Load-aware Handover Decision Algorithm in Next-generation HetNets

Konstantinos Alexandris, Nikolaos Sapountzis, Navid Nikaein, Thrasyvoulos Spyropoulos
Mobile Communications Department
EURECOM, Biot, France 06410
Email: firstname.lastname@eurecom.fr

***Abstract*—In this work we propose a novel handover (HO) algorithm, that considers system performance from both user and network perspective, in the context of heterogeneous networks (HetNets), i.e., networks composed of BSs with asymmetrical transmission power. In such an environment, conventional HO algorithms that consider only the user perspective, e.g., received signal strength (RSS)-based, might offer suboptimal performance, since they mainly push users to cells with high transmission powers. Thus, new algorithms that take into account also the network perspective, e.g., cell load, are needed. In this work, a load-aware algorithm is proposed considering the service delay that a user experiences from the network. In addition, an implementable framework based on Software Defined Networking (SDN) architecture is sketched to support the algorithm. The proposed algorithm is compared with the traditional one we meet in long-term evolution (LTE) systems and a distance-based one. Extracted cell assignment probability and user service delay performance results show that the load-aware approach outperforms both of them.**

I. Introduction

Handover (HO) decision algorithms in heterogeneous networks (HetNets) are emerging with the extreme densification and offloading of cellular systems as the key technologies to get 1000x data rate [1]. Many works have been done in HO decision algorithms considering the overall system performance from the *user* and *network perspective*, in the context of long-term evolution (LTE)/LTE–Advanced (LTE–A) networks [2]. Specifically, from the network perspective, a HO algorithm that adapts the hysteresis and the time-to-trigger (TTT) based on certain network key-metrics (e.g., HO failure or ping-pong ratios) is studied in [3], along with the self-organising networks (SONs). Advanced self-organizing map (SOM) is examined to suppress handovers in regions that coincide with many unnecessary HOs [4]; also the HO offset tuning is performed via network load-balancing for assuring handover to possible target cells and no return to their serving cell [5]. The main focus so far has been to enhance the system performance from the user-perspective. The most common used are based on: (i) received signal strength (RSS), (ii) user speed and (iii) interference-management [6], [7].

However, denser deployments in HetNets experience high spatio-temporal load variations and thus require more advanced HO algorithms that consider both perspectives, *jointly*. In addition, power based algorithms (e.g., RSS-based) proposed in the literature cannot be applied, due to asymmetrical transmission powers among macrocells and picocells (also called small cells (SCs) in the literature). Especially, in such environments, a user equipment (UE) usually remains connected to a macrocell that offers high transmission power (and is potentially overloaded due to its high number of concurrent users), while is placed close to one or more underloaded SCs. According to the latter, authors in [8] propose to scale down macrocell RSS with an appropriate factor to force connection to a SC. This factor is selected optimally based on the maximization of the SC assignment probability. Therefore, it was left as future work to examine policies that consider user throughput, delay or other user service requirements. To this end, we revisit the problem of HO, in the context of future HetNets deployments. Our contributions can be summarized as follows:

1) We construct a framework considering both user (i.e., RSS) and network (i.e., cell load) perspectives. Specifically, the *service delay* of a random flow is analytically derived, as a key QoS criterion for the HO decision algorithm under the assumption of uneven transmission power regimes. This approach overcomes the shortcomings created by only considering RSS criteria in HO decision for HetNets (Section II and III).

2) We propose a HO decision algorithm (Section III) and sketch an implementable framework based on Software Defined Networking (SDN) architecture, as suggested to be a key enabler for the realization of 5G networks [9] (Section IV).

3) We investigate the related trade-offs involved, in different load variation scenarios. Further, the proposed algorithm is compared with the (i) one used in traditional LTE systems and (ii) a distance-based algorithm proposed in [8] and significantly outperforms them (Section V).

II. System model

A. Signal model

This work considers a network consisting of a macrocell and a group of picocells located at given distances D_j from it, where $j \in \{1, \ldots, \mathcal{P}\}$ and $\mathcal{P}$ is the total number of the picocells. The cells, i.e., macrocell/picocells, are denoted by m, p_j, respectively. In addition, the picocells are assumed circular on average, regardless of the fading effects presence. Each base station (BS) is considered to be located in the center of each circular cell $i \in \{m, p_j\}$ of a given radius R_i corresponding to the (x_i, y_i) coordinates. Let $r_m^{\text{dBm}}[k]$ and $r_{p_j}^{\text{dBm}}[k]$ denote the reference signal received power (RSRP) from the macrocell BS and the picocell BS at time k [1] in dBm scale, respectively:

$$r_i^{\text{dBm}}[k] \triangleq P_{T_x,i}^{\text{dBm}} + P_{L,i}^{\text{dB}}(d_i^k) + \psi_i^{\text{dB}}[k],\ i \in \{m, p_j\}, \quad (1)$$

[1] k corresponds to the discretization of the continuous time t sampling at kT_s intervals, where T_s stands for the measurement sampling period.

where $P^{\text{dBm}}_{T_x,i}$ denotes the BS transmission power, $P^{\text{dB}}_{L,i}(d^k_i)$ denotes the path loss in dB, d^k_i represents the distance from the BS to the user that is greater than a reference distance d_r and $\psi^{\text{dB}}_i[k]$ stands for the shadowing fading in dB for each cell i. The pathloss $P^{\text{dB}}_{L,i}(d^k_i) \triangleq 10\log_{10}(K) - 10\eta\log_{10}\left(d^k_i/d_r\right)$, where K is the path loss measured at reference distance d_r and η is the pathloss exponent. The K parameter is given by $K \triangleq \left[c_0/\left(4\pi f_c d_r\right)\right]^2$, where c_0 is the universal speed of light in vacuum and f_c is the carrier frequency. The shadow fading $\psi^{\text{dB}}_i[k] \sim \mathcal{N}\left(0, \sigma^2_{\text{dB},i}/\xi^2\right)$ [2] and is assumed to be independent across i and k, where $\xi \triangleq 10/\ln(10)$. Averaging is performed by an exponential moving average (EMA) filter, i.e., low-pass filter, for smoothing any RSRP abrupt variations. High frequency fluctuations (i.e., Rayleigh fading) are filtered out and can be neglected. Consequently, the output filtered signal is:

$$\bar{r}^{\text{dBm}}_i[k] \triangleq (1-\alpha)\bar{r}^{\text{dBm}}_i[k-1] + \alpha r^{\text{dBm}}_i[k], \qquad (2)$$

where $\alpha \triangleq 2^{-q/2}$ and $q \in \mathbb{N}$.

B. Mobility and Network users

The users are distinguished in two different categories depending on their mobility status: a) **Static users** (SU): Users that are not moving on average and they don't intend to handover. The static users are divided in two subcategories based on their on-going traffic: i) **Active users** (AU): Users that have already associated with a BS and generate new flows. ii) **Disconnected users** (DU): Users that are considered to be switched-off. This implies no association with a specific BS. b) **Mobile users** (MU): Users that are moving and intend to handover. The users mobility can be described with any mobility model.

The SUs are getting active with probability p_{r_i} in each cell i at each time k. On that account, the number of the AUs in each cell i at time k, $N^k_{\text{AU},i}$, is distributed according to the binomial distribution with parameters $N_{\text{SU},i}$ and p_{r_i}, where $N_{\text{SU},i}$ stands for the total number of the static users in each cell i. The exact value of $N^k_{\text{AU},i}$ is considered to be known to the BS at each time k. The MUs are always active and N_{MU} stands for their total number.

C. Traffic model

New flows (all considered as best-effort) by active users are assumed to arrive according to a Poisson process with the total arrival rate in each cell i at time k [10]:

$$\lambda^k_i = \lambda\left(\tilde{N}^k_{\text{MU},i} + N^k_{\text{AU},i}\right), \qquad (3)$$

where λ denotes the flow arrival rate per user. The flow size follows a general distribution with mean Y. We assume the Processor Sharing (PS) scheduling discipline and adopt the stationary M/G/1/PS system [10]. $\tilde{N}^k_{\text{MU},i} \triangleq \mathbf{1}\left(u \notin i\right) + N^k_{\text{MU},i}$, where $\mathbf{1}(\cdot)$ stands for the indicator function and becomes one when a mobile user u that intends to handover is not associated with the cell i (i.e., $u \notin i$) at time k; $N^k_{\text{MU},i}$ denotes the total number of mobile user that are associated with the cell i at time k (i.e., a subset of N_{MU}). $N^k_{\text{MU},i}$ is also acquainted by each cell i BS. If the mobile user u is already associated with the cell i, then $\mathbf{1}(\cdot)$ is zero and u is already included in $N^k_{\text{MU},i}$. Thus, after computing $\tilde{N}^k_{\text{MU},i}$ and plug-inning it to Eq. (3), we get an estimation of the total arrival rate after the HO procedure. This is also used for the prediction of the service delay, as it will be discussed later in Section III.

Algorithm 1 Handover algorithm

```
Input: $\bar{\mathcal{D}}^k_m$: predicted average delay of the macrocell $m$
       $\bar{\mathcal{D}}^k_{p_j}$: predicted average delay of the picocell $p_j$
       $r_{p,th}$: picocell RSRP threshold
Output: user cell association
  $j^* = \arg\max_j \bar{r}_{p_j}$
  if $\left(\bar{r}_{p_{j^*}}[k] > r_{p,th}\right)$ then
    if $\left(\tilde{r}_{j^*}[k] > \bar{r}_m[k]\right)$ then
      connect to picocell
    else
      connect to macrocell
    end if
  end if
```

III. Handover decision algorithm

A well-known criterion, commonly used in conventional HO decision algorithms for mobile communication systems (e.g., 3GPP LTE), is based on RSRPs comparison method in which hysteresis and threshold are included [11]. Specifically, the criterion for the handover scenario is expressed as:

$$\bar{r}^{\text{dBm}}_{p_j}[k] > \bar{r}^{\text{dBm}}_m[k] + \Delta \ \wedge \ \bar{r}^{\text{dBm}}_m[k] < r^{\text{dBm}}_{m,th}, \qquad (4)$$

where Δ denotes the hysteresis and $r_{m,th}$ is the minimum RSRP macrocell threshold value.

There are scenarios where a macrocell is overloaded and it becomes crucial for the user to connect to a picocell that is probably underloaded. Due to significant differences in transmission power among a macrocell and picocells, there is a huge gap between their received powers at the UEs (i.e., $\bar{r}^{\text{dBm}}_m[k] \gg \bar{r}^{\text{dBm}}_{p_j}[k]$). Hence, the above criterion does not often hold and the conventional HO algorithm cannot be triggered to maintain the required QoS (e.g., in terms of rate, delay and throughput). Thus, *load-aware* algorithms development are crucial for uneven received powers regimes found in HetNet scenarios.

The proposed load-aware (LA) policy is described in Algorithm 1. The algorithm is applied only to the mobile users by taking into account also the static active ones. The disconnected users are not taken into account, since they are assumed to be switched-off. $\bar{r}_i[k]$ stands for the RSRP from each cell i at time k in linear scale. The picocell that the user intends to handover is the one with the maximum received power, denoted as p_{j^*}. After the selection of the p_{j^*} picocell, the condition $\bar{r}_{p_{j^*}}[k] > r_{p,th}$, with $r_{p,th}$ standing for the picocell RSRP threshold, ensures that the user stays within the picocell coverage area and the received signal is retained adequately strong. Thus, the algorithm is running only inside the picocell. Subsequently, we define $\tilde{r}_{j^*}[k] \triangleq \bar{r}_{p_{j^*}}[k] + \left[1 - f\left(\bar{\mathcal{D}}^k_m, \bar{\mathcal{D}}^k_{p_{j^*}}\right)\right]\bar{r}_m[k]$. $\bar{\mathcal{D}}^k_i$ denotes

[2] The $x \sim \mathcal{N}\left(\mu, \sigma^2\right)$ denotes that random variable x is Gaussian with mean μ and variance σ^2.

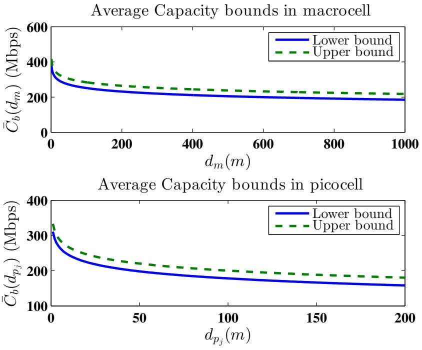

Fig. 1. Upper and lower bounds of the average capacity for the macrocell and the picocell, as a function of the distance.

the *prediction* of the average (service) delay of a new flow of a cell i at time k that the user experiences when staying in the picocell. An analytical expression for $\bar{\mathcal{D}}_i^k$ is given later in Eq. (7). It is noted that the prediction of the average delay $\bar{\mathcal{D}}_i^k$ is performed at time k, but it represents an estimation for the entire time period that the user is staying within the picocell (i.e., user's delay assumed to be $\bar{\mathcal{D}}_i^k$ during this time period). As it is mentioned later, the computations for $\bar{\mathcal{D}}_i^k$ remain consistent with this assumption. $f(\cdot)$ stands for a continuous function and is constructed properly to force the user to connect in the picocell p_{j*}, iff $\bar{\mathcal{D}}_{p_{j*}}^k \ll \bar{\mathcal{D}}_m^k$, despite the fact that $\bar{r}_m[k] \gg \bar{r}_{p_{j*}}[k]$. This approach meets the user needs that cannot be supported by the congested macrocell and its association with a picocell is critical in terms of QoS. The condition that must hold to connect to the picocell is given analytically by the following inequality that applies the above policy:

$$\tilde{r}_{j^*}[k] > \bar{r}_m[k] \Leftrightarrow \bar{r}_{p_{j^*}}[k] > f\left(\bar{\mathcal{D}}_m^k, \bar{\mathcal{D}}_{p_{j^*}}^k\right)\bar{r}_m[k], \quad (5)$$

otherwise the user is connected to the macrocell. Finally, the function $f\left(\bar{\mathcal{D}}_m^k, \bar{\mathcal{D}}_{p_j}^k\right) \triangleq \exp\left[-c\bar{\mathcal{D}}_m^k/\bar{\mathcal{D}}_{p_j}^k\right]$, $c \geq 1$ with a codomain in $[0, 1]$ provides such a property as intended. Specifically, the exponential nature of the function provides quick convergence to zero when $\bar{\mathcal{D}}_{p_{j*}}^k \ll \bar{\mathcal{D}}_m^k$. This manages to scale down the received powers asymmetry in order to hold Eq. (5) and associate the user with the underloaded picocell. Finally, the constant c stands for a tuning parameter. Next, we describe the computations needed for the prediction of $\bar{\mathcal{D}}_i^k$:

The $\tilde{C}(d_{l,i}) \triangleq \left[\bar{C}_L(d_{l,i}) + \bar{C}_U(d_{l,i})\right]/2$ is defined as the arithmetic mean of $\bar{C}_L(d_{l,i}), \bar{C}_U(d_{l,i})$ that represents the analytical lower and upper bounds of the averaged capacity as a function of the l^{th} static active user distance, $d_{l,i}$, for each cell i (see Fig. 1). A detailed description of their analytical expression is given in Appendix A (see Eq. (9), (10)). The latter computations are consistent with the prediction of $\bar{\mathcal{D}}_i^k$ for the entire user's sojourn within the picocell, since $d_{l,i}$ remains invariant for the static users across time k. For the mobile users, there is no prior knowledge of their future positions, thus their corresponding lower and upper bounds of the averaged capacity for each cell i, $\bar{C}_L^i, \bar{C}_U^i$, are averaged over all the possible distances, assuming that the user passes uniformly through all the points within the picocell after multiple visits during its staying within it. The uniform distribution of the points as well as the corresponding capacity analysis are enclosed in Appendix B. Thus, the aforementioned computations are consistent also with the prediction of $\bar{\mathcal{D}}_i^k$ for the total user's sojourn within the picocell. Finally, their respective arithmetic mean is given by $\tilde{C}_i = \left(\bar{C}_L^i + \bar{C}_U^i\right)/2$.

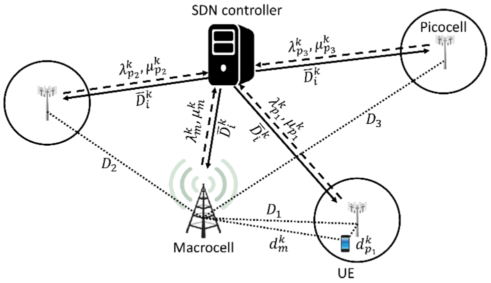

Fig. 2. Network setup and SDN controller.

According to the above computations, the expected average rate [3] at time k for each cell i is determined by [12]:

$$\mathcal{R}_i^k = \frac{\tilde{N}_{\text{MU},i}^k \tilde{C}_i + \sum\limits_{l=1}^{N_{\text{AU},i}^k} \tilde{C}\left(d_{l,i}\right)}{\tilde{N}_{\text{MU},i}^k + N_{\text{AU},i}^k}. \quad (6)$$

Consequently, based on the system assumptions (i.e., stationary M/G/1/PS), the prediction of $\bar{\mathcal{D}}_i^k$ is given by [10]:

$$\bar{\mathcal{D}}_i^k = \frac{1}{\mu_i^k - \lambda_i^k}, \quad (7)$$

where $\mu_i^k = \mathcal{R}_i^k / Y$ stands for the service rate at time k for each cell i.

IV. SDN-based implementation

The proposed handover algorithm decides whether a mobile user should handover from its (currently) associated BS, to another nearby BS that might promise enhanced QoS. To make such a decision, the user should be aware of the cell load or the offered user QoS for *both* BSs. To this end, a centralized entity determines the service delay of each individual cell based on the received measurements and communicates this with the underlying network. SDN architecture facilitates this procedure, since it offers a centralized programmable control for the underlying network [9]. Following the SDN outline, we consider three planes as illustrated in Fig. 2:

Controller tier: Each time k, the controller: a) receives the respective λ_i^k and μ_i^k from all cells, b) computes and sends the corresponding delays $\bar{\mathcal{D}}_i^k$ to all BSs needed (e.g., a certain BS should know not only his corresponding service delay, but also the one from its neighboring BSs).

Network tier: Each time k, BSs: a) send the respective λ_i^k and μ_i^k, b) receive the corresponding delays $\bar{\mathcal{D}}_i^k$, c) send the $\bar{\mathcal{D}}_i^k$ to the UE.

User tier: Each time k, the UE: a) receives the delays $\bar{\mathcal{D}}_i^k$, b) triggers the association procedure based on Algorithm 1.

V. Simulation results

This section investigates the proposed LA HO policy performance compared to other HO policies. The entire framework for the simulation is built in MATLAB. The simulation parameters are provided in Table I [13].

[3] The rate considers a priori the user association, as it is used for the predicted delay. Thus, $\tilde{N}_{\text{MU,i}}^k$ is plug-inned.

TABLE I. SIMULATION SETUP PARAMETERS

parameter	value	parameter	value	parameter	value
R_m	1000 m	f_c	700 MHz	p_{r_m}	0.8
R_{p_j}	200 m	B	10 MHz	$p_{r_{p_j}}$	0.8
$P^{\text{dBm}}_{T_x,m}$	43 dBm	d_r	1 m	Δ	3 dB
$P^{\text{dBm}}_{T_x,p_j}$	21 dBm	q	4	Y	10 MBytes
N_0^{dBm}	-104.5 dBm	η	2	s	50
$r^{\text{dBm}}_{m,\text{th}}$	-90 dBm	$\sigma_{\text{dB},m}$	8 dB	δ	10
$r^{\text{dBm}}_{p,\text{th}}$	-55 dBm	σ_{dB,p_j}	6 dB	c	1

The MUs mobility model is based on a two-dimensional random walk (2D RW) limited in a finite space (i.e., circular macrocell) with step s or s/δ depending on where the user is moving inside (i.e., macrocell/picocell) and $\delta > 0$ denotes a degrading step factor [4]. This variable step simulates the scenario that the users are staying inside the picocell area for a sufficient amount of time compared to the macrocell one (e.g., in airports, malls, etc). As next, Fig. 3–4 present a scenario with three picocells placed in different distances from the macrocell ($D_1 = 250$ m, $D_2 = 350$ m, $D_3 = 450$ m) and one mobile user (i.e., $N_{\text{MU}} = 1$) [5], as shown in Fig. 2.

Fig. 3 demonstrates the mobile user cell assignment probabilities, $P_r(u \in m)$, $P_r(u \in p)$ for the macrocell and the picocell, respectively. These mobile user cell assignment probabilities are presented as a function of the flow arrival rate λ for different number of static users [6] $N_{\text{SU},m} = 200, 400, 600$ and $N_{\text{SU},p_j} = 10$. Their computation is performed via Monte-Carlo simulations within the picocells p_j. Especially, $P_r(u \in p)$ is also averaged out for all picocells p_j. Specifically, with the increment of λ, the total load in the macrocell augments, since the number of the macrocell static users is greater than the picocell ones, causing macrocell overload. Consequently, the difference in the corresponding delays increases (i.e., macrocell delay is greater than that of the picocells). Thus, the LA algorithm associates the user with the underloaded cell (i.e., picocell) or equivalently $P_r(u \in p)$ increases with the augmentation of λ. Also, as $N_{\text{SU},m}$ increases the macrocell becomes overloaded with smaller values of λ, so $P_r(u \in p)$ tends to 1 in different $\lambda = 0.04, 0.06, 0.12$ for $N_{\text{SU},m} = 200, 400, 600$, respectively. What is more, we include the conventional (CONV) algorithm, described in Eq. (4) and compare it with the proposed LA one. In sharp contrast with the LA algorithm, the CONV algorithm always suggests the user to be associated with the macrocell. In case of congestion, this is suboptimal, since the user is placed close to one or more underloaded cells. $P_r(u \in m)$, is also plotted, showing the complementary behavior of the algorithm.

Fig. 4 shows the predicted delays, $\bar{\mathcal{D}}_m, \bar{\mathcal{D}}_p$ for the macrocell and the picocell, as a function of the flow arrival rate λ. Specifically, these are the $\bar{\mathcal{D}}^k_m, \bar{\mathcal{D}}^k_p$ that are also averaged out across time k. Their computation is performed via Monte-Carlo simulations for different number of static users $N_{\text{SU},m} = 200, 400, 600$ and $N_{\text{SU},p_j} = 10$ within the picocells p_j. Especially, $\bar{\mathcal{D}}_p$ is also averaged out for all picocells p_j. It is noticed that the average delay, $\bar{\mathcal{D}}_m$, increases *sharper* for larger values of $N_{\text{SU},m}$ and smaller values of λ, according to Eq. (7). $N_{\text{SU},m}$ and λ affect $\bar{\mathcal{D}}_m$ increment and consequently enlarge the difference between $\bar{\mathcal{D}}_m$, $\bar{\mathcal{D}}_p$. Hence, when their ratio becomes large ($\bar{\mathcal{D}}_m \gg \bar{\mathcal{D}}_p$), $f(\cdot)$ tends to zero and vanishes the gap between the received powers. Thus, the impact of $N_{\text{SU},m}$, λ on $\bar{\mathcal{D}}_m$, $\bar{\mathcal{D}}_p$ ratio exceeds the impact of received powers asymmetry. This validates the LA algorithm functionality that associates the user with the underloaded picocells in the case of an overloaded macrocell ($P_r(u \in p)$ goes to one), regardless the gap in received powers. On the other hand, HO is not triggered if there is not much difference in $\bar{\mathcal{D}}_m$, $\bar{\mathcal{D}}_p$, i.e., the user QoS is satisfied ($P_r(u \in p)$ goes to zero). Consequently, RSS-based HO that maximizes the instantaneous rate of a user (i.e., the best modulation and coding scheme (MCS) is used), reflects user QoS only when the BS is lightly loaded. However, user performance, in terms of per flow delay, may be severely affected if the BS offering the best RSS is congested.

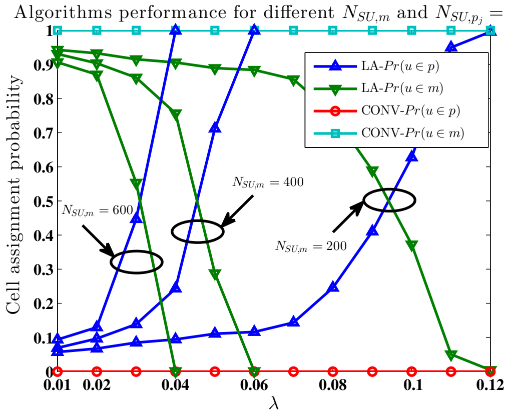

Fig. 3. Cell assignment probability (CONV vs LA) for different number of static macrocell users $N_{\text{SU},m}$ and number of static picocell users $N_{\text{SU},p_j} = 10$.

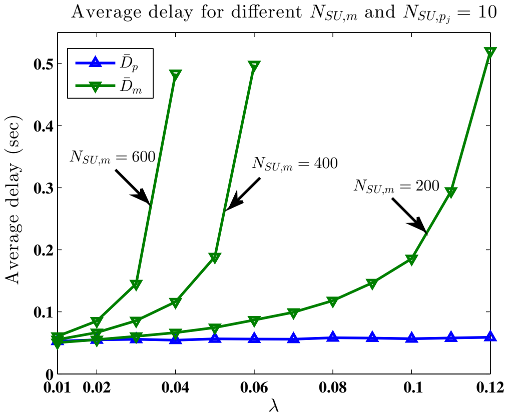

Fig. 4. Average delay in the macrocell/picocell for different number of static macrocell users $N_{\text{SU},m}$ and number of static picocell users $N_{\text{SU},p_j} = 10$.

In Fig. 5, the picocell assignment probability, $P_r(u \in p)$, is depicted as a function of the user's distance from the macrocell for fixed number of $N_{\text{SU},m} = 200$ and $N_{\text{SU},p_j} = 10$ in the picocell. Our algorithm is compared with the CONV and a distance-based (DIST) one provided in [8], for different λ. The latter is applied in hierarchical macrocell/SCs network that is also our case. The DIST approach associates the user

[4] The user coordinates $(x^k_i, y^k_i) \triangleq (x^{k-1}_i + s_i\cos(\phi), y^{k-1}_i + s_i\sin(\phi))$ in each cell i at each time k, where $s_i = \begin{cases} s & , \; i = m, \\ s/\delta & , \; i = p_j, \end{cases}$ and ϕ is uniformly distributed in $[0, 2\pi)$.

[5] The proposed algorithm as well as the performed analysis hold for the general type of cells (i.e., either overlapping or non-overlapping cells) with any radius R_i. Without loss of generality, we relax this assumption considering non-overlapping cells and the same radius for the picocells. The consideration of one mobile user is sufficient enough for studying the behavior of the algorithm without affecting its validity for the general case.

[6] At each time k, $N^k_{\text{AU},i}$ is changing. On average, we have $\mathbb{E}[N^k_{\text{AU},i}] = p_{r_i} N_{\text{SU},i}$.

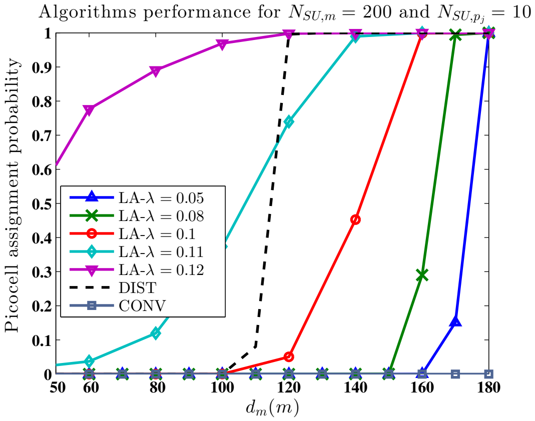

Fig. 5. Picocell assignment probability (CONV vs DIST vs LA) for number of static macrocell users $N_{\text{SU},m} = 200$ and number of static picocell users $N_{\text{SU},p_j} = 10$ for different λ.

TABLE II. DELAY GAIN FOR $N_{\text{SU},m} = 200$, $N_{\text{SU},p_j} = 10$ AND DIFFERENT λ

$\boldsymbol{\lambda}$	**0.05**	**0.08**	**0.1**	**0.11**	**0.12**
$\bar{\mathcal{D}}_{\text{LA}}$ (sec)	0.0701	0.0992	0.1261	0.1125	0.0621
$\bar{\mathcal{D}}_{\text{DIST}}$ (sec)	0.0621	0.0807	0.1109	0.1470	0.2808
$\bar{\mathcal{D}}_{\text{CONV}}$ (sec)	0.0743	0.1160	0.1864	0.2708	0.5496
$\bar{\mathcal{D}}_{\text{DIST}}/\bar{\mathcal{D}}_{\text{LA}}$	0.8861	0.8141	0.8798	1.3066	4.5235
$\bar{\mathcal{D}}_{\text{CONV}}/\bar{\mathcal{D}}_{\text{LA}}$	1.0596	1.1701	1.4790	2.4073	8.8547

with its closest SC regardless the asymmetry in BSs powers. This is achieved using a scale down factor α based on the maximization of the SC assignment probability. Further, some of the system parameters are changed to adapt the simulated scenario in [8] [7]. For the DIST algorithm, $P_r(u \in p)$ is close to 1 after a specific user's distance from the picocell despite the fact that the macrocell is overloaded or not. However, with the LA approach, $P_r(u \in p)$ increases, depending on the macrocell load for different λ, no matter the user's proximity to the picocell. Thus, if the user is close to a picocell, its association with it may be unnecessary, since its QoS requirements may be already met. Finally, the CONV algorithm does not trigger the HO, due to BSs power assymetry.

The overall delay that the user experiences (i.e., $\bar{\mathcal{D}}_m$ from the macrocell and $\bar{\mathcal{D}}_p$ from the picocell depending on where it is associated) is presented in Table II. The setup is considered the same as in Fig. 5. Using the CONV algorithm, it is noted that the overall delay, $\bar{\mathcal{D}}_{\text{CONV}}$, is identical to the delay of the macrocell (see Fig. 4 for $N_{\text{SU},m} = 200$ and $N_{\text{SU},p_j} = 10$), since the user is always connected to the macrocell. The DIST algorithm retains the user connected to the macrocell up to a specific distance threshold and after that it attaches it to the picocell (see Fig. 5) with $P_r(u \in p)$ close to 1. Thus, the augmentation of the macrocell delay, when it becomes overloaded, compared to the picocell ones, increases the overall delay, $\bar{\mathcal{D}}_{\text{Dist}}$. The LA algorithm keeps a conservative policy when the load is low in the macrocell, i.e., the user remains associated with the macrocell. Hence, $P_r(u \in p)$ is low even if the user is close to the picocell. This policy gives slightly higher overall delays when $\lambda \leq 0.1$ compared to the DIST one, since the latter connects the user to the cell with the lower delay (i.e., picocell) earlier. Consequently, unnecessary handovers to the picocell and related signaling overhead are avoided with the LA approach. On the other hand, if the load is high in the macrocell, the LA algorithm associates the user with the underloaded picocell earlier than the DIST algorithm that is load-unaware giving significant performance to the overall delay, $\bar{\mathcal{D}}_{\text{LA}}$. The respective ratios of the CONV/DIST algorithms over the LA algorithm (i.e., $\bar{\mathcal{D}}_{\text{CONV}}/\bar{\mathcal{D}}_{\text{LA}}$, $\bar{\mathcal{D}}_{\text{DIST}}/\bar{\mathcal{D}}_{\text{LA}}$) overall delays are demonstrated to show the corresponding gains in Table II. The same trend holds for higher values of $N_{\text{SU},m}$, but with different values of λ, as explained in Fig. 4. Finally, significant gains are provided in high load scenarios (i.e., $\simeq 4$ and $\simeq 8$ compared to the CONV and DIST approach).

VI. Conclusion & Future work

This work focuses on user-centric HO decision algorithms for HetNets. It was shown that the proposed load-aware algorithm can significantly enhance the system performance by considering both user and network perspective in order to improve user QoS. To this end, it was compared with a conventional and a distance-based LTE-handover algorithm and we showed that it can outperform both of them according to the extracted cell assignment probability and user service delay performance results. As future work, we plan to investigate more complex scenarios with more mobile users and study the impact of our load-aware algorithm on user-distribution and further on *load-balancing*. Finally, although we treated all flows as best-effort, we plan to consider also *dedicated* flows based on different QoS metrics (e.g., blocking probabilities).

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme under grant agreement no 612050 (FLEX Project).

References

[1] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5g be?" *IEEE Journal on Selected Areas in Communications*, 2014.

[2] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility management for femtocells in LTE-advanced: Key aspects and survey of handover decision algorithms," *IEEE Communications Surveys Tutorials*, 2014.

[3] T. Jansen, I. Balan, J. Turk, I. Moerman, and T. Kürner, "Handover parameter optimization in LTE self-organizing networks," in *IEEE 72nd Vehicular Technology Conference Fall (VTC 2010-Fall)*, 2010.

[4] N. Sinclair, D. Harle, I. Glover, J. Irvine, and R. Atkinson, "An advanced som algorithm applied to handover management within LTE," *IEEE Transactions on Vehicular Technology*, 2013.

[5] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan, "Load balancing in downlink LTE self-optimizing networks," in *IEEE 71st Vehicular Technology Conference (VTC 2010-Spring)*, 2010.

[6] H. Zhang, W. Ma, W. Li, W. Zheng, X. Wen, and C. Jiang, "Signalling cost evaluation of handover management schemes in LTE-advanced femtocell," in *IEEE 73rd Vehicular Technology Conference (VTC Spring)*, 2011.

[7] D. Lopez-Perez, A. Ladanyi, A. Jüttner, and J. Zhang, "Ofdma femtocells: Intracell handover for interference and handover mitigation in two-tier networks," in *2010 IEEE Wireless Communications and Networking Conference (WCNC)*, 2010.

[8] J.-M. Moon and D.-H. Cho, "Novel handoff decision algorithm in hierarchical macro/femto-cell networks," in *2010 IEEE Wireless Communications and Networking Conference (WCNC)*, 2010.

[7] Specifically, one macrocell and one picocell are considered in distance $D_1 = 200$ m, $a^* = 0.96$ as the optimal a for the specific D_1, $\Delta = 0$, $R_p = 150$ m and $r_{p,\text{th}} = -51$ dBm. In addition, the mobility model is changed, since the authors in [8] assume that one mobile user moving across a line from the macrocell to the picocell coverage area.

[9] P. Rost, C. Bernardos, A. Domenico, M. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, "Cloud technologies for flexible 5g radio access networks," *IEEE Communications Magazine*, 2014.

[10] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems*. Cambridge University Press, 2010.

[11] G. Pollini, "Trends in handover design," *IEEE Communications Magazine*, 1996.

[12] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, and U. Salim, "Reducing the energy consumption of small cell networks subject to QoE constraints," in *Proceedings of the Global Communications Conference, GLOBECOM 2014, Texas, USA*, 2014.

[13] 3GPP. TR 36.931 version 9.0.0 Release 9.

Appendix A
Capacity analysis-I

The average capacity, assuming shadowing and Rayleigh flat fading [8], is expressed as:

$$\bar{C}\left(d_i\right)=\mathbb{E}_{\psi_i,|h|^2}\left[B\log_2\left(1+\rho_i K\left(d_i/d_r\right)^{-\eta}\psi_i\,|h|^2\right)\right], \quad (8)$$

where $\rho_i \triangleq P_{T_x,i}/N_0$ denotes the transmit signal-to-noise ratio (SNR) [9], $P_{T_x,i}$ is the transmission power in linear scale, d_i represents the distance from the BS cell i to the user that is greater than a reference distance d_r, B stands for the available bandwidth, N_0 represents the aggregated noise power over B in linear scale [10], $\psi_i \sim \text{Log-}\mathcal{N}\left(0,\sigma^2_{\text{dB},i}/\xi^2\right)$ [11] represents the shadowing fading in linear scale and $h\sim\mathcal{CN}\left(0,1\right)$ [12] defines the Rayleigh flat fading channel coefficient. To the best of our knowledge, an analytical expression for the averaged capacity does not exist. Therefore, its lower and upper bound, $\bar{C}_b\left(d_i\right)$, $b\in\{L,U\}$, can be easily computed, as follows.

A. Lower bound

Exploiting the concavity of the logarithmic function, an analytical lower bound is given by:

$$\begin{aligned}
&\bar{C}\left(d_i\right)=\mathbb{E}_{\psi_i,|h|^2}\left[B\log_2\left(1+\rho_i K\left(d_i/d_r\right)^{-\eta}\psi_i\,|h|^2\right)\right]\Leftrightarrow\\
&\bar{C}\left(d_i\right)=\mathbb{E}_{\psi_i}\left[\mathbb{E}_{|h|^2}\left[B\log_2\left(1+\rho_i K\left(d_i/d_r\right)^{-\eta}\psi_i\,|h|^2\right)\right]\right]\Leftrightarrow\\
&\bar{C}\left(d_i\right)\geq\mathbb{E}_{\psi_i}\left[\mathbb{E}_{|h|^2}\left[B\log_2\left(\rho_i K\left(d_i/d_r\right)^{-\eta}\psi_i\,|h|^2\right)\right]\right]\Leftrightarrow\\
&\bar{C}\left(d_i\right)\geq\underbrace{B\left\{\log_2\left(\rho_i K\left(d_i/d_r\right)^{-\eta}\right)-\gamma/\ln\left(2\right)\right\}}_{\bar{C}_L\left(d_i\right)}, \qquad (9)
\end{aligned}$$

where γ denotes the Euler-Mascheroni constant.

B. Upper bound

Using Jensen's inequality for concave functions, the upper bound is computed analytically as follows:

$$\begin{aligned}
&\bar{C}\left(d_i\right)=\mathbb{E}_{\psi_i,|h|^2}\left[B\log_2\left(1+\rho_i K\left(d_i/d_r\right)^{-\eta}\psi_i\,|h|^2\right)\right]\Leftrightarrow\\
&\bar{C}\left(d_i\right)=\mathbb{E}_{\psi_i}\left[\mathbb{E}_{|h|^2}\left[B\log_2\left(1+\rho_i K\left(d_i/d_r\right)^{-\eta}\psi_i\,|h|^2\right)\right]\right]\Leftrightarrow\\
&\bar{C}\left(d_i\right)\leq\underbrace{B\log_2\left(1+\rho_i K\left(d_i/d_r\right)^{-\eta}\mu_{\psi_i}\right)}_{\bar{C}_U\left(d_i\right)}, \qquad (10)
\end{aligned}$$

where $\mu_{\psi_i}=\mathbb{E}\left[\psi_i\right]=\exp\left[\sigma^2_{\text{dB},i}/2\xi^2\right]$.

Appendix B
Capacity analysis-II

The picocell points x,y are distributed uniformly over a circle according to the following distribution:

$$f_{X,Y}(x,y)=\begin{cases}\frac{1}{\pi\left(R^2_{p_j}-d^2_r\right)} & ,\ d_r\leq r\leq R_{p_j},\\ 0 & ,\ o/w,\end{cases} \qquad (11)$$

where $r=\sqrt{(x-x_{p_j})^2+(y-y_{p_j})^2}$. Average capacity bounds are obtained by averaging over all the possible distances (i.e., over all the possible points) within the picocell on the respective lower and upper capacity bound for each cell i. This can be analytically described as:

$$\bar{C}^i_b=\iint\limits_{XY}\bar{C}_b\left(d_i\left(x,y\right)\right)f_{X,Y}\left(x,y\right)dxdy,\ b\in\{L,U\}. \qquad (12)$$

A. Lower bound

The average lower bound of the average capacity by applying cartesian to polar coordinates transformation in Eq. (12) is computed as:

$$\begin{aligned}
\bar{C}^i_L&=\frac{1}{\pi\left(R^2_{p_j}-d^2_r\right)}\int_0^{2\pi}\int_{d_r}^{R_{p_j}}r\bar{C}_L\left(\tilde{d}_i\left(r,\theta\right)\right)drd\theta=\\
&=B\left[\log_2\left(\rho_i\right)+\right.\\
&+\frac{1}{\pi\left(R^2_{p_j}-d^2_r\right)}\underbrace{\int_0^{2\pi}\int_{d_r}^{R_{p_j}}r\log_2\left(K\left[\tilde{d}_i\left(r,\theta\right)/d_r\right]^{-\eta}\right)drd\theta}_{J_1}\\
&\left.-\gamma/\ln\left(2\right)\right]. \qquad (13)
\end{aligned}$$

B. Upper bound

Applying the same methodology used for the lower bound, the corresponding average upper bound is given by:

$$\begin{aligned}
\bar{C}^i_U&=\frac{1}{\pi\left(R^2_{p_j}-d^2_r\right)}\int_0^{2\pi}\int_{d_r}^{R_{p_j}}r\bar{C}_U\left(\tilde{d}_i\left(r,\theta\right)\right)drd\theta\\
&=\frac{B}{\pi\left(R^2_{p_j}-d^2_r\right)}\times\\
&\times\underbrace{\int_0^{2\pi}\int_{d_r}^{R_{p_j}}r\log_2\left(1+\rho_i K\left[\tilde{d}_i\left(r,\theta\right)/d_r\right]^{-\eta}\mu_{\psi_i}\right)drd\theta}_{J_2}.
\end{aligned} \qquad (14)$$

The function $\tilde{d}_i\left(r,\theta\right)$ is given by:

$$\tilde{d}_i\left(r,\theta\right)=\begin{cases}\sqrt{\left[r\cos\left(\theta\right)-x^*\right]^2+\left[r\sin\left(\theta\right)-y^*\right]^2} & ,\ i=m,\\ r & ,\ i=p_j,\end{cases} \qquad (15)$$

where $x^*=x_m-x_{p_j}$ and $y^*=y_m-y_{p_j}$.

This function represents the user distance from the macrocell and the picocell BS in polar coordinates, when the user is located within the picocell. Finally, the J_1 and J_2 integrals are numerically computed.

[8] It is noted that the capacity is based on the received signal before the filtering process, thus Rayleigh fading effect is taken into account. [9] SINR could also be introduced, but would make the analysis complex. [10] N_0^{dBm} in Table I stands for the N_0 in dBm scale. [11] $x\sim\text{Log-}\mathcal{N}\left(\mu,\sigma^2\right)$ denotes that random variable x is distributed according to the log-normal distribution with parameters μ and σ. [12] $x\sim\mathcal{CN}\left(\mu,\sigma^2\right)$ denotes that random variable x is complex Gaussian with mean μ and variance σ^2.