Load Balancing in Cellular Networks with User-in-the-loop: A Spatial Traffic Shaping Approach

Ziyang Wang, Rainer Schoenen, Halim Yanikomeroglu, and Marc St-Hilaire
Department of Systems and Computer Engineering, Carleton University, Canada
Email: {wangzi, rs, halim}@sce.carleton.ca, marc_st_hilaire@carleton.ca

***Abstract*—Long term user rate in cellular networks is the product of spectral efficiency achieved and the resources (time/frequency slots) allocated. The former is related to the received SINR, while the latter is limited by the load of the associated cell. The max-SINR cell association strategy has been used in cellular networks from GSM to LTE. This strategy maximizes the possible achieved spectral efficiency but fails to account for the load imbalance. Recently, there have been several investigations on load-aware cell association as an approach to match the traffic demand with the traffic supply, in which a user may associate to a less loaded cell, even though it does not necessarily provide the maximum SINR. In other words, a user is associated with a cell to get more share of resources at the cost of lower spectral efficiency.**

This paper goes beyond that by proposing a new load balancing approach that can simultaneously increase the user received SINR and the share of allocated resources. This is achieved by the user-in-the-loop (UIL) paradigm, which encourages the user to move to a new location that maximizes the utility function considering the received SINR, cell load and the probability of moving. Numerical results show that the UIL can increase the mean user rate substantially in comparison to the max-SINR or the load-aware cell association strategy, and also results in a more balanced load across the network.

***Index Terms*—User-in-the-loop, load balancing, cell association, load-aware, heterogeneous cellular networks.**

I. Introduction

Cellular networks are becoming increasingly heterogeneous in two different dimensions. First in the traffic supply dimension, the architecture of the network is becoming increasingly heterogeneous, with small cells (such as picocells and femtocells) layered upon traditional macrocells. Second, from the traffic demand side, new applications with diversified traffic patterns are emerging everyday with the proliferation of smart mobile devices (e.g., smart phones, tablets and smart watches). Due to the disparities between macrocell base stations (BSs) and small-cell BSs in terms of transmit power, antenna gain, and antenna height, the coverage sizes of these two types of BSs are massively different. As such, the conventional max-SINR (associate the user to the cell whose SINR is maximum) cell association strategy results in significant traffic imbalance in heterogeneous networks (HetNets); this is a major source of performance degradation [1].

Recently, load-aware cell association has been considered extensively in the literature as an approach to load balancing. In [2], the handoff and cell association are formulated into an iterative optimization problem and solved with a distributed load-awareness scheme. In [3] and [4], user-cell associations are obtained from an optimization problem with sum log-utility as the objective function. The duality theory is used to tackle the optimization problem, in which the loads are interpreted as the prices of the BSs. In [5], a heuristic solution to the joint problem of cell association and linear beamformer design in a MIMO HetNets is given.

This paper adopts a different approach to achieve load balancing. Under the paradigm of user-in-the-loop (UIL) [6], users are suggested opportunistically to move to a new location according to operator recommendation and incentive displayed on the user terminal (UT). Users can choose to or not to comply with the suggestion for better service and better rewards (depending on the content of the incentive). In this paper, the new location is the place where the user will receive higher SINR and/or more resource share, and the user is likely to move to. The suggestion is given by the UIL controller (see Figure 1) based on the utility function, which considers SINR maps of all cells (the potential spectral efficiencies of the new locations in different cells), the load factor of each cell and the probability of each user to move to different locations (related to application class, user behavior, and so on). Load balancing is achieved by the spatial movement of users which comply with the suggestions. By shaping the traffic spatial distribution, the traffic demand is controlled to meet the traffic supply better, the distribution of which usually stays unchanged once the placement of BSs has been completed.

A. User-in-the-loop

The recently developed UIL concept [6]–[9] aims at controlling the user ("layer-8") behavior in a wireless system to achieve a better performance of both the user and the network by convincing the users to move from one location to a better one or to avoid traffic congestion by postponing session traffic out of the busy hours. Based on the impact dimension, the approach is called spatial or temporal UIL control (this paper only involves spatial UIL). In both cases, the user is within, as part of, a closed-loop control system.

In [7], the authors show that substantial cell spectral efficiency gain is obtained with the use of UIL. In [8], the questions about what type of incentives will lead to what type of user reactions are answered based on survey results. Thus we are able to quantitatively describe the user behavior in a system theoretic framework. In [9], the economic aspect of the UIL concept is investigated in order to find relevant business cases for the operators and the wireless subscribers. This paper

This work is supported in part by Huawei Canada Co., Ltd., and in part by the Ontario Ministry of Economic Development and Innovations ORF-RE (Ontario Research Fund - Research Excellence) program.

goes further with a utility function that incorporates cell-level load factor, potential SINR and the user moving probability, which leads to a novel spatial traffic shaping approach to load balancing.

B. Contributions and Organization

In this paper, a novel load balancing approach with spatial traffic shaping by UIL is proposed. We show that the spectral efficiency for the users and the load balancing for the system can be improved at the same time with this approach, resulting in significant network performance enhancement.

The rest of this paper is organized as follows. First, in Section II, the system model for spatial traffic shaping is introduced. The diagram inside explains how the user is incorporated as part of a closed-loop control system. In Section III, a user model is introduced, which includes user traffic class, resource allocation scheme and user spatial distribution. The load balancing approach with spatial traffic shaping is proposed in Section IV, and a load-aware cell association approach for load balancing without users' movement is introduced in Section V. Numeric results are presented in Section VI. Finally, the paper is concluded in Section VII.

II. System Model for Spatial Traffic (Demand) Shaping

The UIL system theoretic model for traffic shaping is shown in Figure 1. Instead of assuming users being a traffic generator only, the UIL framework allows a control input into the user block, through which the user receives suggestions and incentives (i.e., progressive tariffs, reward points, higher access rates, or even environmental indicators) in order to convince him to move to a new location; this is what we refer to as spatial traffic (demand) shaping. The control information (CI) is sent from the UIL controller in form of suggestions on the UT graphical interface (e.g., a map with directions). The suggestion is opportunistic (not mandatory), and users can choose to comply or not. The action of the users (cooperate or not) is then returned to the cellular network. The suggestions are the main output of the UIL controller based on the utility function discussed in Section IV, and the incentive is set by the operator considering factors such as the current tariffs, the marketing policy, the network congestion degree, and so on.

According to the construct in this paper, the input of the UIL controller includes three components: (1) The map information from the database, (2) the user probability of moving, P, from the user behavior learning center, and (3) the cell load information, L, from the cellular network. The map information that facilitates the movement of users, and the spectral efficiency map is used to calculate the utility function. The spectral efficiency map is generated from the network with the measurement from all the UTs being accumulated and statistically averaged. It is relatively constant and only needs to be updated when the network configuration or the city landscape changes.

P is the output of the user behavior learning center, which learns the user behavior under different circumstances. For example, the probability of user u with quality of service (QoS) q to move distance d with incentive i within the user's context c can be formulated as $P_u(d, q, i, c)$. A QoS level may

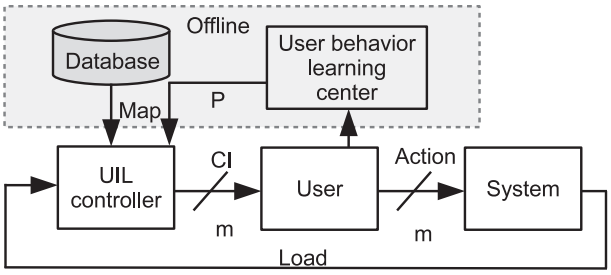

Fig. 1. The UIL system theoretic model is a closed loop with the user included in the system to control. The user's output is the action of complying or not with the suggestions. As we have multiple users in a cell, the arcs between controller, user, and the system are vectorized for m users.

be real-time, non-real-time, or differentiated based on contract (e.g., gold, silver, or bronze). The incentive may take different forms: Financial bonus, penalty surcharge, extra capacity, or even environmental indicators. The user's context c can be in various forms. The user may be known to be immobile, for example in a stadium. The user may be known to pay all penalties and discard all incentives. The user options may be set to suppress all UIL suggestions. The connectivity of the user to other peers nearby may be part of the context as well. Other human, social, or technical aspects may also be part of the context. With the evolution of machine learning, more specific output could be possible, e.g., the probability of one specific user at a specific location to move to another specific spot at a specific time.

While the map information and the value of P are being provided from offline databases, the cell load L is included in the closed-loop and is updated simultaneously when a new user session arrives to or departs from the system.

III. User Model

A. User Traffic Class

Unlike most papers in the literature, which assume one type of users only (best effort in most cases), we model users with two different traffic classes: guaranteed bit rate (GBR) and best effort (BE). An example of these two traffic types could be real-time video application and FTP/HTTP download. The GBR users have a guaranteed rate and a higher priority, while BE users share all of the resources in the cell that are not used by GBR users. In this paper, we assume that all the GBR users are guaranteed a fixed rate r when the service is available or receive no service at all (i.e., outage) if the resources are not sufficient for the user to reach the guaranteed rate r.

B. Resource Allocation

We denote by $\mathcal{C}$ the set of all cells (including macrocells and small-cells). The total resources of a cell $j \in \mathcal{C}$, denoted by W_j, is shared among all the users associated with it. The resources can be time or frequency slots, or both. In our model, GBR users have higher priority and are allocated the exact amount of resources they need. For a GBR user i with spectral efficiency s_{ij} (when associating with cell j), the amount of resources needed from cell j for user i to reach the guaranteed rate r is

$$w_{ij} = \frac{r}{s_{ij}}. \tag{1}$$

The spectral efficiency s_{ij} is derived from the Shannon formula,

$$s_{ij} = \log\left(1 + \frac{P_j g_{ij}}{\sum_{h \in C, h \neq j} P_h g_{ih} + \sigma^2}\right), \quad (2)$$

where P_j is the transmit power of cell j, and g_{ij} denotes the channel gain between user i and cell j. The channel gain includes antenna gain and path loss (including shadowing). Fast fading is not considered here as an averaged SINR is assumed over the session length. In (2), σ^2 represents the noise power.

When a new GBR user i arrives to cell j, the user will be allocated the exact amount of resources w_{ij} (see (1)) that is needed if

$$W_j - \sum_{i' \in \mathcal{U}^g(i)} a_{i'j} w_{i'j} > w_{ij}, \quad (3)$$

where $\mathcal{U}^g(i)$ is the set of the existing GBR users at the time when user i arrives to the system, and $a_{i'j}, i' \in \mathcal{U}^g(i)$, is the association indicator (i.e., $a_{i'j}$ = 1, if user i' is associated with cell j, $a_{i'j}$ = 0 otherwise). The left side of the inequality represents the remaining resources after all the existing GBR users in cell j are satisfied. If the inequality (3) does not hold, this GBR user will be blocked, i.e., a service outage occurs.

The BE users equally share the remaining resources that are not used by the GBR users. For example, when a new BE user k arrives to cell j, the resources will be reallocated to these $n_j^b(k) + 1$ users, where $n_j^b(k)$ is the number of the existing BE users associated with cell j at the time when user k arrives. The amount of resources allocated to user k is

$$w_{kj} = \frac{W_j - \sum_{i \in \mathcal{U}^g(k)} a_{ij} w_{ij}}{n_j^b(k) + 1}. \quad (4)$$

Because of the outage mechanism of the GBR users, w_{kj} is always positive, which means each BE user will be served no matter how little resources are allocated. Minimum rate demand for BE users can be added to the model, which will result in possible outage for BE users as well. Note that the arrival/departure of a GBR user i to/from cell j also triggers the reallocation of resources for all the existing $n_j^b(i)$ BE users.

C. Probability of Moving: The P Function

This paper adopts the P function from [8], in which the probability of user moving is modeled as a function of moving distance, QoS, and incentive (i.e., $P_u(d, q, i)$) without considering c, the user's context as introduced in Section II. The functions are obtained from regression analysis on survey results. Three types of services (QoS) were included in the questionnaire: Data, video, and voice. We take the results of data and video as the behavior model for the BE and GBR users, respectively. The fitting exponential functions for video traffic (GBR users) and data traffic (BE users) are $p = e^{(-0.0285-0.0265\cdot\delta)\cdot d}$ and $p = e^{(-0.0327-0.0310\cdot\delta)\cdot d}$, respectively, where δ is the discount incentive [8].

D. User Spatial Distribution

Besides the traffic class, the user spatial distribution is also non-uniform (i.e., heterogeneous). This paper adopts the method in [10], in which users are randomly and heterogeneously distributed modeled with a log-Gaussian Cox process (LGCP). Cox process is a generalization of the Poison point process (PPP), with the intensity itself being a Stochastic process [11]. LGCP is a Cox process with a log-Gaussian parent process. By changing the standard deviation σ of the Gaussian process, the distribution of users are controlled smoothly from homogeneous to extremely heterogeneous [10]. When $\sigma = 0$, the LGCP falls back to a homogeneous PPP.

To better captures the level of the user spatial heterogeneity, this paper adopts the metric introduced in [12], in which the measures based on Voronoi and Delaunay tessellations are proposed. The user spatial heterogeneity can be represented by a non-negative real number C (normalized coefficient of variation). This formulation results in a C value that is greater than 1 in super-Poisson processes, equal to 1 in PPP, and between 0 and 1 in sub-Poisson processes. The LGCP brings more heterogeneity to the PPP, so it is a super-Poisson process with $C \geq 1$.

IV. Proposed UIL Load Balancing Approach

For each user, the UIL controller outputs the control information as the suggestion of a potential location based on the maximization of the utility function. The objective utility function combines three factors: The SINR value of different locations from each cell, the load of each cell, and the probability of moving for this user from the current location to the potential locations.

A. Utility Function of GBR Users

The utility function of a GBR user i with cell j is formulated as

$$U_{ij}(x, y) = p_i(x, y) \cdot s_j(x, y) \cdot (1 - \rho_j^b(i)), \quad (5)$$

where $p_i(x, y)$ is the probability of user i moving from his current location to the new location with coordinates (x, y), $s_j(x, y)$ is the spectral efficiency map of cell j, and $\rho_j^b(i)$ is the load factor ($\in [0, 1]$) of cell j for the existing GBR users at the time when user i arrives to the system. $\rho_j^b(i)$ is defined as

$$\rho_j^b(i) = \frac{\sum_{i' \in \mathcal{U}^g(i)} a_{i'j} w_{i'j}}{W_j}, \quad (6)$$

where $\mathcal{U}^g(i)$, $a_{i'j}$, and $w_{i'j}$ have the same definitions as in Section III-B. So for each new user arriving to the system, we get a three dimensional matrix. The first dimension is the cell index, and the other two dimensions are the coordinates of the map.

B. Utility Function of BE Users

The utility function of a BE user k with cell j is formulated as

$$U_{kj}(x, y) = p_k(x, y) \cdot s_j(x, y) \cdot \frac{(1 - \rho_j^b(k))}{n_j^b(k) + 1}, \quad (7)$$

where $\rho_j^b(k)$, as defined in (6), is the load factor of cell j of the existing GBR users when BE user k arrives to the system.

C. Sequential Optimization

Due to the fact that each user (or session) arrives to the system sequentially and takes action (move or not after receiving UIL suggestion) independently, it is unrealistic to formulate the movement suggestions of all users in a one-shot optimization problem and find the global optimum suggestions for all users. In this paper, we use sequential optimization to find the optimal location for each new user based on the network situation at the time the user arrives. Different from dynamic programming, the sequential optimization never reconsiders its choices, i.e., a user will not receive a second suggestion during one data session even though the network load changes (and it may be better for him to move to another location) after he receives the first suggestion.

For a new GBR user i, we conduct an exhaustive search on the utility function $U_{ij}(x, y)$ of all the cells and locations based on the current load information. The optimization problem is formulated as

$$\begin{aligned} \max_{j,x,y} \quad & U_{ij}(x,y) \\ \text{s.t.} \quad & W_j - \sum_{i' \in \mathcal{U}^g(i)} a_{i'j} w_{i'j} > \frac{r}{s_j(x,y)} \\ & x \in [0, X] \\ & y \in [0, Y] \\ & j \in \mathcal{C}, \end{aligned} \tag{8}$$

where x and y take discrete values from the geographic range X and Y based on the resolutions of the map. $\frac{r}{s_j(x,y)}$ gives the resources needed to reach the rate r at different locations of cell j with spectral efficiency $s_j(x, y)$. The exhaustive search over the three dimensional matrix provides the optimal utility $U_i^\star$ for user i, and the corresponding values of the variables, i.e., the optimal associated cell $j^\star$ and the optimal location $(x^\star, y^\star)$. The probability of user i to move to the optimal location ($x^\star$, $y^\star$) is obtained from the P function of this user, $p_i(x^\star, y^\star)$, which we denote as $p_i^\star$.

The first constraint in (8) guarantees that there are sufficient resources in the potential cell for the GBR user i to reach the rate r. If no cell is feasible because of the constrains, it means either there is no location in any cell that can provide the spectral efficiency and the required resources to make user i reach the guaranteed bit rate, or while such a location does exist, this user has zero probability to move there based on the P function (e.g., too far from the current location). This results in an outage for this GBR user and no movement suggestion will be given.

The optimization problem for the BE users is similar to the GBR users except that it comes without the first constraint that guarantees the amount of resources available as in (8).

In the simulation, the action of the user can be modeled as a Bernoulli trial. A random number between 0 and 1 is generated and is compared to $p_i^\star$. This user moves if the random number is smaller than $p_i^\star$, and stays otherwise. If user i moves, the location of this user will be changed to ($x^\star$, $y^\star$), and cell $j^\star$ will be assigned as the associated cell for this user, i.e., $a_{ij^\star} = 1$. If the user stays according to the Bernoulli trial, he will be associated with the cell that provides the highest SINR.

D. Outlook: Include Mobility Model

If a user mobility model is included in the system, (6) should be modified to include network load variation during the time period that the current user is moving to the potential location. For example, when a new user i arrives to the system at time t, the loads of the potential cells used in the utility function are not the loads at time t, $\rho_j(t)$, $j \in C$, but the load at the time when user i arrives to the potential cell, $\rho_j(t+\tau_j)$, $j \in C$, where τ_j is the time needed for user i to move to cell j. The users considered in the load $\rho_j(t + \tau_j)$ should include the coming users and exclude the leaving users during the time interval τ_j, for each cell j.

V. Load-aware Cell Association

In this section, we develop a baseline load balancing approach without involving users' movement. Users are associated with cells not purely based on the received SINR, but also on the load of each cell.

Load-aware cell association has been researched in the literature intensively as an approach to perform load balancing, yet only one class of users is assumed (in most cases, best effort users). To better understand the effect of spatial traffic shaping by UIL, we formulate the load-aware association problem with the same user model and the same optimization method as in the UIL approach, and compare the performance of them in Section VI.

Without including the movement of users, only two factors are considered in the utility function, the spectral efficiency and the load of the potential cell. The utility function of GBR user i with cell j is defined as

$$\hat{U}_{ij} = s_{ij} \cdot (1 - \rho_j^b(i)), \tag{9}$$

where s_{ij} denotes the spectral efficiency of cell j at the current location of user i, and $\rho_j^b(i)$ has the same definition as in the UIL scheme (the load of GBR users of cell j when user i arrives to the system). Similarly, the utility function of BE user k with cell j is defined as

$$\hat{U}_{kj} = s_{kj} \cdot \frac{(1 - \rho_j^b(k))}{n_j^b(k) + 1}. \tag{10}$$

For each new GBR user, an exhaustive search is conducted on all the cells to find the maximum utility defined in (9). The optimization problem is formulated as

$$\begin{aligned} \max_{j} \quad & \hat{U}_{ij} \\ \text{s.t.} \quad & W_j - \sum_{i' \in \mathcal{U}^g(i)} a_{i'j} w_{i'j} > w_{ij} \\ & j \in \mathcal{C}. \end{aligned} \tag{11}$$

The GBR user i is associated to $j^\star$ from this optimization problem, or associated to the best-SINR cell if there is no cell feasible. For BE user k, the associated cell $j^\star$ is $\text{argmax}_j \hat{U}_{kj}$ with no constrains.

TABLE I. SIMULATION PARAMETERS

Parameter	Assumption
Macrocell layout	Hexagonal grid of 19 × 3 = 57 macrocells. Inter-site distance = 500 m
Picocell layout	57 × 2 = 114 picocells, uniformly and randomly deployed
System bandwidth	10 MHz (FDD) at 2 GHz
Average user number	10 users / cell × 171 cells = 1710 users
Percentage of GBR users	50%
Average session length	300 s
Guaranteed bit rate for GBR users	1 Mbps
Traffic model	BE and GBR
Discount incentive (δ) in UIL	-80%

VI. NUMERICAL RESULTS

A. Simulation Setup

We use 3GPP case 6.2 [13] in release 9 as the scenario for the HetNet. 19 macrocells sites (each of 3 cells) with inter site distance (ISD) of 500 meters, are configured in the system. The locations of the macrocell sites are fixed and form a hexagonal grid layout. The HetNet consists of two tiers with outdoor picocells overlaid on the same area of macrocells. The distribution of picocells is random and uniform. A wrap-around technique is applied on both macro and small-cells to eliminate the boundary effect.

The channel follows the model 2 in [13] for both macrocells and outdoor picocells, in which a LOS and NLOS path loss model is used. The downlink signal experiences path loss (including shadowing), while the fast fading is assumed to be averaged out. Both macrocells and picocells share the same bandwidth, and no interference coordination or cancellation technique is used. Table I shows the key parameters used.

Users arrive the system according to a Poisson process and depart the system after the session length. In each drop of the simulation, the system starts from zero users. All the performances are evaluated based on the snapshots of the system when the user number is in steady state, and each snapshot is taken from an individual drop, i.e., the snapshots are totally independent.

The metrics are evaluated with respect to the increasing spatial traffic demand heterogeneities under three scenarios: (1) No load balancing with best-SINR association strategy, (2) load balancing approach with load-aware association strategy as introduced in Section V, and (3) load balancing approach with the UIL scheme as proposed in Section IV. We change the value of σ in LGCP to get user distributions with different spatial heterogeneities, which are measured by C. When $\sigma = 0$, the user distribution becomes a PPP, which results in $C = 1$.

B. Loads among Different Cells

The load of GBR users is defined as ρ in (6) in Section IV-A. As the GBR users have higher priority, the standard deviation of ρ_j ($j \in C$), denoted as σ_ρ, can be used as a measure to indicate the degree of network-wide load balance. The lower the σ_ρ is, the more balanced the network is. Figure 2 shows σ_ρ with respect to user spatial heterogeneities under different load balancing scenarios. First, it is obvious that the network is becoming more imbalanced when the user spatial heterogeneity (i.e., traffic demand heterogeneity) increases in all the three scenarios. Second, σ_ρ is the lowest in the UIL scheme, highest in the best-SINR scheme, and the load-aware cell association scheme is in the middle. It shows that the UIL approach and the load-aware cell association approach are both effective in load balancing, and the former performs better.

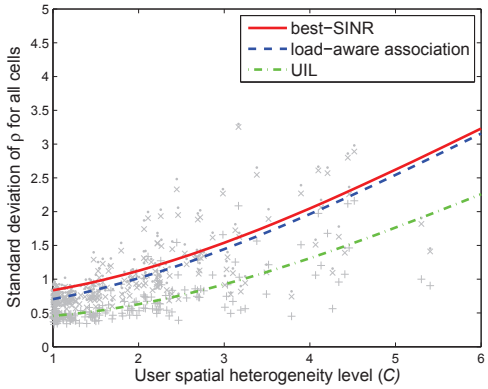

Fig. 2. The standard deviation of GBR user load (ρ) of all the cells with respect to user spatial heterogeneities under different load balancing scenarios. Each dot, cross, and plus sign represents the result of a simulation drop in best-SINR, load-aware cell association and UIL scenario, respectively. The lines are the polynomial fittings with $n = 3$. Same rule applies to other figures.

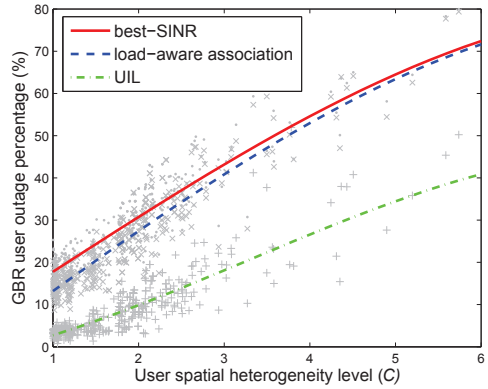

Fig. 3. The outage percentage of GBR users with respect to user spatial heterogeneities under different load balancing scenarios.

C. GBR Users Outage

As shown in Figure 3, the GBR users' outage percentage is the lowest (best) with the UIL scenario and the highest (worst) with the best-SINR scenario. The load-aware cell association strategy as an approach to load balancing without the involvement of user relocations performs in the middle. The outage percentages rise quickly with the increase of user spatial heterogeneity (i.e., traffic demand heterogeneity) in all the three scenarios.

The reason is when the user spatial heterogeneity increases, the traffic demand may be highly clustered in some areas that exceed the capacities of the associated cells, which results in higher GBR user outage. Load-aware cell association approach takes the load factor into account, and thus performs better than the best-SINR strategy without load balancing. The UIL

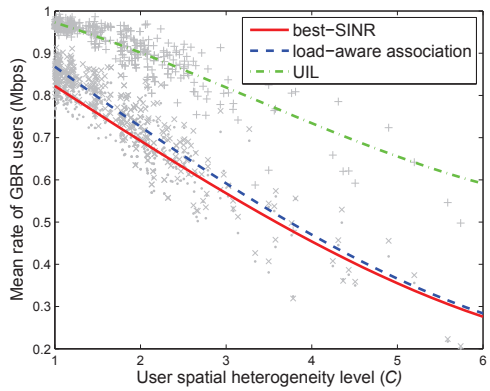

Fig. 4. Mean rate of all the GBR users in the system with respect to user spatial heterogeneities under different load balancing scenarios.

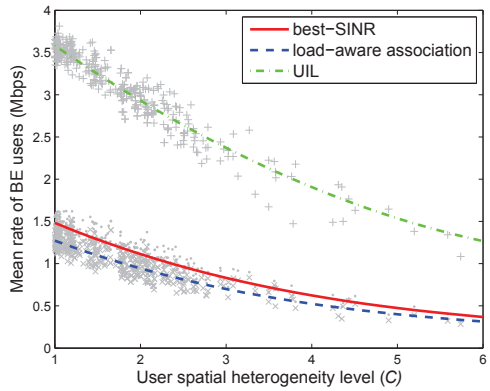

Fig. 5. Mean rate of all the BE users in the system with respect to user spatial heterogeneities under different load balancing scenarios.

scheme outperforms the load-aware cell association by combining user relocation and load factor together.

D. Mean User Rate

The mean user rates of GBR users and BE users are calculated separately as shown in Figure 4 and Figure 5, respectively. They all have the same downward trends with the increase of user spatial heterogeneity. For GBR users, the UIL scheme performs the best, the load-aware association scheme comes next, and the best-SINR strategy with no load balancing performs the worst. This can be explained with the same reason as for the outage percentage. However, for BE users, the UIL scheme still performs the best (more than twice as the best-SINR strategy), but the load-aware cell association strategy does not surpass the best-SINR strategy. This is due to the fact that the load-aware strategy may associate a GBR user to a lightly loaded neighbor cell with lower received SINR, which results in a higher resource consumption. Because of the priority of GBR users, less resources are left for the BE users, which brings a worse performance for the load-aware association strategy in terms of mean user rate of BE users.

VII. Conclusion and Outlook

In this paper, a novel load balancing approach in cellular networks is proposed. User-in-loop as the spatial traffic shaping method is the enabler of the approach.

A user model consisting of GBR and BE is considered in this paper with corresponding resource allocation policy. To better evaluate the performance, a load-aware cell association strategy is introduced with the same user traffic model and resource allocation policy as the UIL scheme. Numerical results show that the proposed load balancing approach with UIL outperforms the load balancing approach with load-aware cell association strategy and the non-load-balancing approach with best-SINR association strategy.

The proposed load balancing approach can be extended with the temporal UIL. Similar to the policies that are implemented in power supply companies, temporal UIL encourages users to postpone a heavy data application in busy hours. The UIL concept can also be used in machine-type of communications if the machine nodes can be controlled to move.

VIII. Acknowledgement

The authors would like to thank Dr. Gamini Senarath and Dr. Ngoc Dao (Huawei Technologies Canada Co., LTD.) for their valuable comments.

References

[1] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in hetnets: old myths and open problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, April 2014.

[2] A. Sang, X. Wang, M. Madihian, and R. Gitlin, "A load-aware handoff and cell-site selection scheme in multi-cell packet data systems," in *IEEE GLOBECOM*, vol. 6, Nov 2004, pp. 3931–3936 Vol.6.

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

[4] K. Shen and W. Yu, "Downlink cell association optimization for heterogeneous networks via dual coordinate descent," in *IEEE ICASSP*, May 2013, pp. 4779–4783.

[5] M. Sanjabi, M. Razaviyayn, and Z.-Q. Luo, "Optimal joint base station assignment and downlink beamforming for heterogeneous networks," in *IEEE ICASSP*, March 2012, pp. 2821–2824.

[6] R. Schoenen and H. Yanikomeroglu, "User-in-the-loop: Spatial and temporal demand shaping for sustainable wireless networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 196–203, Feb. 2014.

[7] R. Schoenen, H. Yanikomeroglu, and B. Walke, "User in the loop: Mobility aware users substantially boost spectral efficiency of cellular ofdma systems," *IEEE Communications Letters*, vol. 15, no. 5, pp. 488–490, May 2011.

[8] R. Schoenen, G. Bulu, A. Mirtaheri, T. Beitelmal, and H. Yanikomeroglu, "Quantified user behavior in user-in-the-loop spatially and demand controlled cellular systems," in *18th European Wireless Conference (EW)*, April 2012.

[9] R. Schoenen and H. Yanikomeroglu, "Economics of user-in-the-loop demand control with differentiated qos in cellular networks," in *IEEE PIMRC*, Sept 2012, pp. 1131–1136.

[10] Z. Wang, R. Schoenen, H. Yanikomeroglu, and M. St-Hilaire, "The impact of user spatial heterogeneity in heterogeneous cellular networks," in *IEEE GlOBECOM Workshop on Heterogeneous and Small Cell Networks*, Austin, USA, Dec. 2014.

[11] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 2nd ed. John Wiley & Sons, 1995.

[12] M. Mirahsan, Z. Wang, R. Schoenen, H. Yanikomeroglu, and M. St-Hilaire, "Unified and non-parameterized statistical modeling of temporal and spatial traffic heterogeneity in wireless cellular networks," in *IEEE ICC Workshop on 5G Technologies*, Sydney, Australia, June 2014.

[13] 3GPP, "E-UTRA: Further advancements for E-UTRA physical layer aspects (release 9)," TR 36.814, Mar. 2010.