are used for transporting a mixture of data, voice and video services, each with its own QoS requirement.

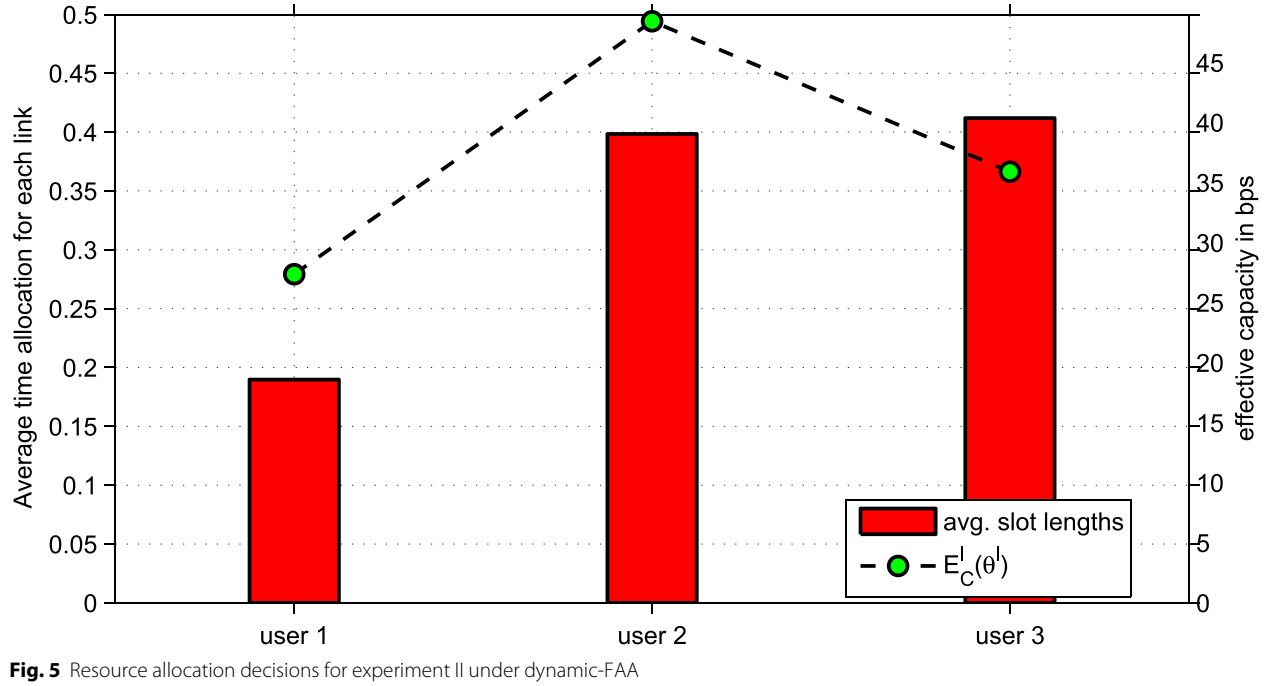

Fig. 5 Resource allocation decisions for experiment II under dynamic-FAA

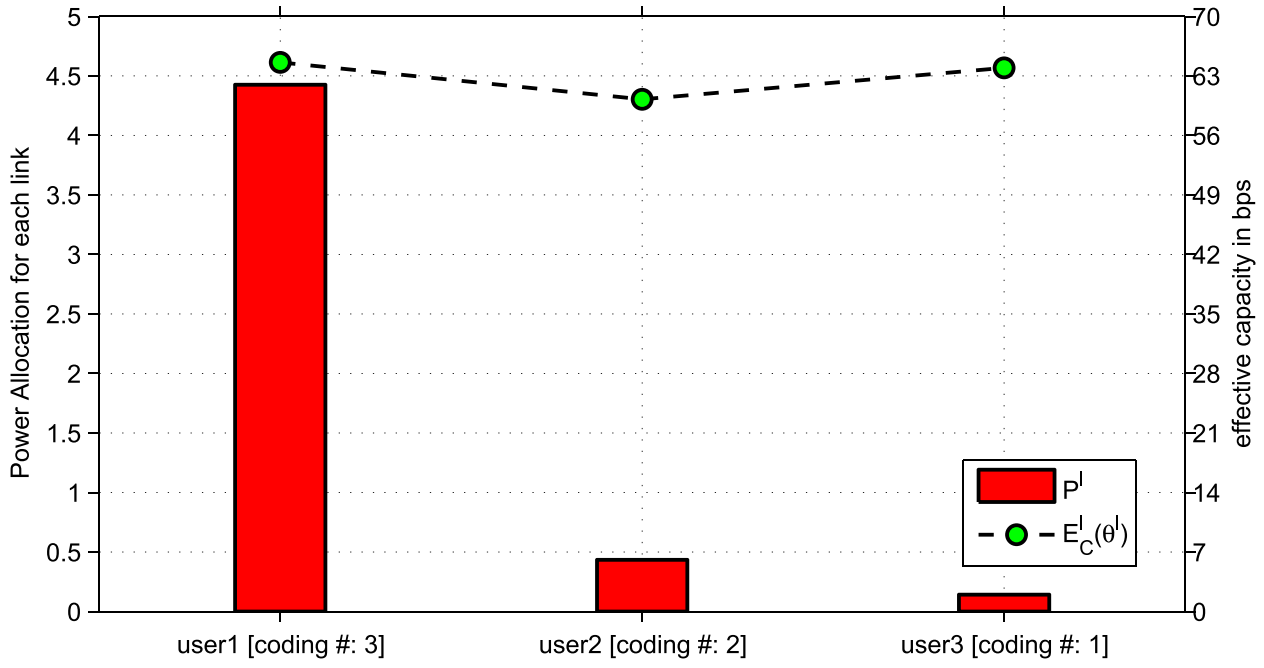

Fig. 6 Resource allocation decisions for experiment II under PAA

7 Conclusions

In this paper, we have investigated the cross-layer resource allocation problem for providing diverse QoS guarantees over MIMO downlink networks. Effective capacity of MIMO links are calculated under two different resource allocation regimes, where either time or fixed power resources are allocated among users. We have developed two resource allocation algorithms FAA and PAA under these two regimes, as solutions of network utility maximization formulations. We demonstrate in detail, the

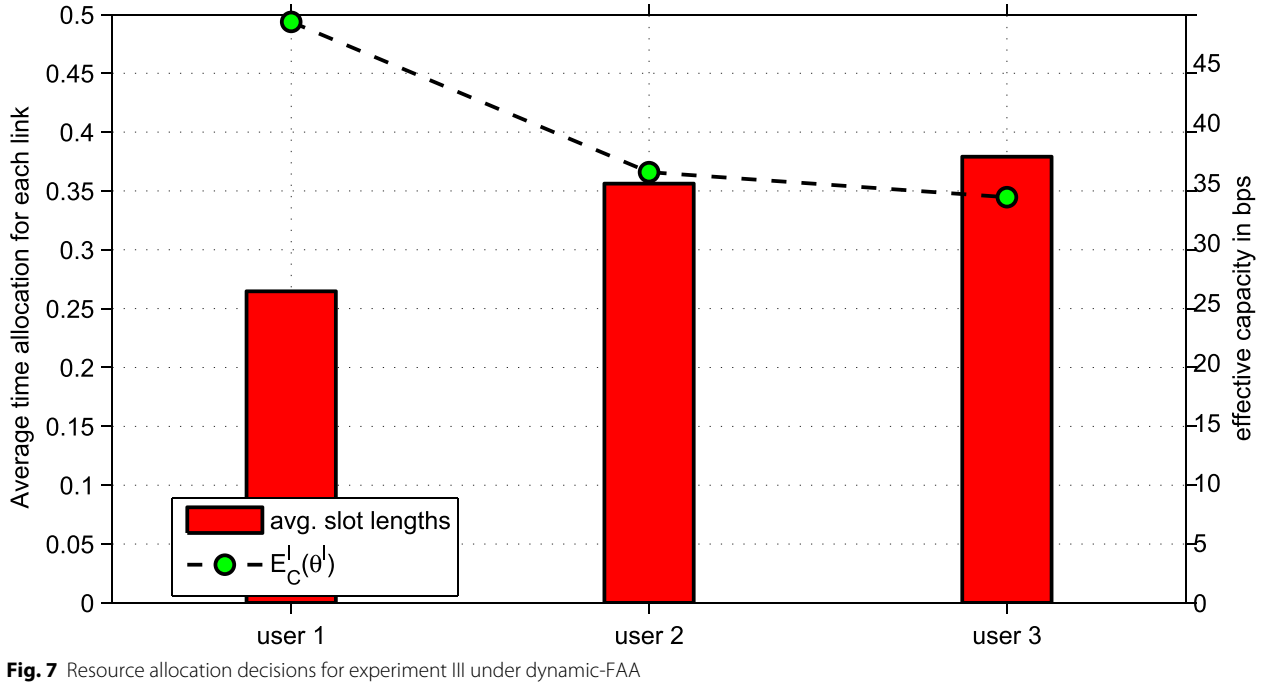

Fig. 7 Resource allocation decisions for experiment III under dynamic-FAA

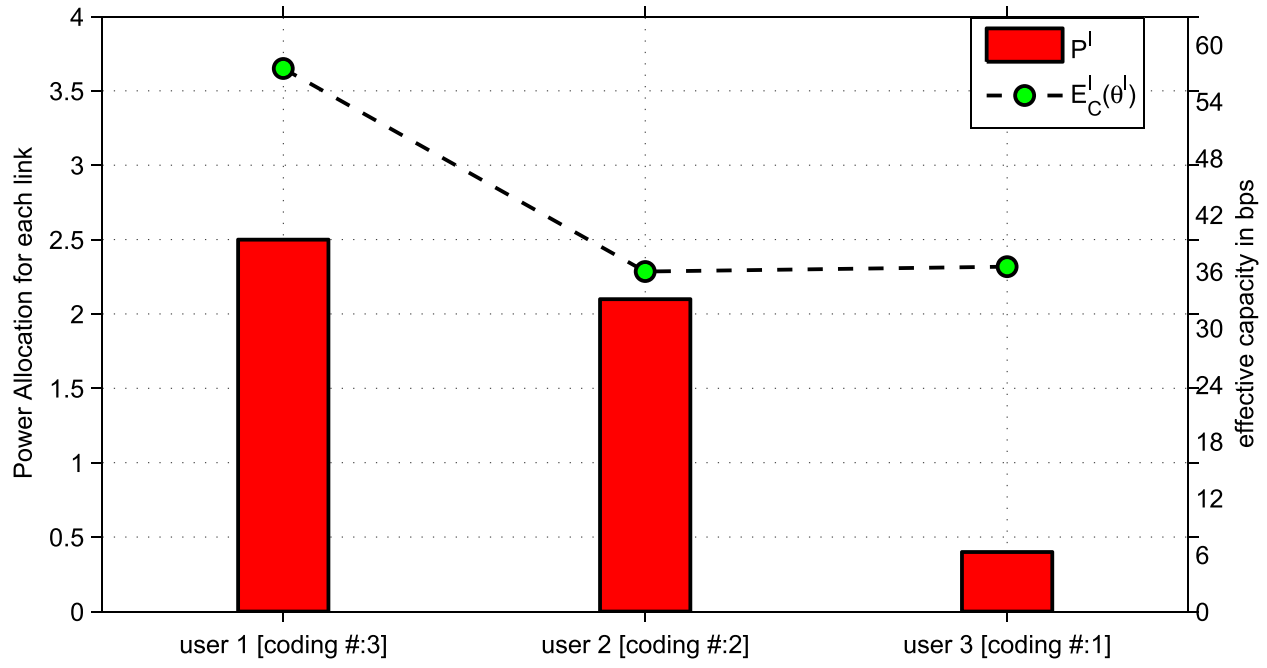

Fig. 8 Resource allocation decisions for experiment III under PAA

effects of QoS parameters and channel conditions on the resource allocation decisions via numerical experiments. In particular, we observe that both FAA and PAA achieve larger improvement when QoS requirements are heterogeneous as opposed to when channel conditions are different. As a future work, we aim to investigate the practical applications and implementation of the developed algorithms in IEEE 802.11n/ac wireless networks.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported in part by TUBITAK grant No:109E242.

Author details
[1]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey. [2]School of Computer Engineering, Iran University of Science & Technology, Tehran, Iran.

Received: 12 February 2015 Accepted: 9 September 2015
Published online: 22 September 2015

References

1. IE Telatar, Capacity of multi-antenna gaussian channels. European Trans. Telecomm. **10**(6), 585–596 (1999)
2. S Shakkottai, Effective capacity and QOS for wireless scheduling. IEEE Trans. Autom. Control. **53**(3), 749–761 (2008)
3. AA Khalek, C Caramanis, RW Heath Jr, in *IEEE GLOBECOM*. Video quality-maximizing resource allocation and scheduling with statistical delay guarantees (IEEE New York, NY, USA, 2013)
4. MC Gursoy, Mimo wireless communications under statistical queueing constraints. IEEE Trans. Inf. Theory. **57**(9), 5897–5917 (2011)
5. D Wu, R Negi, Effective capacity: a wireless link model for support of quality of service. IEEE Trans. Wirel. Commun. **2**(4), 630–643 (2000)
6. Z Ji, C Dong, Y Wang, J Lu, in *IEEE ICC*. On the analysis of effective capacity over generalized fading channels (IEEE New York, NY, USA, 2014), pp. 1977–1983
7. X Guo, L Dong, Y Li, L Wang, in *13th Canadian Workshop on Information Theory (CWIT)*. Effective capacity of MIMO MRC system with constant and variable power loading (IEEE New York, NY, USA, 2013), pp. 117–121
8. J Li, X Weiwei, P Martins, L Shen, Low complexity user scheduling for multi-antenna gaussian broadcast systems with quality of service requirements. IET Commun. **8**(10), 1820–1830 (2014)
9. Y Chen, L Dong, I Darwazeh, in *9th IEEE International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP)*. Effective capacity-based delay performance estimators for LTE radio bearer QOS provision (IEEE New York, NY, USA, 2014)
10. M Kashif, A Rizk, Y Jiang, in *IEEE ICC*. On the flow-level delay of a spatial multiplexing MIMO wireless channel (IEEE New York, NY, USA, 2011)
11. M Zorzi, J Zeidler, A Anderson, B Rao, J Proakis, AL Swindlehurst, M Jensen, S Krishnamurthy, Cross-layer issues in MAC protocol design for MIMO ad hoc networks. IEEE Wireless Commun. **13**(4), 62–76 (2006)
12. J Liu, Y Shi, YT Hou, in *IEEE INFOCOM*. A tractable and accurate cross-layer model for multi-hop MIMO networks (IEEE New York, NY, USA, 2010), pp. 1–9
13. K Mahmood, M Vehkapera, Y Jiang, in *IEEE ICCCN*. Delay constrained throughput analysis of a correlated MIMO wireless channel (IEEE New York, NY, USA, 2011), pp. 1–7
14. K Sundaresan, R Sivakumar, MA Ingram, TY Chang, Medium access control in ad hoc networks with MIMO links: optimization considerations and algorithms. IEEE Trans. Mob. Comput. **3**(4), 350–365 (2004)
15. K Sundaresan, R Sivakumar, A unified MAC layer framework for ad-hoc networks with smart antennas. IEEE/ACM Trans. Netw. **15**(3), 546–559 (2007)
16. B Hamdaoui, P Ramanathan, A cross-layer admission control framework for wireless ad-hoc networks using multiple antennas. IEEE Trans. Wirel. Commun. **6**(11), 4014–4024 (2007)
17. YH Lin, T Javidi, RL Cruz, LB Milstein, in *IEEE Fortieth Asilomar Conference on Signals, Systems and Computers*. Distributed link scheduling power control and routing for multi-hop wireless MIMO networks (IEEE New York, NY, USA, 2006), pp. 122–126

18. K Sundaresan, R Sivakumar, Routing in ad-hoc networks with MIMO links: Optimization considerations and protocols. Comput. Netw. **52**(14), 2623–2644 (2008)
19. B Mumey, J Tang, T Hahn, in *IEEE International Conference on Communications*. Joint stream control and scheduling in multihop wireless networks with MIMO links (IEEE New York, NY, USA, 2008), pp. 2921–2925
20. T Elbatt, in *IEEE MILCOM*. Towards scheduling MIMO links in interference-limited wireless ad hoc networks (IEEE New York, NY, USA, 2007), pp. 1–7
21. MO Pun, W Ge, D Zheng, J Zhang, VH Poor, in *IEEE International Conference on Communications*. Distributed opportunistic scheduling for MIMO ad-hoc networks (IEEE New York, NY, USA, 2008), pp. 3689–3693
22. M Zhao, M Ma, Y Yang, in *Managing Traffic Performance in Converged Networks*. Opportunistic medium access control in MIMO wireless mesh networks (Springer Berlin, 2007), pp. 458–470
23. X Liu, NB Shroff, EKP Chong, Opportunistic scheduling: An illustration of cross-layer design. Telecommun. Rev. **14**(6), 947–959 (2004)
24. H Yin, H Liu, Performance of space-division multiple-access (SDMA) with scheduling. IEEE Trans. Wirel. Commun. **1**(4), 611–618 (2002)
25. W Cheng, Z Xi, H Zhang, QOS-aware power allocations for maximizing effective capacity over virtual-MIMO wireless networks. IEEE J. Selected Areas Commun. **31**(10), 2043–2057 (2013)
26. W Cheng, Z Xi, H Zhang, Joint spectrum and power efficiencies optimization for statistical QOS provisionings over SISO/MIMO wireless networks. IEEE J. Selected Areas Commun. **31**(5), 903–915 (2013)
27. K Higuchi, Y Kishiyama, in *Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th*. Non-orthogonal access with random beamforming and intra-beam sic for cellular MIMO downlink (IEEE New York, NY, USA, 2013), pp. 1–5
28. B Kimy, S Lim, H Kim, S Suh, J Kwun, S Choi, C Lee, S Lee, D Hong, in *Military Communications Conference, MILCOM 2013-2013 IEEE*. Non-orthogonal multiple access in a downlink multiuser beamforming system (IEEE New York, NY, USA, 2013), pp. 1278–1283
29. N Nonaka, A Benjebbour, K Higuchi, in *Communication Systems (ICCS), 2014 IEEE International Conference On*. System-level throughput of noma using intra-beam superposition coding and sic in MIMO downlink when channel estimation error exists (IEEE New York, NY, USA, 2014), pp. 202–206
30. A Benjebbour, Y Saito, Y Kishiyama, A Li, A Harada, T Nakamura, in *Intelligent Signal Processing and Communications Systems (ISPACS), 2013 International Symposium On*. Concept and practical considerations of non-orthogonal multiple access (noma) for future radio access (IEEE New York, NY, USA, 2013), pp. 770–774
31. Y Saito, Y Kishiyama, A Benjebbour, T Nakamura, A Li, K Higuchi, in *Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th*. Non-orthogonal multiple access (noma) for cellular future radio access (IEEE New York, NY, USA, 2013), pp. 1–5
32. Y Saito, A Benjebbour, Y Kishiyama, T Nakamura. Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium On (IEEE New York, NY, USA, 2013), pp. 611–615
33. M-R Hojeij, J Farah, CA Nour, C Douillard, New optimal and suboptimal resource allocation techniques for downlink non-orthogonal multiple access. Wirel. Pers. Commun, 1–31 (2015). http://link.springer.com/article/10.1007/s11277-015-2629-2
34. S Timotheou, I Krikidis, Fairness for non-orthogonal multiple access in 5g systems. Signal Process. Lett. IEEE. **22**(10), 1647–1651 (2015)
35. Z Ding, Z Yang, P Fan, HV Poor, On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users. Signal Process. Lett. IEEE. **21**(12), 1501–1505 (2014)
36. S Vassaki, AD Panagopoulos, P Constantinou, Effective capacity and optimal power allocation for mobile satellite systems and services. Commun. Lett. IEEE. **16**(1), 60–63 (2012)
37. L Mao, S Xu, T Fu, Q Huang, in *Vehicular Technology Conference (VTC Fall), 2012 IEEE*. Game theory based power allocation algorithm in high-speed mobile environment (IEEE New York, NY, USA, 2012), pp. 1–5
38. W Cheng, X Zhang, H Zhang, Joint spectrum and power efficiencies optimization for statistical qos provisionings over SISO/MIMO wireless networks. Selected Areas Commun. IEEE J. **31**(5), 903–915 (2013)
39. W Cheng, X Zhang, H Zhang, Qos-aware power allocations for maximizing effective capacity over virtual-MIMO wireless networks. Selected Areas in Commun. IEEE J. **31**(10), 2043–2057 (2013)
40. D Qiao, MC Gursoy, S Velipasalar, Transmission strategies in multiple-access fading channels with statistical QOS constraints. Inf. Theory, IEEE Trans. **58**(3), 1578–1593 (2012)
41. CS Chang, JA Thomas, Effective bandwidth in high-speed digital networks. IEEE J. Selected Areas Commun. **13**(6), 1091–1100 (1995)
42. L Massoulie, J Roberts, in *Proceedings of IEEE Infocom*. Bandwidth sharing: objectives and algorithms (IEEE New York, NY, USA, 1999), pp. 1395–1403
43. J Liu, Y Shi, YT Hou, in *IEEE INFOCOM*. A tractable and accurate cross-layer model for multi-hop MIMO ad-hoc networks (IEEE New York, NY, USA, 2010), pp. 1–9
44. CS Chang, *Performance Guarantees in Communication Networks*. (Springer, Berlin, 2000)
45. A Bennatan, D Burshtein, G Caire, S Shamai, Superposition coding for side-information channels. IEEE Trans. Inf. Theory. **52**(5), 1872–1889 (2006)
46. I Krikidis, JS Thompson, Mimo two-way relay channel with superposition coding and imperfect channel estimation. J. Netw. Comput. Appl. **35**(1), 510–516 (2012)
47. B Soret, MC Aguayo-Torres, JT Entrambasaguas, Capacity with explicit delay guarantees for generic sources over correlated rayleigh channel. IEEE Trans. Wirel. Commun. **9**(6), 1901–1911 (2010)
48. O Ercetin, MO Memis, Comments on "capacity with explicit delay guarantees for generic sources over correlated rayleigh channel" (2011). arXiv preprint arXiv 1112.5152
49. FP Kelly, *Notes on effective bandwidths, royal statistical society lecture notes series*, vol. 4. (Oxford University Press, London, 1996), pp. 141–168