# Load Balancing in Small-Cell Access Point Placement

Govind R. Gopal*, Bhaskar D. Rao*, and Gabriel Porto Villardi†
*Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, United States
†Wireless Network Research Center, National Institute of Information and Communications Technology, Yokosuka, Japan
*ggopal@ucsd.edu, *brao@ucsd.edu, †gpvillardi@nict.go.jp

***Abstract*—We address the uplink small-cell access point (AP) placement problem for optimal throughput, while considering load balancing (LB) among the APs. To consider LB and consequently incorporate fairness in user spectral access, i.e., the frequency of user-to-AP communications, we modify the Lloyd algorithm from vector quantization so that delays incurred by the existence of a large number of users in a cell are accounted for in the AP placement process. Accordingly, we present two methods, the first of which involves the incorporation of weights proportional to the cell occupancy, hence called the Occupancy Weighted Lloyd algorithm (OWLA). The second method adds a new step to the Lloyd algorithm, which involves re-assigning users from higher to lower occupancy cells, and the adoption of a distance threshold to cap the throughput lost in the assignment process. This formulated Lloyd-type algorithm is called the Cell Equalized Lloyd Algorithm-$\alpha$ (CELA-$\alpha$) where $\alpha$ is a factor that allows for throughput and spectrum access delay trade-off. Extensive simulations show that both CELA-$\alpha$ and OWLA algorithms provide significant gains, in comparison to the standard Lloyd algorithm, in 95%-likely user spectral access. For the $\alpha$ values considered in this paper, CELA-$\alpha$ achieves gains up to 20.83%, while OWLA yields a gain of 12.5%. Both algorithms incur minimal throughput losses of different degrees, and the choice of using one algorithm over the other for AP placement depends on system LB as well as throughput requirements.**

***Index Terms*—Base station placement, Beyond 5G, fairness, Lloyd algorithm, throughput optimization, user cell association.**

## I. Introduction

In the recent decades, massive multiple-input multiple-output (MIMO) has attracted much attention since it will enable higher throughput and will be an integral part of 5G and Beyond systems [1], [2]. Especially, distributed antenna systems (DASs), comprising distributed MIMO are popular as they enable higher average rates over co-located systems. DASs constitute small-cell (non-cooperative) and cell-free (co-operative) systems, with the latter providing higher throughput, but requiring enhanced backhaul [3]. Small-cell systems are thus still favored in practice [4]. To further improve system throughput, a degree of freedom that can also be exploited is access point (AP) placement, especially with the advent of unmanned aerial vehicle (UAV) based non-terrestrial networks. This bears the question: *How to optimally place the APs given the distributions of users?* Such endeavors are even more useful in situations where user densities change over time, e.g., conferences and stadiums, and in emergency scenarios where existing APs have been destroyed. Previous work on AP placement have involved the works [5]–[10] and references therein. The standard Lloyd algorithm (with squared Euclidean distance) from vector quantization (VQ) is used to solve the DAS ergodic capacity maximization problem in [5]. Average rate is maximized in [6] to place circularly arranged antennas. Deployment of UAVs for increased throughput has also garnered interest [7], [8]. A recent work by our group [9] adds inter-cell interference (ICI) to the throughput optimization problem and proposes Lloyd-type algorithms to yield improvements in 95%-likely rates over the standard Lloyd algorithm that maximizes SNR alone.

Besides throughput, delay in spectrum access is also a relevant system design parameter, especially for the delay sensitive applications expected in the deployment of Beyond 5G networks. In systems where throughput alone is optimized, either by using the Lloyd algorithm for SNR optimization only or the Lloyd-type algorithm by incorporating ICI in the problem formulation, the cells have unequal occupancies, that is to say, unequal number of users, after the algorithm converges, thus yielding placement with sub-optimal fairness in spectral usage. In other words, this results in an unbalanced distribution of users across the cells, which in turn leads to users of cells with lower occupancy having more opportunities to access the spectrum over users of other cells. Naturally, the question that arises is: *How to efficiently perform user-cell association so that users are ensured opportunity to access spectrum without undue delay?* Defining spectrum access delay as the time that a user waits for its opportunity to communicate with its assigned AP, we create a metric called *spectral access fraction* in order to allow us to quantify the access delay of the proposed algorithms. One approach [10] is to equalize the occupancy of each cell by re-assigning users from cells with higher-than-average occupancy (among all cells), to cells with lower-than-average occupancy. The objective of this procedure is balancing the cell loads, following the motivation behind cell breathing [11]–[14]. Nevertheless, the main drawback of this strategy is that users can be moved to far away cells and therefore suffer a significant reduction in overall throughput. To this end, our work aims to create a desirable and flexible trade-off between throughput reduction and increase in the spectral access fraction. We therefore devise algorithms for non-cooperative small-cell systems that modify the Lloyd algorithm yielding AP placements that maximize throughput while minimizing spectrum access delay, thereby considering load balancing (LB) among the small-cells.

The work of Govind R. Gopal and Bhaskar D. Rao was supported in part by the Center for Wireless Communications (CWC), University of California San Diego, in part by Qualcomm Inc. through the Faculty-Mentor-Advisor program, and in part by the National Science Foundation (NSF) under Grant CCF-2124929. The work of Gabriel Porto Villardi was partly carried out when he was a visiting scholar with the Qualcomm Institute of Calit2, University of California San Diego, La Jolla, CA 92093 USA.

*Contributions.* To the best of our knowledge, solutions to the small-cell AP placement problem based on the Lloyd algorithm and that jointly address throughput and spectrum access delay (incorporating LB), have not been provided in literature. Hence, in this work, our contributions are as follows.

- The Lloyd algorithm from VQ is modified to incorporate weights chosen to prevent users from associating with APs having a large occupancy. This weighted Lloyd algorithm is hereafter referred to as the Occupancy Weighted Lloyd Algorithm (OWLA), and considers LB and throughput altogether.
- An alternate LB procedure that re-assigns users between cells is proposed. By prioritizing and re-assigning users from higher to lower occupancy cells, the joint effect of throughput and delay is addressed. Moreover, in order to control the trade-off between throughput and spectrum access delay, the distance threshold used incorporates a factor $\alpha$. The Lloyd-type algorithm created is called the Cell Equalized Lloyd Algorithm-$\alpha$ (CELA-$\alpha$).

## II. System Model

We use the small-cell system model outlined in [9], [10], [15], [16]. $K$ single-antenna users are distributed with a probability density function $f_{\mathbf{P}}(\mathbf{p})$, with $\mathbf{p} \in \mathbb{R}^2$ denoting the user position, over a geographical area. $M$ single-antenna APs serve these users, where $\mathbf{q} \in \mathbb{R}^2$ is the AP location. A narrowband user-AP fading channel is considered with $m = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, K$ as $g_{mk} = \sqrt{\beta_{mk}} h_{mk}$, where $\beta_{mk}$ and $h_{mk} \sim \mathcal{CN}(0, 1)$ are the large- and small-scale fading coefficients, respectively, independent of each other and over coherent intervals. All APs are connected to a network controller via error-free backhaul links, which knows all AP and user positions. We assume an uplink model where each AP serves a subset of the users (cell $\mathcal{C}_m$ for AP $m$). Users are scheduled in a round robin fashion according to time-division multiple access (TDMA), and each AP serves only one user in a time slot. The received signal at AP $m$ with $k_m$ denoting the user index associated with AP $m$ is

$$y_m = \sum_{m'=1}^{M} \sqrt{\rho_r} g_{mk_{m'}} s_{k_{m'}} + w_m, \tag{1}$$

where $\rho_r$ is the transmit power, $s_{k_m}$ is the data symbol with $\mathbb{E}\{|s_{k_m}|^2\} = 1$, and $w_m \sim \mathcal{CN}(0, 1)$ is the additive noise. The data symbol $s_{k_m}$ as estimated by a matched filter at AP $m$ is $\hat{s}_{k_m} = (g^*_{mk_m}/|g_{mk_m}|) y_m$. The signal-to-interference-plus-noise ratio (SINR) achieved by user $k_m$ at AP $m$ is then

$$\phi_{k_m} = \frac{\rho_r \beta_{mk_m} |h_{mk_m}|^2}{1 + \rho_r \sum\limits_{\substack{m'=1 \\ m' \neq m}}^{M} \beta_{mk_{m'}} |h_{mk_{m'}}|^2}. \tag{2}$$

## III. Vector Quantization and Lloyd Algorithm

We review VQ and the Lloyd algorithm as applied to small-cell AP placement. The VQ framework considers a random vector (position of a single user $\mathbf{p}$) that is quantized using an encoder and decoder. The encoder splits the domain (geographical area) into Voronoi regions (cells), which are each assigned a codepoint (AP location) by the decoder. To optimize the quantizer, the Lloyd algorithm alternates between finding the regions keeping the codepoints fixed, called the Nearest Neighbor Condition (NNC), and finding the codepoints keeping the regions fixed, called the Centroid Condition (CC). The optimization problem minimizes the mean squared error $\mathbb{E}_{\mathbf{p}}\left\{d_{\text{SE}}(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})})\right\}$, where $d_{\text{SE}}(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})}) = ||\mathbf{p} - \mathbf{q}_{\mathcal{E}(\mathbf{p})}||^2$ is the squared Euclidean distortion measure, $\mathcal{E}$ is the encoder, and $\mathcal{E}(\mathbf{p})$ denotes the index of the cell of user at $\mathbf{p}$. The Lloyd algorithm for AP placement can be found in [9].

## IV. Accounting for Load Balancing

As mentioned in Section I, the Lloyd and Lloyd-type algorithms (in [9]) for throughput optimality result in unequal cell occupancies, where users of cells with a lower occupancy would unfairly get more opportunities to access the spectrum than users of cells with a higher occupancy. Hence, when user delay is measured by the frequency of user-to-AP access (the *spectral access fraction*), these algorithms result in significantly varied spectral access profiles. For delay sensitive applications, LB capabilities are required in order to achieve a certain application-dependent degree of similarity in spectral access profile for all users. This means that we should strive to equalize the occupancy of each cell, the degree of which is determined by the specific application. Below, we outline the OWLA and CELA-$\alpha$ AP placement algorithms, both generating Lloyd-type algorithms that improve system fairness, with CELA-$\alpha$ having a flexible throughput-delay trade-off adaptable to application requirements.

**Algorithm 1** OWLA

1: Initialize random AP locations $\mathbf{q}_m^{(0)}$, $\forall m$.
2: Use the NNC to determine the cells $\mathcal{C}_m^{(i+1)}$, $\forall m$

$$\mathcal{C}_m^{(i+1)} = \left\{\mathbf{p}_k : w_m^{(i)} d_{\text{SE}}\left(\mathbf{p}_k, \mathbf{q}_m^{(i)}\right) \leq w_l^{(i)} d_{\text{SE}}\left(\mathbf{p}_k, \mathbf{q}_l^{(i)}\right), \forall l \neq m\right\}.$$

3: Use the CC to determine the AP locations $\mathbf{q}_m^{(i+1)}$, $\forall m$

$$\mathbf{q}_m^{(i+1)} = \frac{1}{\left|\mathcal{C}_m^{(i+1)}\right|} \sum_{\mathbf{p}_k \in \mathcal{C}_m^{(i+1)}} \mathbf{p}_k.$$

4: Repeat from step 2 until convergence.

### A. Occupancy Weighted Lloyd Algorithm (OWLA)

The access rate using per-user SNR $\psi_{k_{\mathcal{E}(\mathbf{p})}}$, which can be obtained by neglecting ICI from SINR $\phi_{k_{\mathcal{E}(\mathbf{p})}}$ in (2), is

$$R^{\text{acc}}_{k_{\mathcal{E}(\mathbf{p})}} = \frac{1}{N_m} \mathbb{E}\left\{\log_2\left(1 + \psi_{k_{\mathcal{E}(\mathbf{p})}}\right)\right\}, \tag{3}$$

where $\mathcal{E}(\mathbf{p})$ indexes the AP closest to user $k$ and $N_m$ is the number of users in cell $\mathcal{C}_m$. Note that the achievable rate $R_{k_{\mathcal{E}(\mathbf{p})}} = \mathbb{E}\left\{\log_2\left(1 + \phi_{k_{\mathcal{E}(\mathbf{p})}}\right)\right\}$ does not account for the delay incurred by a user as it waits to transmit to its AP with TDMA scheduling. Therefore, the rate is normalized using

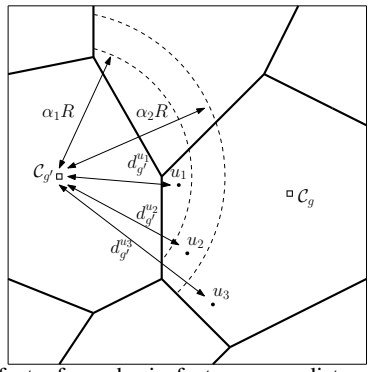


Fig. 1: The effect of emphasis factor $\alpha$ on distance thresholds in CELA-$\alpha$. In this case, $R$ is the communications radius for cell $\mathcal{C}'_g$.

the resource sharing factor $1/N_m$. The logarithm in (3) can be well fitted with a third degree polynomial $\log_2(1+x) \approx a_1 + a_2 x + a_3 x^2 + a_4 x^3$, and therefore (3) can be rewritten as

$$R^{\text{acc}}_{k_{\mathcal{E}(\mathbf{p})}} = \frac{1}{N_m}\mathbb{E}\left\{a_1 + a_2\psi_{k_{\mathcal{E}(\mathbf{p})}} + a_3\psi^2_{k_{\mathcal{E}(\mathbf{p})}} + a_4\psi^3_{k_{\mathcal{E}(\mathbf{p})}}\right\}. \quad (4)$$

Focusing first on the linear term $a_2 x$, we can write

$$\mathbb{E}_{\mathcal{A},\mathbf{p}}\left\{\frac{a_2}{N_m}\psi_{k_{\mathcal{E}(\mathbf{p})}}\right\} \geq \mathbb{E}_{\mathcal{A}}\left\{\frac{a_2\rho_r c_1 |h_{\mathcal{E}(\mathbf{p})}|^2 z_{\mathcal{E}(\mathbf{p})}}{\left(\mathbb{E}_{\mathbf{p}}\left\{N_m^{\frac{2}{\gamma}}\left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^2\right\}\right)^{\frac{\gamma}{2}}}\right\}, \quad (5)$$

using Jensen's inequality, where $z_{\mathcal{E}(\mathbf{p})}$ is the shadow fading coefficient, $c_1$ is the pathloss constant, $\gamma$ is the pathloss exponent, $\beta_{\mathcal{E}(\mathbf{p})}$ in $\psi_{k_{\mathcal{E}(\mathbf{p})}}$ approximated as in [9], and $\mathcal{A} = \{h_{\mathcal{E}(\mathbf{p})}, z_{\mathcal{E}(\mathbf{p})}\}$. Note that we wish to maximize (4) and correspondingly minimize the denominator in (5). Following the VQ framework, the distortion function for the linear term is $d\left(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right) = N_m^{2/\gamma}\left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^2$. Applying the same technique to all terms in (4), the new distortion function is defined by summing the distortion from each term

$$d\left(\mathbf{p}, \mathbf{q}_{\mathcal{E}(\mathbf{p})}\right) = \left(N_m^{\frac{2}{\gamma}} + N_m^{\frac{2}{2\gamma}} + N_m^{\frac{2}{3\gamma}}\right)\left|\left|\mathbf{p}-\mathbf{q}_{\mathcal{E}(\mathbf{p})}\right|\right|^2, \quad (6)$$

where the squared Euclidean distance measure of the standard Lloyd algorithm has been multiplied with a weight $w_m = N_m^{2/\gamma} + N_m^{2/2\gamma} + N_m^{2/3\gamma}$. Also, note that curve fitting in (3) with a higher order polynomial is unnecessary since the growth of the weight $w_m$ diminishes as the polynomial order grows. The Lloyd-type algorithm corresponding to this weighted distortion measure is termed OWLA and is outlined in Algorithm 1.

### *B. Cell Equalized Lloyd Algorithm-$\alpha$ (CELA-$\alpha$)*

In the previous work [10], a hard criterion of equal user access was set, resulting in all cells ending up with the same occupancy, however, causing significant throughput loss due to users being re-assigned to far away cells. Further, a distance threshold to prevent such undesirable user re-assignments was introduced, however, having two main drawbacks in its re-assignment procedure. Firstly, only the distortion values are considered to re-assign users. That is, although the cell occupancies are checked prior to user re-assignment, they are not used in the decision of the order in which users are re-assigned to other cells. Thus, there is a need to jointly consider both distortion and cell occupancies in this decision process. Secondly, depending on system requirements, a trade-off between delay and throughput might be necessary. Therefore, we next address the above needs, leading to a more comprehensive algorithm called CELA-$\alpha$, where $\alpha$ is the trade-off factor allowing flexibility between throughput and delay.

*1) Allowing throughput and delay trade-off:* By pre-multiplying the abovementioned cell-specific distance thresholds, i.e., $R^{\text{th}}_m$, with the trade-off factor $\alpha$, then the new threshold becomes $\alpha R^{\text{th}}_m$. We know that if the distance threshold is set to 0, i.e., $R^{\text{th}}_m = 0$, then the algorithm then becomes the standard Lloyd algorithm. On the other hand, high distance thresholds would enable completely equal user access due to the equal occupancy in each cell, but it would result in reduced throughput owing to the large distances between select users and their APs. Here, $\alpha$ enables us to adjust the threshold between these two extremes. In Fig. 1, we illustrate the discussions about CELA-$\alpha$. Two cells $\mathcal{C}_g$ and $\mathcal{C}_{g'}$ are shown along with three users $u_1$, $u_2$, and $u_3$ in cell $\mathcal{C}_g$. Assume that the three users in $\mathcal{C}_g$ with excess of users are to be moved to $\mathcal{C}_{g'}$ with low occupancy. For simplicity, $R^{\text{th}}_{g'} = R$, the communication radius of the AP in $\mathcal{C}_{g'}$ and the distance between the three users and $\mathcal{C}_{g'}$ are $d^{u_1}_{g'}$, $d^{u_2}_{g'}$, and $d^{u_3}_{g'}$, respectively; $\alpha_1$ and $\alpha_2$ represent two trade-off factors. Under $\alpha_1$, since $d^{u_1}_{g'} < \alpha_1 R < d^{u_2}_{g'} < d^{u_3}_{g'}$, user $u_1$ will be moved to $\mathcal{C}_{g'}$ while users $u_2$ and $u_3$ will remain in $\mathcal{C}_g$. On the other hand, under $\alpha_2$, we have $d^{u_1}_{g'} < d^{u_2}_{g'} < \alpha_2 R < d^{u_3}_{g'}$, which implies that users $u_1$ and $u_2$ will be moved to $\mathcal{C}_{g'}$ while user $u_3$ will remain in $\mathcal{C}_g$. Hence, it is evident that if the value of the trade-off factor $\alpha$ is increased, more users are re-assigned to $\mathcal{C}_{g'}$ and hence lesser spectrum access delay is obtained in $\mathcal{C}_g$ at the expense of some overall throughput loss.

*2) Addition of cell occupancy to re-assignment:* As mentioned above, considering the joint influence of distance and cell occupancy in order to determine the priority with which to re-assign users to other cells would involve updating the algorithm as follows. For each cell with excess users, the distance between the users in the cell and all other APs is multiplied with the occupancy of the corresponding cell and the user with the lowest such value is considered first.

The complete CELA-$\alpha$ detailing the above two modifications is provided in Algorithm 2. Note that the vector $\overline{\mathbf{v}}_{u_g}$ represents the ordered set of cells which should be followed when re-assigning the user $u_g$. On the other hand, vector $\mathbf{y}_g$ provides the order in which each user in cell $\mathcal{C}_g$ has to be re-assigned to its respective cell $\mathcal{C}_{g'}$.

## V. Simulation Results

The simulation setup consists of $M = 8$ APs and $K = 2000$ users over a $2\times 2$ km$^2$ area. The user distribution is a Gaussian mixture model (GMM), whose parameters are taken from [10]. Also, the pathloss model and parameters are obtained from [9]. One random user from each cell transmits to its serving AP

**Algorithm 2** CELA-$\alpha$

1: Initialize random AP locations $\mathbf{q}_m^{(0)}$, $\forall m$.

2: Use the NNC to determine the cells $\mathcal{C}_m^{(i+1)}$, $\forall m$

$$\mathcal{C}_m^{(i+1)} = \left\{\mathbf{p}_k : d_{\text{SE}}\left(\mathbf{p}_k, \mathbf{q}_m^{(i)}\right) \leq d_{\text{SE}}\left(\mathbf{p}_k, \mathbf{q}_l^{(i)}\right), \forall l \neq m\right\}.$$

3: Perform the re-assignment procedure:

3.1: Find all the cells that have number of users $> N$ and arrange them in descending order. Let the ordered set of cells generated be $\mathcal{C}^G$.

3.2: Iterate through the cells in $\mathcal{C}^G$ and perform the following process for each cell $\mathcal{C}_g \in \mathcal{C}^G$:

a: For each user $u_g$ associated with $\mathcal{C}_g$, generate a vector $\mathbf{v}_{u_g}$ containing distances to *all other* APs. Multiply each element of $\mathbf{v}_{u_g}$ with the occupancy of the corresponding cell. Arrange these composite values (product of user-AP distance and cell occupancy) in ascending order within the vector to generate $\overline{\mathbf{v}}_{u_g}$.

b: Take the first element of all vectors $\overline{\mathbf{v}}_{u_g}$, $\forall u_g \in \mathcal{C}_g$ and sort them in a new vector $\mathbf{y}_g$ in ascending order.

c: Iterate through the elements of $\mathbf{y}_g$ and for both corresponding user $u_g$ and cell $\mathcal{C}_{g'}$, the two following conditions have to be met to allow user $u_g$ to be assigned to cell $\mathcal{C}_{g'}$:

- Occupancy of cell $\mathcal{C}_{g'}$, $N_{g'} < N$
- User-AP distance for cell $\mathcal{C}_{g'}$, $d(\mathbf{p}_{u_g}, \mathbf{q}_{g'}) < \alpha R_{g'}^{\text{th}}$

If either condition is not satisfied, $u_g$ is not re-assigned to $\mathcal{C}_{g'}$ and remains in $\mathcal{C}_g$.

d: Once all elements of $\mathbf{y}_g$ are considered, use the next (second, third, ...) element of every vector $\overline{\mathbf{v}}_{u_g}$ of users who have not been re-assigned and repeat from step 3.2-b.

3.3: Repeat from step 3.2-a for the next cell in $\mathcal{C}^G$ until all cells have been considered.

4: Use the CC to determine the AP locations $\mathbf{q}_m^{(i+1)}$, $\forall m$

$$\mathbf{q}_m^{(i+1)} = \frac{1}{\left|\mathcal{C}_m^{(i+1)}\right|} \sum_{\mathbf{p}_k \in \mathcal{C}_m^{(i+1)}} \mathbf{p}_k.$$

5: Repeat from step 2 until convergence.

with power $\rho_r = 200$ mW. The placement algorithms are identically initialized for unbiased comparison. For performance evaluation of the presented algorithms, we use the following *per-user* metrics: (a) *access rate* $R_{k_m}^{\text{acc}}$, defined in (3), using $\phi_{k_m}$ from (2), and (b) *spectral access fraction* $U_{k_m} = 1/N_m$, which is a measure of the frequency with which user $k_m$ communicates with its serving AP $m$.

In our numerical simulations, we obtain the AP placements, access rates, and the spectral access fractions of the LB-aware OWLA and CELA-$\alpha$, and compare them to those of the Lloyd algorithm. In CELA-$\alpha$, three trade-off values $\alpha = 0.9$, 1, and 1.75 are used. We choose the threshold as the distance of the AP to its nearest AP (cell-specific distance threshold). The AP locations obtained after the algorithms converge are

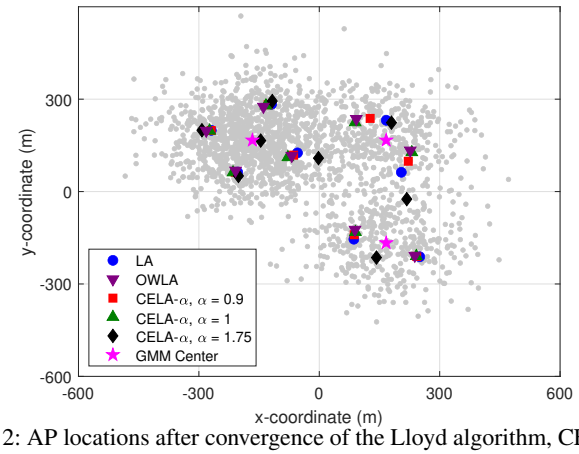


Fig. 2: AP locations after convergence of the Lloyd algorithm, CELA-$\alpha$ with $\alpha = 0.9$, 1, 1.75, and OWLA for $M = 8$ and GMM-2.

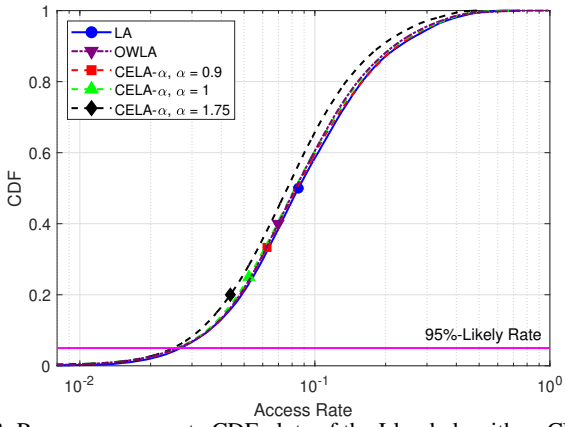


Fig. 3: Per-user access rate CDF plots of the Lloyd algorithm, CELA-$\alpha$ with $\alpha = 0.9$, 1, 1.75, and OWLA for $M = 8$, and GMM-2.

shown in Fig. 2. OWLA results in APs that are placed away from those of the Lloyd algorithm, due to the LB-promoting weighted distortion measure. From CELA-$\alpha$, we know that a higher value of $\alpha$ results in more user re-assignments. This is also evidenced by the fact that the AP locations are more different from those of the Lloyd algorithm as $\alpha$ increases. In order to quantitatively show the degree to which CELA-$\alpha$ perform user re-assignments, the occupancy of every cell for each of the considered $\alpha$ is provided in Table I. We observe that while the occupancy level for equal occupancy would be $2000/8 = 250$, the occupancy levels vary significantly for the Lloyd algorithm. For smaller values of $\alpha$, i.e., $\alpha = 0.9$, it is observed that those cells with occupancy higher than the target in the Lloyd algorithm have their occupancy lowered. The opposite effect occurs for larger $\alpha$ values. As more users are re-assigned with the increase in $\alpha$, more cells are able to attain the target value of 250 users. Particularly, for $\alpha = 1$, two cells and for $\alpha = 1.75$, an additional three cells attain this target occupancy. The occupancy observed in OWLA is also shown in the Table I and similarly to CELA-$\alpha$, OWLA mitigates the issue of unbalanced loads in the cells.

Next, we show in Fig. 3 the cumulative distribution functions (CDFs) of the access rate. The rate curve corresponding to OWLA is observed to have a slightly inferior performance than that of the Lloyd algorithm. Also, as expected, when

TABLE I: Cell Occupancy of LA, CELA-$\alpha$, and OWLA

| Algorithm | $\mathcal{C}_1$ | $\mathcal{C}_2$ | $\mathcal{C}_3$ | $\mathcal{C}_4$ | $\mathcal{C}_5$ | $\mathcal{C}_6$ | $\mathcal{C}_7$ | $\mathcal{C}_8$ |
|---|---|---|---|---|---|---|---|---|
| LA | 282 | 278 | 327 | 337 | 236 | 180 | 157 | 203 |
| $\alpha = 0.9$ | 266 | 269 | 321 | 331 | 235 | 198 | 174 | 206 |
| $\alpha = 1$ | 250 | 277 | 302 | 314 | 250 | 213 | 179 | 215 |
| $\alpha = 1.75$ | 250 | 250 | 264 | 262 | 250 | 250 | 224 | 250 |
| OWLA | 281 | 288 | 302 | 286 | 221 | 219 | 189 | 214 |

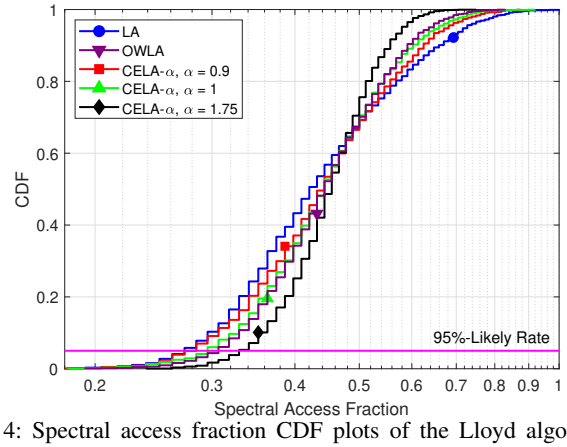


Fig. 4: Spectral access fraction CDF plots of the Lloyd algorithm, CELA-$\alpha$ with $\alpha = 0.9$, 1, 1.75, and OWLA for $M = 8$, and GMM-2

TABLE II: Percentage Improvements in 95%-Likely Achievable and Access Rates and Spectral Access Fraction for CELA-$\alpha$ and OWLA

| Algorithm | Access Rate | Spectral Access Fraction |
|---|---|---|
| $\alpha = 0.9$ | -1.46% | 4.17% |
| $\alpha = 1$ | -2.96% | 8.33% |
| $\alpha = 1.75$ | -6.96% | 20.83% |
| OWLA | -2.28% | 12.5% |

$\alpha$ increases to prioritize spectral access fairness among users within the cells, some degree of throughput loss is observed. Notice that the worst rate loss happens when $\alpha = 1.75$. Finally, Fig. 4 shows the CDFs of the spectral access fractions for OWLA and for the three values of $\alpha$ in CELA-$\alpha$ under consideration. It can be seen that this metric increases significantly with $\alpha$ and OWLA provides a performance slightly higher to that of $\alpha = 1$. Table II shows the percentage improvements for the access rate and spectral access fraction. Although rate reduction is observed to a degree, there is a significant improvement in spectrum access fraction. That is, the magnitude of access increase is much higher than that of the rate decrease. For instance, while the access rate suffers a reduction of 6.96%, the increase in spectral access fraction is nearly three-fold of that amount, at 20.83%, for $\alpha = 1.75$. Additionally, a key observation is that OWLA is able to achieve a higher spectral access fraction improvement at the cost of a lower access rate decrease, compared to CELA-$\alpha$ when $\alpha = 1$. As such, one should favor OWLA over CELA-$\alpha$ if a spectral access fraction of up to 12.5% above the one provided by Lloyd algorithm is required, for the considered user configuration. Above this mark, CELA-$\alpha$ is to be preferred. Finally, CELA-$\alpha$ has an inherent flexibility that enables the performance of the system to be governed by the trade-off factor $\alpha$ which can be based on specific system requirements, which OWLA cannot provide.

## VI. Conclusion

In this work, we have addressed the aspect of load balancing (LB) in throughput optimal small-cell access point placement. To account for LB in the placement process, we modified the Lloyd algorithm from vector quantization and presented two methods, namely the Occupancy Weighted Lloyd algorithm (OWLA) and Cell Equalized Lloyd algorithm-$\alpha$ (CELA-$\alpha$), both of which yield increases in user access with minimal throughput loss. While OWLA utilizes a weighted distortion function, CELA-$\alpha$ adds an additional step in the Lloyd framework to achieve a degree of LB. Results show that both proposed algorithms achieve higher spectral access (up to around 21%) while suffering a relatively minor reduction in throughput (up to around 7%), and CELA-$\alpha$, through its trade-off factor $\alpha$, allows for flexibility in deciding the degree of LB.

## References


[1] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[2] R. Chataut and R. Akl, "Massive MIMO systems for 5G and beyond networks — overview, recent trends, challenges, and future research direction," in *Sensors*, vol. 20, no. 10, May 2020, Art. ID 2753.

[3] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[4] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5G evolution: A view on 5G cellular technology beyond 3GPP release 15," *IEEE Access*, vol. 7, pp. 127 639–127 651, Sept. 2019.

[5] X. Wang, P. Zhu, and M. Chen, "Antenna location design for generalized distributed antenna systems," *IEEE Commun. Lett.*, vol. 13, no. 5, pp. 315–317, May 2009.

[6] E. Park, S. Lee, and I. Lee, "Antenna placement optimization for distributed antenna systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 7, pp. 2468–2477, July 2012.

[7] C. Lai, C. Chen, and L. Wang, "On-demand density-aware UAV base station 3D placement for arbitrarily distributed users with guaranteed data rates," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 913–916, June 2019.

[8] J. Guo, P. Walk, and H. Jafarkhani, "Optimal deployments of UAVs with directional antennas for a power-efficient coverage," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5159–5174, Aug. 2020.

[9] G. R. Gopal, E. Nayebi, G. P. Villardi, and B. D. Rao, "Modified vector quantization for small-cell access point placement with inter-cell interference," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6387–6401, Aug. 2022.

[10] G. R. Gopal and B. D. Rao, "Throughput and delay driven access point placement," in *Proc. 2019 53rd Asilomar Conf. Signals, Syst., Comput.*, Nov. 2019, pp. 1010–1014.

[11] A. Jalali, "On cell breathing in CDMA networks," in *Proc. 1998 IEEE Int. Conf. Commun. (ICC)*, vol. 2, June 1998, pp. 985–988.

[12] T.-C. Tsai and C.-F. Lien, "IEEE 802.11 hot spot load balance and QoS-maintained seamless roaming," in *Proc. Nat. Comput. Symp. (NCS)*, Jan. 2003.

[13] Y. Bejerano and S. Han, "Cell breathing techniques for load balancing in wireless LANs," *IEEE Trans. Mobile Comput.*, vol. 8, no. 6, pp. 735–749, June 2009.

[14] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[15] E. Nayebi and B. D. Rao, "Access point location design in cell-free massive MIMO systems," in *Proc. 2018 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 985–989.

[16] G. R. Gopal, G. P. Villardi, and B. D. Rao, "Is vector quantization good enough for access point placement?" in *Proc. 2021 55th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2021.