Queue Aware Resource Optimization in Latency Constrained Dynamic Networks

Inosha Sugathapala*, Savo Glisic*, Markku Juntti*, A. Shams Shafigh* and Le-Nam Tran†
*Centre for wireless Communication, University of Oulu, Finland, †University College Dublin, Ireland.
Email:{*inosha.sugathapala, *savo.glisic, *markku.juntti, *alireza.shamsshafigh}@oulu.fi, †nam.tran@ucd.ie.

***Abstract*—Low latency communications is one of the key design targets in future wireless networks. We propose a queue aware algorithm to optimize resources guaranteeing low latency in multiple-input single-output (MISO) networks. Proposed system model is based on dynamic network architecture (DNA), where some terminals can be configured as temporary access points (APs) on demand when connected to the Internet. Therein, we jointly optimize the user-AP association and queue weighted sum rate of the network, subject to limitations of total transmit power of the APs and minimum delay requirements of the users. The user-AP association is viewed as finding a sparsity constrained solution to the problem of minimizing ℓ_q-norm of the difference between queue and service rate of users. Finally, the efficacy of the proposed algorithm in terms of network latency and its fast convergence are demonstrated using numerical experiments. Simulation results show that the proposed algorithm yields up to two-fold latency reductions compared to the state-of-the-art techniques.**

***Index Terms*—Dynamic networks, user association, SOCP, queue weighted sum rate, convex optimization, low latency.**

I. Introduction

The ubiquitous use of advanced wireless devices such as smart phones and laptops laid the foundation for the concept of dynamic network architecture (DNA): configuring smart devices to operate as access points (APs) if required [1]. Such networks do not require additional pre-installed infrastructure to accommodate excess traffic, and large numbers of users and their dynamic availability make DNA highly adaptive to traffic variations in the network [2]. At the same time, wireless beamforming techniques have gradually matured to a level that they can be integrated into many wireless systems [3]. When equipped with multiple antennas, APs provide more degrees of freedom, which can be exploited to improve the spectral efficiency of the system through spatial channel reuse. In this regard, Vu *et al.* [3] proposed an algorithm to optimize the rate in low latency network while Venkatraman *et al.* [4] presented a different algorithm to optimize resource allocation for queue aware network by minimizing ℓ_q-norm of the difference between queues and service rates in MISO networks. In [3], the delay bounds guaranteed with certain probability are analyzed as a function of current queue length and packet arrival rate. We state, however, the above-mentioned algorithms are accounted for a fixed user-AP association. Funabiki *et al.* [5] highlight that maximizing the number of active APs in the network may not always guarantee to achieve higher QoS or network throughput unless proper user-AP selection is performed. For our DNA framework, where the numbers of users and the available APs can change in time and space, a user-AP selection mechanism is required. For cellular wireless networks, the problem of joint optimization of power and AP selection has been investigated in some pioneering works. Hanly [6] as well as Yates and Huang [7] investigated the problem of joint user-base station association and power optimization for uplink transmission. In [8], the problem of joint AP selection and power allocation for a multi-carrier wireless network with multiple APs and mobile users was considered. On the other hand, this paper highlights some problems regarding the use of non-cooperative game theoretical approaches to solve resource allocation problems in wireless networks. Therefore, different from the earlier studies, we propose a queue aware algorithm to optimize beamformers of the network while guarantee the required network latency using convex approximations. Our **contributions** are as follows:

- We propose a low-complexity iterative algorithm to jointly optimize the user-AP association and the queue weighted sum rate. As far as an optimal solution to the user-AP association problem is concerned, an exhaustive search over all possible combinations is normally required, which is impractical even for a small number of users and APs. A common solution in practice is to assign a user to the closest AP based on the path-loss (PL) [9]. We demonstrate by numerical results that the algorithm proposed in this paper outperforms the PL-based assignment. Moreover, for a small number of users and APs, the sub-optimal association scheme obtained from our proposed algorithm performs nearly close to the exhaustive search.
- The queue weighted sum-rate maximization problem considered in this paper also introduces probabilistic constraint to guarantee low latency transmission. The original problem is non-convex, and, thus, convex-concave procedure (CCP) [10] and Markov inequality are used to derive an iterative algorithm to achieve locally optimal solution. Our contribution in this regard is to develop second-order cone programs (SOCPs) in each iteration of the proposed iterative procedure to jointly optimize the user-AP association and queue weighted sum-rate while guaranteeing low latency requirement of the users (see [11] for the details on the convexification method and its convergence). Moreover, the SOCP based proposed

This publication has emanated from research supported in part by the academy of Finland 6Genesis Flagship (grant no. 318927) and a grant from Science Foundation Ireland under Grant number 17/CDA/4786.

algorithm is fast converging algorithm and output is demonstrated by comparing our algorithm output with the resource allocation algorithm proposed in [4] for known user-AP association. According to the numerical experiments, our proposed CCP based algorithm is 80% more efficient than the considered algorithm in literature.

Notations: Boldface lower and upper case letters are used to denote vectors and matrices, respectively. $\mathrm{Re}(\mathbf{x})$ and $\mathrm{Im}(\mathbf{x})$ represent the real and imaginary parts of a complex vector $\mathbf{x}$, respectively. $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ represent the space of real and complex matrices of dimensions given in superscript, respectively; $\mathbf{X}^T$ and $\mathbf{X}^H$ are the transpose and Hermitian transpose of $\mathbf{X}$, respectively; $[y]^+ \triangleq \max\{y, 0\}$. The absolute value of a scalar y is defined by $|y|$, and $\|\mathbf{y}\|_q$ represents the ℓ_q-norm of a vector $\mathbf{y} = [y_1, y_2, \dots y_n]$, i.e., $\|\mathbf{y}\|_q = (\sum_{i=1}^{n} |y_i|^q)^{1/q}$. For two vectors $\mathbf{x}$ and $\mathbf{y}$ of the same size, their inner product is denoted by $\langle \mathbf{x}, \mathbf{y}\rangle$, i.e., $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^T\mathbf{y}$; $[\mathbf{x}]_i$ denotes the ith element of a vector $\mathbf{x}$ and $(\mathbf{X})_i$ represents the ith row of a matrix $\mathbf{X}$. The complex normal distribution is denoted by $\mathcal{CN}$.

II. Network Model and Problem Formulation

At a given time instant we consider the downlink transmission of the DNA, where APs communicate with users in a single-hop manner. The network model consists of sets of $\mathcal{J} = \{1, 2, \dots, J\}$ single antenna users and $\mathcal{A} = \{1, 2, \dots, A\}$ APs with T transmit antennas. It is assumed that data is not shared among APs. The problem of interest is to minimize the back-log queue and delay while optimizing the user-AP association by designing the beamformers for each user.

The complexity of the proposed system increases with the number of APs and users. Therefore, in order to make the optimization process feasible, network clustering is introduced to reduce the size of the network under consideration [12]. Frequency reuse factor is considered as 3 for network partitioning; thus, inter-cluster interference is ignored. Fig. 1 illustrates a simple network model for a single cluster with the sets of users and APs. Furthermore, it is assumed that the central server handles each independent cluster and full channel state information (CSI) of the network is available. Although the assumption may be sometimes optimistic, the results provide an important performance benchmark. Moreover, time-division duplex allows the CSI acquisition utilizing the channel reciprocity. When architecture is based on well-known protocols like universal plug and play (UPnP) and devices profile for web services (DPWS) to control and exchange information, all the APs can transmit CSI to the central processor with little overhead. Therefore, we propose an algorithm for a cluster [12] to optimize queue weighted sum rate and user-AP association.

The channel between AP a and user j is denoted by a complex row vector $\mathbf{h}_{aj} = \sqrt{\eta\gamma}\tilde{\mathbf{h}}_{aj} \in \mathbb{C}^{1\times T}$, where η represents the log-normal shadowing, i.e., $\log(\eta) \sim \mathcal{CN}(\mathbf{0}, \zeta^{\mathbf{2}})$, γ is the path loss, and $\tilde{\mathbf{h}}_{aj}$ is the small-scale fading modeled as $\mathcal{CN}(\mathbf{0}, \mathbf{I})$. The message for user j is linearly weighted by a beamforming column vector $\mathbf{w}_{aj} \in \mathbb{C}^{T\times 1}$, before being transmitted from AP a.

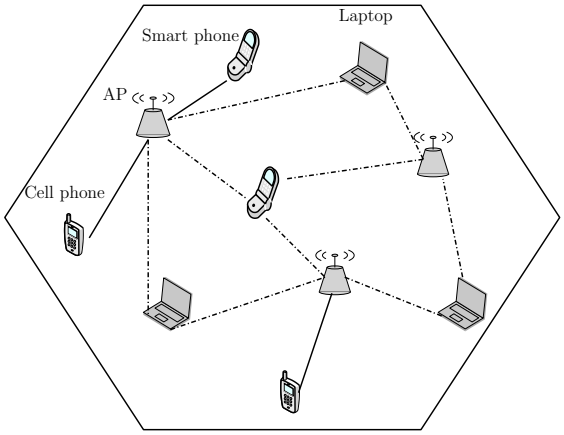

Fig. 1. A single cluster based scenario of our system model. The black dash dotted lines indicate potential connections between a user and an APs and black solid lines refer to the selected connection between a user and an AP.

The objective of this paper is to maximize the queue weighted sum rate to reduce queue length and transmission delay of the network, i.e.,

$$\mathbb{U} = \max f(\mathbf{r}, \mathbf{Q}) = \max \textstyle\sum_{\forall j\in\mathcal{J}} Q_j r_j \tag{1}$$

where Q_j is the number of backlog bits[1] destined for user j with $\mathbf{Q} \triangleq [Q_1, Q_2 \dots Q_J]$ and r_j is the allocated transmission data rate for user j with $\mathbf{r} \triangleq [r_1, r_2 \dots r_J]$, respectively. Meanwhile, it is shown in [4], that minimizing the ℓ_q-norm of the difference between users' queues and users' transmission rates is equivalent to maximizing the queue weighted sum rate. Thus, to achieve better tractability, the objective function is modified as $\mathbb{U} = \min f_q(\mathbf{r}, \mathbf{Q}) = ||\mathbf{v}||_q$, where $\mathbf{v} \in \mathbb{R}^{J\times 1}$ is the vector with entries $v_j = Q_j - r_j, \quad \forall \mathcal{J}$.

The data symbol transmitted to user j from AP a represents $m_{aj} \in \mathbb{C}$. We assume that the information symbols for different users are independent, i.e., $\mathbb{E}\{m_{aj}m_{al}^*\} = 0$ for all $l, j \in \mathcal{J}$ where $l \neq j$; and we also assume that m_{aj} is normalized $\mathbb{E}|m_{aj}|^2 = 1$. Therefore, when user j is associated with a given AP $\tilde{a}$, the received signal for user j can be written as,

$$y_j = \underbrace{m_{\tilde{a}j}\mathbf{h}_{\tilde{a}j}\mathbf{w}_{\tilde{a}j}}_{\text{Desired Signal}} + \underbrace{\sum_{\forall a\in\mathcal{A}}\sum_{\forall l\in\mathcal{J}\setminus\{j\}} m_{al}\mathbf{h}_{aj}\mathbf{w}_{al}}_{\text{Interference}} + \underbrace{n_j}_{\text{Noise}} \tag{2}$$

The additive white Gaussian noise (AWGN), n_j has the complex normal distribution $\mathcal{CN}(\mathbf{0}, \sigma_j^2)$ and σ_j^2 is the noise power. Therefore, the data rate of user j is $r_j = \log(1 + \hat{\gamma}_{\tilde{a}j}), \quad \forall j \in \mathcal{J}$, where

$$\hat{\gamma}_{\tilde{a}j} = \frac{|\mathbf{h}_{\tilde{a}j}\mathbf{w}_{\tilde{a}j}|^2}{\sigma_j^2 + \sum_{a\in\mathcal{A}}\sum_{l\in\mathcal{J}\setminus\{j\}} |\mathbf{h}_{aj}\mathbf{w}_{al}|^2}. \tag{3}$$

For the formulation of AP-user association problem, let us express another form of SINR of user j in a general way as

$$\gamma_j = \frac{\sum_{a\in\mathcal{A}} |\mathbf{h}_{aj}\mathbf{w}_{aj}|^2}{\sigma_j^2 + \sum_{a\in\mathcal{A}}\sum_{l\in\mathcal{J}\setminus\{j\}} |\mathbf{h}_{aj}\mathbf{w}_{al}|^2}. \tag{4}$$

[1]As long as the arrival and the transmission data rates units are same, the unit can either be bits or packets.

Moreover, we define $\mathbf{h}_j = [\mathbf{h}_{1j}, \mathbf{h}_{2j}, ...\mathbf{h}_{Aj}] \in \mathbb{C}^{1\times AT}$ to be the aggregated channel vector and $\mathbf{w}_j = [\mathbf{w}_{1j}^T, \mathbf{w}_{2j}^T, ...\mathbf{w}_{Aj}^T]^T \in \mathbb{C}^{AT\times 1}$ to be the aggregated beamformer vector for user j. The equivalent representation of (4) can be written as $\gamma_j = \frac{|\mathbf{h}_j\mathbf{w}_j|^2}{\sigma_j^2+\sum_{l\in\mathcal{J}\setminus\{j\}}|\mathbf{h}_j\mathbf{w}_l|^2}$. The numerator of (4) includes all the possible signal powers from every AP $a \in \mathcal{A}$ to user j. In order to (4) to be a valid SINR expression for user j (i.e., in the form of (3)), we need to force all other beamformers $\mathbf{w}_{ij}$'s related to user j to be zero, except $\mathbf{w}_{\tilde{a}j}$ for the selected AP $\tilde{a}$. To clarify this point, we introduce a new matrix $\mathbf{\Psi}_j$,

$$\mathbf{\Psi}_j = \begin{bmatrix} [\mathbf{w}_{1j}]_1 \cdots [\mathbf{w}_{1j}]_T \\ \vdots \qquad\quad \vdots \\ [\mathbf{w}_{\tilde{a}j}]_1 \cdots [\mathbf{w}_{\tilde{a}j}]_T \\ \vdots \qquad\quad \vdots \\ [\mathbf{w}_{Aj}]_1 \cdots [\mathbf{w}_{Aj}]_T \end{bmatrix} \in \mathbb{C}^{A\times T} \tag{5}$$

where the $\tilde{a}$th row of $\mathbf{\Psi}_j$ corresponds to the set of beamformers for user j associated with the $\tilde{a}$th AP. To select an appropriate AP to communicate with user j, all rows of $\mathbf{\Psi}_j$ should be nulled out except the $\tilde{a}$th row. In the context of sparsity-constrained optimization, this is equivalent to imposing some degree of sparsity on $\mathbf{\Psi}_j$. More precisely, we impose group sparsity on $\mathbf{\Psi}_j$, such that all rows of $\mathbf{\Psi}_j$ except one are encouraged to be zero as discussed in [13]. This can be done by applying a mixed ℓ_1/ℓ_2 norm to the matrix $\mathbf{\Psi}_j$. Specifically, the mixed ℓ_1/ℓ_2 norm acts as ℓ_2 norm for the rows of $\mathbf{\Psi}_j$ and ℓ_1 norm for the columns of $\mathbf{\Psi}_j$. The mixed ℓ_1/ℓ_2 norm for $\mathbf{\Psi}_j$ can be written as

$$\|\mathbf{\Psi}_j\|_{1,2} = \textstyle\sum_i \|(\mathbf{\Psi}_j)_i\|_2 = (\sum_i(\sum_z |[\mathbf{w}_{ij}]_z|)^2)^{1/2}. \tag{6}$$

We redefine the objective function $\mathbb{U}$ to optimize user-AP association in the queue weighted sum rate maximization problem as

$$\mathbb{U} = \min \|\mathbf{Q} - \mathbf{r}\|_q + \rho \|\mathbf{\Psi}_j\|_{1,2} \tag{7}$$

where ρ is a positive constant which controls the degree of sparsity of the solution and $r_j = \log(1+\gamma_j), \quad \forall j \in \mathcal{N}$.

We model the queues using the same queue-aware scheduling model as that in [14]. The number of bits per unit frequency arriving at the central server to transmit user j, at slotted discrete time t is denoted as $x_j(t)$. Here, a time slot corresponds to the channel coherence time within which the channel is assumed to be constant. Moreover, channels over time slots are independent from one another. By assuming the arrivals are i.i.d. over time slots, the mean arrival rate can be denoted by $\mathbb{E}[x_j(t)] = \lambda_j$ with the upper bound $x_j^{\max}$. Moreover, the central server has the knowledge of the queue backlog $Q_j(t)$ bits of each user associated with it. Hence, the queue time evolution for user j is $Q_j(t+1) = [Q_j(t) - r_j(t)]^+ + x_j(t)$. Furthermore, within the network data may be transmitted at maximum rates to avoid transmission delay. However, this might force the network to over-allocate the resources [3], and, thus, we introduce a maximum transmission rate constraint for each user. Practically, this should be equal to backlog queue of the user j, i.e., $Q_j(t)$ to minimize the idling time at APs. Together with the minimum rate ($r_j^{\min}(t)$) requirement to guarantee QoS of the users, the instant rate constrained can be defined as

$$Q_j(t) \geq r_j(t) \geq r_j^{\min}(t), \quad \forall j \in \mathcal{J}. \tag{8}$$

According to the Little's formula [15], the average delay is proportional to $\lim_{\tau\to\infty}\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}[Q_j(\tau)]/\lambda_j$. Thereby, we refer $Q_j(t)/\lambda_j$ as the transmission delay for user j with allowed upper bound $d_j^{\max}$. Hence, the probabilistic constraint is imposed on user j as

$$\Pr\{Q_j(t)/\lambda_j \geq d_j^{\max}\} \leq \epsilon_j, \quad \forall t \tag{9}$$

to minimize the transmission delay [3]. Moreover, the target probability should be sufficiently low to guarantee high reliability, i.e., $\epsilon_j << 1$. Furthermore, the Markov's inequality claims, $\Pr\{Q_j(t) \geq \lambda_j d_j^{\max}\} \leq \mathbb{E}[Q_j(t)]/\lambda_j d_j^{\max}$ [16]; thus, when the system satisfy,

$$\mathbb{E}[Q_j(t)] \leq \lambda_j d_j^{\max}\epsilon_j, \quad \forall j \in \mathcal{J} \tag{10}$$

it will always satisfy the (9). Therefore, conservative approximation of (9) can be represented with (10). Further, assuming $\{x_j(t)|\forall t \geq 1\}$ is a Poisson arrival process, average queue size for time frame t is equal to the difference between the number of bits arrived and transmitted for unit frequency, i.e., $E[Q_j(t)] = t\lambda_j - \sum_{\tau=1}^{t} r_j(\tau)$ [16]. Therefore, (10) becomes

$$r_j(t) \geq t\lambda_j - \lambda_j d_j^{\max}\epsilon_j - \textstyle\sum_{\tau=1}^{t-1} r_j(\tau). \qquad \forall j \in \mathcal{J} \tag{11}$$

Together with (8), the QoS constraint for user j can be modified as

$$Q_j(t) \geq r_j(t) \geq \max\{r_j^{\min}(t), t\lambda_j - \lambda_j d_j^{\max}\epsilon_j - \textstyle\sum_{\tau=1}^{t-1} r_j(\tau)\}. \tag{12}$$

Furthermore, there is a limited power budget at a given AP a, which is given by

$$\textstyle\sum_{j\in\mathcal{J}} \|\mathbf{w}_{aj}(t)\|_2^2 \leq p_a^{\max}, \ \forall a \in A. \tag{13}$$

In summary, the joint optimization problem of delay minimization and user-APs association is stated as

$$\min_{\mathbf{w}} \ \mathbb{U} \tag{14a}$$

$$\text{sub. to} \ \ Q_j \geq r_j \tag{14b}$$

$$\textstyle\sum_{j\in\mathcal{J}} \|\mathbf{w}_{aj}\|_2^2 \leq p_a^{\max} \tag{14c}$$

$$r_j = \log\left(1+\gamma_j\right) \tag{14d}$$

$$r_j \geq \max\{r_j^{\min}, t\lambda_j - \lambda_j d_j^{\max}\epsilon_j - \sum_{\tau=1}^{t-1} r_j(\tau)\} \tag{14e}$$

Note that for notational simplicity we have excluded the time index t from (14) and the rest of the discussion.

III. SOCP based Algorithm to Optimize Queue Weighted Sum Rate and User-AP Association.

Generally, problem (14) is NP hard and difficult to solve even for a small network. In practice, we can find a suboptimal solution for NP hard problems based on a convexification for which a polynomial time algorithm is possible. Therefore, in

this section we describe the method to solve problem (14) to achieve a local optimal solution.

First, the non-convex constraint (14d) need to be approximated by convex constraints to arrive at a convex optimization problem. We can relax (14d) into the inequality constraint given by $r_j \leq \log\big(1+\frac{|\mathbf{h}_j\mathbf{w}_j|^2}{\sigma_j^2+\sum_{l\in\mathcal{J},l\neq j}|\mathbf{h}_j\mathbf{w}_l|^2}\big)$. The reason is that, the above inequality constraint holds with equality at optimality [17]; otherwise we can always increase r_j without violating the constraints and obtain a strictly lower objective. With the same argument, without loss of optimality, the rate constraint in (14) can be replaced with following three constraints:

$$r_j \leq \log(z_j), \tag{15a}$$

$$|\mathbf{h}_j\mathbf{w}_j|^2 \geq (z_j-1)u_j, \tag{15b}$$

$$u_j \geq (\sigma_j^2+\textstyle\sum_{\forall l\in\mathcal{J},l\neq j}|\mathbf{h}_j(t)\mathbf{w}_l|^2), \tag{15c}$$

where $\mathbf{z}\in\mathbb{R}^{J\times 1}$ and $\mathbf{u}\in\mathbb{R}^{J\times 1}$ are newly introduced auxiliary variables. Before proceeding further, we remark that (15a) and (15c) are convex. Clearly, the non-convexity in (15b) is due to the inner product of the two involved variables. To convexify (15b), the non-convex constraint, we recall the well-known equality: $4\langle\mathbf{x},\mathbf{y}\rangle = ||\mathbf{x}+\mathbf{y}||_2^2 - ||\mathbf{x}-\mathbf{y}||_2^2$, which is a difference of two convex terms. In the light of CCP, the concave term $-(z_j-1-u_j)^2$ linearizes to obtain convex term. For the description purposes, we denote $x^{(n)}$ as the value of an optimization variable x after iteration n of the proposed iterative algorithm described in **Algorithm 1**. In iteration $n+1$, using the first order Taylor series expansion, we can approximate (15b) as

$$\begin{aligned}(z_j-1+u_j)^2 \leq (z_j{}^{(n)}-1-u_j^{(n)})^2+4|\mathbf{h}_j\mathbf{w}_j^{(n)}|^2+\\ 8(\mathrm{Re}\big(\mathbf{w}_j^{(n)T}\mathbf{h}_j^T\mathbf{h}_j(\mathbf{w}_j-\mathbf{w}_j^{(n)})\big)+\mathrm{Im}\big(\mathbf{w}_j^{(n)T}\mathbf{h}_j^T\mathbf{h}_j(\mathbf{w}_j-\mathbf{w}_j^{(n)})\big))\\ +2\langle z_j{}^{(n)}-1-u_j^{(n)}, z_j-z_j{}^{(n)}-u_j+u_j^{(n)}\rangle. \end{aligned}\tag{16}$$

Therefore, the convex optimization problem at iteration $n+1$ of the proposed iterative method is given by

$$(\mathcal{P}_{(n+1)}) \triangleq \begin{cases}\min\limits_{\mathbf{w},\mathbf{r},\mathbf{u},\mathbf{z}} & ||\mathbf{Q}-\mathbf{r}||_q+\rho\,\|\mathbf{\Psi}_j\|_{1,2} \quad (17a)\\ \text{sub. to} & \text{(14e) - (14c), (15a), (15c), (16).} \quad (17b)\end{cases}$$

To conclude this section, we outline the proposed algorithm in **Algorithm 1**.

IV. Numerical Results

In this section, we evaluate the performance of the proposed algorithm by numerical experiments. We consider one cluster based simulation model with different numbers of users, APs and transmission antennas, i.e., $J\in\{1,\ldots,6\}$, $A\in\{2,3\}$ and $T\in\{3,4\}$. In the channel modeling, log-normal shadowing is considered with standard deviation of 8 dB. The path loss is modeled as $\gamma = 10^{-\kappa/10}$, where κ is given in dB by $35\log(\tilde{d})$ and $\tilde{d}$ is the distance in meters [19]. The maximum power $p^{\max}$ is 30 dBm for all the APs and σ_j^2 = -110 dBm for all the users. $r_j^{\min}$ is set to be $0.8\lambda_j$. The parameters ρ and q are set to 2, unless otherwise stated. The proposed algorithm is implemented in MATLAB environment using the conic solver SeDuMi [20] through the parser CVX [21]. Moreover, Intel® Core i5-6300U @ 2.4 GHz Processor and 8GB RAM workstation is used to run the simulations. The stopping criterion for **Algorithm 1** is when the increase in the objective values between two successive iterations is less than 10^{-4}.

In the first set of numerical experiments, the convergence rate of the algorithm is shown in Fig. 2 for $A=T=3$ with different configurations. The black color curves represent the objective value of the two different channel realizations ($CH_r = 1,2$) with $J=4$, while two red curves represent the objective value for $\rho=3$ and $q=\infty$ with $J=4$ and $CH_r=1$. The blue color curve represents the objective value of the system with $J=3$. It is clear that the convergence rate of the algorithm depends on the density of the network and the channel realizations. However, Fig. 2 shows that the convergence rate of **Algorithm 1** is nearly independent of the choice of penalty values ρ and q. After **Algorithm 1** converges, the obtained beamformers are used to calculate the actual $||\mathbf{Q}-\mathbf{r}||_q$ (and of course ρ is ignored in the objective). Further,

Algorithm 1 Joint optimization for queue minimization and user-AP association in low latency DNA

Initialization:

1: Set $n=0$ and find a feasible solution $(\mathbf{w}^{(0)},\mathbf{u}^{(0)},\mathbf{z}^{(0)})$ by solving the following convex feasibility problem with pre-defined user-AP association. To increase the chance of obtaining a feasible solution, a reasonable way is to assign users to APs with higher channel gains.

$$\text{find } \mathbf{w} \tag{18a}$$

$$\text{subject to } Q_j \geq r_j \geq r_j^{\min} \tag{18b}$$

$$\textstyle\sum_{j\in\mathcal{J}}\|\mathbf{w}_{aj}\|_2^2 \leq p_a^{\max} \tag{18c}$$

$$r_j \leq \log(1+\hat{\gamma}_{\tilde{a}j}). \tag{18d}$$

For a known user-AP association, (18d) is SOC representable [18]. If the solution is feasible, use the obtained $\mathbf{w}$ along with association to calculate rest of the initial variables, $(\mathbf{w}^{(0)},\mathbf{u}^{(0)},\mathbf{z}^{(0)})$ by setting the inequalities in the constraint to their corresponding equality where they appear. If it is infeasible, change the association and repeat until a feasible point is achieved.

Main loop:

2: **repeat**

3: Solve $(\mathcal{P}_{(n)})$ and denote an optimal solution as $(\mathbf{w}^*,\mathbf{u}^*,\mathbf{z}^*,\mathbf{r}^*)$.

4: Update $(\mathbf{w}^{(n+1)},\mathbf{u}^{(n+1)},\mathbf{z}^{(n+1)}) = (\mathbf{w}^*,\mathbf{u}^*,\mathbf{z}^*)$ and $n\rightarrow n+1$.

5: **until convergence**

Post-processing:

6: The selected AP for user j is the one associated with the row having the largest l_2 norm. After fixing the user-AP association, it is required to re-run the algorithm to find the actual beamformer values.

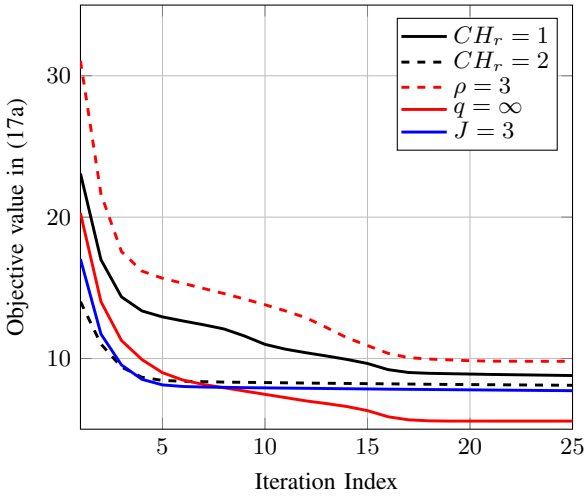

Fig. 2. Convergence property of the Algorithm for $A = T = 3$.

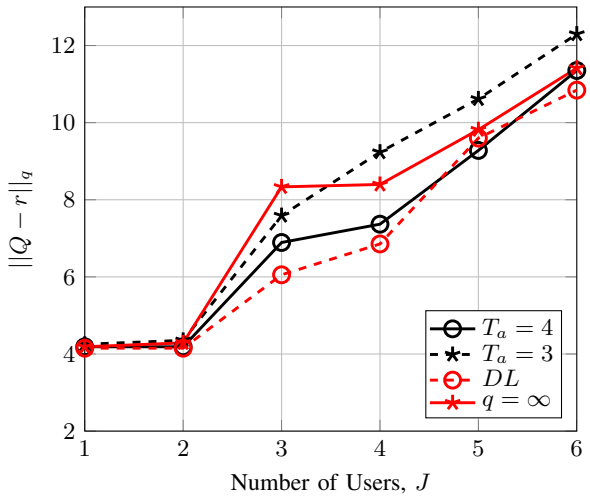

Fig. 3. Final $||\mathbf{Q} - \mathbf{r}||_q$ for different numbers of users with $A = 3$ at $t = 1$s.

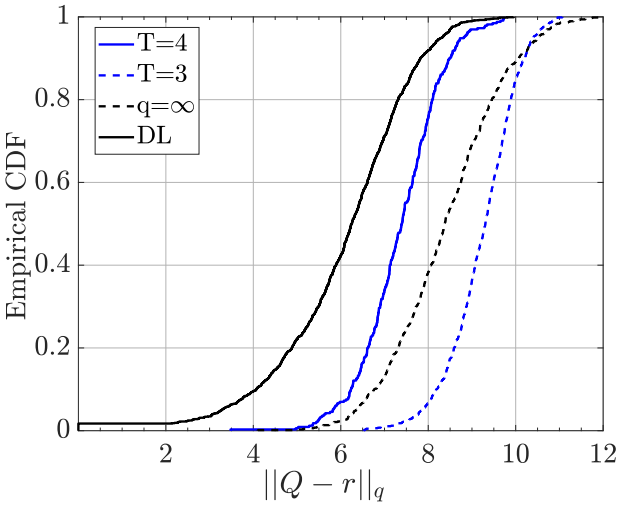

Fig. 4. CDF of the $||\mathbf{Q} - \mathbf{r}||_2$ of the algorithm for $N = 4$.

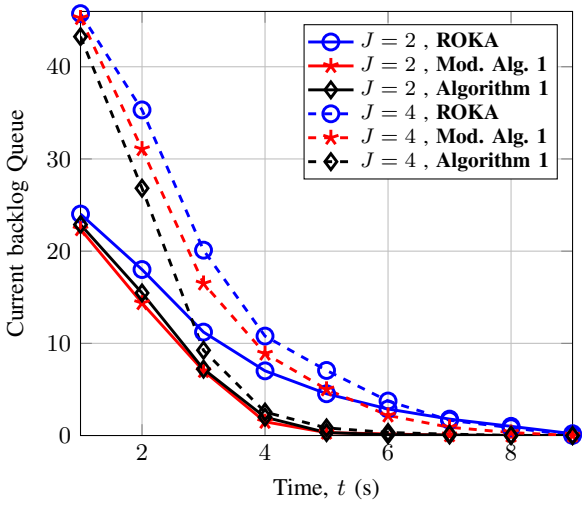

Fig. 5. Comparison of the summation of the current back-log queues in the network of the proposed algorithm with existing algorithms. For $J = 4$, $\mathbf{Q}(2) = [14, 16, 16, 18]$ kB and $\lambda = 1.2$ kB/s/Hz and for $J = 2$, $\mathbf{Q}(3) = [16, 18]$ kB and $\lambda = 2$ kB/s/Hz.

it can be seen that the **Algorithm 1** converges within a few iterations and for the considered network setup, per iteration run time varies between 0.4 and 3.5 seconds.

In Fig. 3, the average optimal output value of $||\mathbf{Q} - \mathbf{r}||_q$ is provided for different network settings with 500 channel realizations. The performance of the algorithm is plotted against the number of users in the network for λ_j = [1.1 1 1.2 1 1.2 1.1] kB/s/Hz with t = 1s time frame. The initial queue is set to $\mathbf{Q}(1)$=[7 7 8 8 9 9] kB for $j \in [1, 2, \ldots, 6]$. In practice, the degree of freedom in finding the best AP reduces with the number of users. Since the users share the available resources, the service rate of individual users may be dropped by increasing the backlog queue of each user. Moreover, when comparing the two black curves in the Fig. 3, we note that more bits can be transmitted with a higher number of transmit antennas, specially when $J > A$. The $q = \infty$ curve illustrates the variation of $||\mathbf{Q} - \mathbf{r}||_\infty$ with $T = 4$. When compared to the two solid curves in Fig. 3, it is clear that $q = 2$ scheme out performs $q = \infty$ scheme. Except the red dashed curve in Fig. 3, all other three curves are obtained for fixed user and AP locations, i.e., path-losses between the users and APs are fixed for all channel realizations. The red dashed curve illustrates the average value of the $||\mathbf{Q} - \mathbf{r}||_2$ over different APs and users' locations (DL) with $T = 4$, i.e., the path-losses between users and APs vary for each channel realization. However, on average, the achieved throughput is nearly equal in both scenarios. Moreover, we consider the $J = 4$ network settings in Fig. 3 and illustrates the CDF of the $||\mathbf{Q} - \mathbf{r}||_q$ in Fig. 4. To compute an empirical CDF of $||\mathbf{Q} - \mathbf{r}||_q$, we consider 10000 channel realizations and obtain the final utility gains to find the variation over considered channel realizations. The achieved rates tend to fluctuate but not much beyond the average and, thus, we can guarantee the efficiency of the algorithm over

perfect channel estimations.

In Fig. 5, we consider two different network settings to benchmark our proposed algorithm against an existing algorithm in [4] for fixed user-AP connections. In this model, for **Algorithm 1**, at $t = 1$s, we obtain the optimal user-AP association and, assume that there is no dynamic changes during the considered time frame. Moreover, there are new arrivals for user j with λ_j arrival rate. Parameters $d_j^{\max}$ and ϵ_j are set to be 4s and 10^{-4}, respectively. i.e., constraint (9) is equal to $\Pr\{\text{delay} \geq 4\text{s}\} \leq 10^{-4}$ for all the users. For our reference, the algorithm in [4] is referred to as resource optimization with known associations (**ROKA**). Obviously, we can modify our algorithm to optimize resources for a known user-AP association scenario and we refer to the variant as modified **Algorithm 1 (Mod. Alg. 1)**. For the first scenario, the numbers of users and APs are all set to be 2 and the data arrival rate for any user is considered as 2 kB/s/Hz. Under this setup, user-AP association for **Mod. Alg. 1** and **ROKA** is obtained with an exhaustive search; i.e., globally optimal association has considered in these two methods. According to the numerical results, our proposed algorithm transmit 98.67% of the backlog queue when $t = 4$s and it is similar to **Mod. Alg. 1**. This implies that the user-AP association of our algorithm is very close to the optimal selection and in terms of latency only 1.2% of the queue is remain to transmit when $t = 4$s. Furthermore, algorithm **ROKA** transmit only 70% of the backlog queue when $t = 4$s. However, the brute-force algorithms are inefficient in practice and, thus, the PL based method is considered in **Mod. Alg. 1** and **ROKA** for the second set of simulations. The PL based method is where a user is associated with the AP of smallest path loss. For this comparison, we consider the network with $J = 4$, $A = 3$ and 1.2kB/s/Hz arrival rates. It can be seen that the proposed algorithm transmits 98% of the queue within 4ms, while **Mod. Alg. 1** and **ROKA** transmit only 86.9% and 81% of the backlog queues, respectively. Moreover, Fig. 5 indicates that **ROKA** in both settings and path loss based **Mod. Alg. 1** require $t = 8$s and $t = 7$s service times to transmit 98% of the queue compared to the shorter service time of $t = 4$s needed with **Algorithm 1**. In this view, our algorithm yields up to two-fold reductions in the latency over **Mod. Alg. 1** and **ROKA**.

V. Conclusion

We proposed an algorithm based on clustering to optimize queue weighted sum rate and user-AP association to ensure latency in dynamic networks. According to the provided numerical results, our proposed algorithm is fast converging and the ℓ_2-norm has a higher impact on reducing the network latency, compared to ℓ_∞-norm. For small numbers of users and APs, the user-AP association of the proposed algorithm is very close to the optimal user-AP selection. We numerically demonstrate that **Algorithm 1** can be up to two-fold efficient than the known solutions in terms of network latency.

References

[1] I. Sugathapala, L.-N. Tran, M. F. Hanif, B. Lorenzo, S. Glisic, and M. Juntti, "SOCP based joint throughput maximization and user association in dynamic networks," in *Proc. IEEE ICC Workshop on MIMO and cognitive radio technologies in multihop network (MIMOCR)*, London, UK, June 2016, pp. 573–578.

[2] B. Lorenzo, F. J. Gonzalez-Castano, and Y. Fang, "A novel collaborative cognitive dynamic network architecture," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 74 – 81, 2017.

[3] T. K. Vu, C.-F. Liu, M. Bennis, M. Debbah, M. Latva-aho, and C. S. Hong, "Ultra-reliable and low latency communication in mmWave-enabled massive MIMO networks," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2041–2044, 2017.

[4] G. Venkatraman, A. Tolli, M. Juntti, and L.-N. Tran, "Traffic aware resource allocation schemes for multi-cell MIMO-OFDM systems," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2730–2745, 2016.

[5] N. Funabiki, J. Shimizu, T. Nakanishi, and K. Watanabe, "A proposal of an active access-point selection algorithm in wireless mesh networks," in *Proc. International Conference on Network-Based Information Systems (NBiS)*, Tirana, Albania, Sep. 2011, pp. 112–117.

[6] S. Hanly, "An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1332–1340, 1995.

[7] R. D. Yates and C.-Y. Huang, "Integrated power control and base station assignment," *IEEE Trans. Veh. Technol.*, vol. 44, no. 3, pp. 638–644, 1995.

[8] M. Hong, A. Garcia, J. Barrera, and S. G. Wilson, "Joint access point selection and power allocation for uplink wireless networks," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3334–3347, 2013.

[9] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless Commun. Mag.*, vol. 21, no. 2, pp. 18–25, 2014.

[10] S. You, L. Chen, and Y. E. Liu, "Convex-concave procedure for weighted sum-rate maximization in a MIMO interference network," in *Proc. IEEE Global Communications Conference*, Dec. 2014, pp. 4060 – 4065.

[11] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *Journ. of Global Optimization*, vol. 47, no. 1, pp. 29–51, 2010.

[12] E. Karami and S. Glisic, "Self-management of mobile clouds in advanced wireless networks," in *Proc. IEEE Network Operations and Management Symposium (NOMS)*, Maui, HI, USA, Apr. 2012, pp. 1054–1060.

[13] Q.-D. Vu, L.-N. Tran, M. Juntti, and E.-K. Hong, "Energy-efficient bandwidth and power allocation for multi-homing networks," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1684–1699, 2015.

[14] J. Chen and V. K. N. Lau, "Large deviation delay analysis of queue-aware multi-user MIMO systems with multi-timescale mobile-driven feedback," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4067–4076, 2013.

[15] J. D. C. Little and S. C. Graves, "Little's Law," in *Proc. Building Intuition: Insights From Basic Operations Management Models and Principles*, 2008, pp. 81–100.

[16] A. Mukherjee, "Queue-aware dynamic on/off switching of small cells in dense heterogeneous networks," in *Proc. IEEE Global Commun. conf. workshop*, Dec. 2013, pp. 182–187.

[17] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra and its Applications*, vol. 284, no. 1-3, pp. 193–228, 1998.

[18] O. Tervo, L.-N. Tran, and M. Juntti, "Energy-efficient transmit beamforming for MISO downlink via sequential convex approximation," in *International Workshop on Signal Processing Advances in Wireless Communications*, June 2015, pp. 415–419.

[19] S. He, Y. Huang, S. Jin, F. Yu, and L. Yang, "Max-min energy efficient beamforming for multicell multiuser joint transmission systems," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1956–1959, 2013.

[20] J. F. Sturm, "Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones," 1998.

[21] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, 2014.