Hybrid User Association with Proactive Auxiliary Intervention for Multitier Cellular Networks

Antti Anttonen, Aarne Mämmelä, and Tao Chen
VTT Technical Research Centre of Finland Ltd
P.O. Box 1100, FI-90571 Oulu, Finland
{antti.anttonen, aarne.mammela, tao.chen}@vtt.fi

***Abstract*—In this paper, we consider a hybrid user association (HUA) problem for load balancing of multitier cellular networks. The proposed hierarchical HUA approach builds on a combination of decentralized user association (DUA) and auxiliary intervention of a central control unit (CCU). A major challenge with the CCU intervention is the time interval determined by a selected CCU control cycle during which the DUA must accept all users that satisfy the prevailing association criterion while proactively mitigating potential resource depletions. Consequently, the primary focus of this work is on relating the control cycle of the CCU intervention with the incipient resource depletions, according to a maximum allowed resource depletion probability. By uniquely combining a set of mathematical tools from stochastic geometry and queueing theory, we present a novel HUA method which evaluates the association bias values of the DUA according to a CCU-optimized load vector and enables tier-based resource depletion probability provisioning over finite control cycles. The trade-offs between the proposed HUA method and the standard DUA approach are demonstrated via network simulations with flow-level spatiotemporal dynamics.**

***Index Terms*—Load balancing, network heterogeneity, predictive congestion control, queueing theory, stochastic geometry.**

I. Introduction

User association (UA), where a user is connected with a particular base station (BS), has become an evolving research topic after the multitier network concept emerged as one of the key enabling technologies of a modern cellular network [1], [2]. In essence, hierarchical control is vital for the next-generation networks to ensure high resource efficiency [3]. The multitier concept represents a specific hierarchical control structure where BSs exhibit heterogeneous features regarding, e.g., their transmit power and spatial density. The cellular multitier architecture, e.g. the two-tier architecture with macro BSs and pico or femto BSs, can readily be extended to different access technologies, such as local area networks [4].

The UA approaches and their decision metrics for heterogeneous multitier networks are largely determined by their control centralization degree and modeling assumptions of the network topology [5]. Fully centralized control or decentralized control schemes lead to well-known problems such as extensive amount of required control data and optimization convergence problems, respectively. As a result, numerous hybrid user association (HUA) approaches have been proposed to mitigate these drawbacks. Distributed schemes enable near-optimal UA without using centralized control [6], [7]. Auxiliary centralized control has been used to mitigate the convergence implications of iterative distributed schemes [8], [9]. The word auxiliary means that the users are in charge of making UA decisions but the decisions are affected by central information of the slowly changing load conditions of the network. In particular, users can be biased away from a BS offering the maximum received signal power to obtain a more balanced network load. Different irregular network topology and load modeling frameworks have been further proposed addressing spatial randomness [1], [4], temporal randomness [10], [11], and comprehensive spatiotemporal randomness [12]–[15]. The inclusion of spatiotemporal randomness is seen important for the next-generation networks to allow adequate performance predictability and proactive resource control.

In this work, we focus on the HUA using an auxiliary central control unit (CCU) to coordinate user association decisions for dynamic load balancing in a spatiotemporal random multitier network model. Contrary to the aforementioned related work, the primary focus of this paper is on relating the expected time interval, which is determined by the selected CCU intervention cycle, with the resource depletion probability (RDP) in the BSs. Conventionally used steady state queueing analysis approaches abstract away the dependence of RDP on time. Therefore, we analyze finite-horizon transient congestion by monitoring first-passage resource depletion events rather than steady state congestion in order to incorporate the effect of CCU cycle on the maximum allowed RDP.

By combining tools from stochastic geometry and queueing theory, we present a novel HUA method which evaluates the association bias values of the decentralized user association (DUA) in order to proactively mitigate potential resource depletions. The HUA approach enables RDP provisioning over finite control cycles with the help of a CCU-optimized load vector. The trade-offs between the proposed HUA method and the standard decentralized approach are demonstrated via network performance simulations with flow-level spatiotemporal dynamics. These include the RDP analysis of the resource blocks (RBs) of a typical BS and outage probability of signal-to-interference-and-noise ratio (SINR) of a typical user.

The rest of the paper is organized as follows. The target system model is presented in Section II. The analytical RDP evaluation methods are given in Section III. The proposed hybrid control approach is described in Section IV, followed by some illustrative numerical examples in Section V. Finally, the conclusions are given in Section VI.

II. System Overview

A. Spatiotemporal Multitier Network Model

Following the spatiotemporal modeling approach from [12]–[15], we consider a region of a two-dimensional space $\mathcal{L}$ where a set of user equipment (UE) is served by BSs $\mathcal{B}_k$ from the kth tier of a multitier network where $k \in (1, 2, \ldots, K)$. The kth tier BS units are randomly located following independent homogeneous Poisson point processes (PPPs) Φ_k with an intensity ϕ_k determined by BS density in units/km^2. We focus on the downlink of the multitier network where each BS of the kth tier transmits with power P_k. The total bandwidth $W = N_{\text{rb}} W_{\text{rb}}$ is divided into N_{rb} RBs of width W_{rb}.

For a given realization of Φ_k, new data flow requests from the UE set $\mathcal{U}$ arrive independently and fall within the space $\mathcal{L}$ and time window $(t_0, t_0 + t)$. We address low mobility hotspot scenarios where it is assumed that the locations of BSs and users do not change significantly during the selected finite time window. The arrival process of the UE flow units follows a space-time PPP Θ with a spatiotemporal intensity θ determined by UE flow density in units/(s$\cdot$km^2). The temporal mean UE flow arrival rate of the network is $\lambda_{\text{net}} = \theta\mathcal{L}$. Let (k, i) be the pair denoting the ith BS of the kth tier used for short notation as (k, i)th BS and $(\cdot)_{ki}$ for subindexing. Using the results from [1], [14], [15], the UE arrival process after the UA in the (k, i)th BS is Poisson with intensity $\lambda_{ki} = p_{ki}\theta$ determined by units/s where p_{ki} is the scaling factor according to the effective coverage area of the (k, i)th BS.

We assume that the departure or service times of the UE flows from the (k, i)th BS are iid random variables for which only the mean μ_{ki} and the variance σ^2_{ki} are needed to be known or estimated (cf. [14], [15]). The moments of the flow departure times depend on the prevailing flow file size and cell data rate in different tiers. For a given SINR and allocated bandwidth, the data rate can be evaluated using the practical truncated and weighted Shannon capacity equation from [16]. The possible coupling between the UE departure times of the BSs can be incorporated via the BS-specific moments which corresponds to commonly used time-averaged interference method [17], [18].

B. BS Load Model

Following the approach from [10], the load state of a BS is modeled as the number of remaining RBs in the resource pool yet to be scheduled at time $t_0 + t$. The number of available RBs in the (k, i)th BS can be represented as

$$q_{ki}(t_0+t) = \min\{\max\{b_{ki}+x_{ki}(t_0+t)-y_{ki}(t_0+t), 0\}, B_{ki}\} \quad (1)$$

where B_{ki} is the maximum number of RBs, $b_{ki} > 0$ is the initial number of remaining RBs at time t_0, and $b_{ki} \leq B_{ki} \leq N_{\text{rb}}$. The random processes $x_{ki}(t)$ and $y_{ki}(t)$ denote, respectively, the cumulative numbers of RB arrivals and RB departures in the (k, i)th BS. Without the loss of generality, we can assume that $t_0 = 0$. If $B_{ki} = N_{\text{rb}}, \forall k, i$, the spectrum is fully reused by the BSs with the reuse factor of one. We assume that B_{ki} is a large number which complies with the current trend of modern cellular systems where the frequency ranges have rapidly increased [5]. However, a large B_{ki} does not prevent congestion in hotspot scenarios where also a high number of users is requesting a service. If there are no UE associated in the (k, i)th BS, $q_{ki}(t) = B_{ki}$. On the other hand, if $q_{ki}(t) = 0$, a resource depletion event is declared.

It is obvious that the arrival process of the available RBs is determined by the departure process of the set of UE from the cell whereas the RB departure process is determined by the UE arrival process. If the number of allocated RBs per UE is one for all users, there is one-to-one correspondence between an RB departure and a UE arrival. However, there can be M UE service classes where each class $m \in (1, 2, \ldots, M)$ determines the required number of RBs per UE arrival Q_m, according to a selected bandwidth allocation strategy summarized, e.g., in [19]. The mean and variance of the requested number of RBs per arriving UE (i.e. a batch arrival of RBs) are denoted as ϑ_{ki} and ν^2_{ki}, respectively. In Section III, we link these moment parameters together.

C. User SINR Model

The downlink signal between the (k, i)th BS and the jth UE experiences path loss and random channel fading so that the instantaneous SINR within a single RB is represented as (cf. [1], [7], [13])

$$\gamma_{kij} = \frac{P_k h_{kij} d_{kij}^{-\eta_k}}{\sum_l \sum_{m \in \mathcal{B}_l \setminus \{i|l=k\}} \kappa_{lm} P_{lmj} h_{lmj} d_{lmj}^{-\eta_l} + \frac{W_{\text{rb}} N_0}{L_0}} \quad (2)$$

where h_{kij} denotes the random channel fluctuation coefficient, d_{kij} is the distance between the jth UE and the (k, i)th BS, $\kappa_{ki} \in (0, 1)$ is a binary variable indicating if the BS is transmitting on a given RB, η_k is the path loss exponent, L_0 is the path loss at a reference distance, and N_0 is the noise power spectral density. The averaged long-term SINR can be obtained by averaging γ_{kij} over a desired time-frequency range.

III. Evaluation of Finite-Horizon RDP in Heterogeneous Networks

Here we present the RDP analysis methods in different hierarchy levels (i.e. BS, tier, and network) and relate the user arrival rate with the maximum allowable RDP and finite horizon. The results are then applied to the HUA approach in Section IV as well as numerical verification in Section V.

A. RDP Analysis at a BS Level

In essence, the RDP acts as a statistical quality of service parameter to indicate if the service is going to be temporally blocked due to lack of RBs in the BSs. It is extremely difficult to obtain exact analytical approaches for the finite-horizon RDP with the RB model presented in Section II. Consequently, we apply the diffusion approximation approach which has been successfully used to find a closed form solution to many multidisciplinary queueing theoretic problems [20]. In this approach, a discrete queue model is replaced by a continuous-time Brownian model which is characterized with the first- and second-order moments of the arrival and departure processes

of the queue. Our analysis problem is in part similar to that of a video streaming application in [21] where the evolution of video packets is studied instead of RBs. Using the notation presented in Section II, the RDP over a finite time horizon T of the (k,i)th BS can be represented as (cf. [21])

$$\begin{aligned}\mathcal{P}_{ki} &= \Pr(\tau_{ki} < T) \\ &= \int_0^T \frac{b_{ki}}{\sqrt{2\pi\alpha_{ki}t^3}} \exp\left[-\frac{(\beta_{ki}t + b_{ki})^2}{2\alpha_{ki}t}\right] \mathrm{d}t \qquad (3) \\ &\leq \exp\left(-\frac{2\beta_{ki}b_{ki}T + b_{ki}^2}{2\alpha_{ki}T}\right) \qquad (4)\end{aligned}$$

where $\tau_{ki} = \inf(t \geq 0 | q_{ki}(t) = 0, q_{ki}(0) = b_{ki})$ is the first passage time of $q_{ki}(0) > 0$ to $q_{ki}(t) = 0$, $\inf(\cdot)$ is the infimum function, and β_{ki} and α_{ki} are, respectively, the mean and variance of the applied RB diffusion process. However, [21] does not consider the batch queueing required for our RB evolution model. Fortunately, this limitation can be overcome by redefining the diffusion moments β_{ki} and α_{ki} as follows.

Lemma 1: For the RB queue model presented in Section II, the diffusion moments in (3) and (4) are given as

$$\beta_{ki} = \vartheta_{ki}(\mu_{ki} - \lambda_{ki}), \qquad (5)$$

$$\alpha_{ki} = \nu_{ki}^2\mu_{ki} + \vartheta_{ki}^2\sigma_{ki}^2\mu_{ki}^3 + \nu_{ki}^2\lambda_{ki} + \vartheta_{ki}^2\lambda_{ki}. \qquad (6)$$

Proof: To find the diffusion moments β_{ki} and α_{ki} that are adequate for the batch RB queue model at hand, we apply the results presented in [22]. By using the assumptions that the statistics of the RB arrival and departure processes of the batch sizes are the same and the UE arrivals defined in Section II-A follow a Poisson process, it is easy to show that the moments presented in [22, Eqs. (7) and (8)] reduce, respectively, to (5) and (6). ∎

B. Maximum User Arrival Rate

Proposition 1: Let $0 < \rho_{ki} < 1$ denote the maximum allowed RDP for the (k,i)th BS within time horizon T and define $\bar{\rho}_{ki} = \ln(\rho_{ki})$. If the UE arrival rate $\lambda_{ki} \leq \Lambda_{ki}$, where

$$\Lambda_{ki} = \frac{b_{ki}\vartheta_{ki}\mu_{ki} + b_{ki}^2(2T)^{-1} + \bar{\rho}_{ki}\nu_{ki}^2\mu_{ki} - \bar{\rho}_{ki}\vartheta_{ki}^2\sigma_{ki}^2\mu_{ki}^3}{b_{ki}\vartheta_{ki} - \bar{\rho}_{ki}\nu_{ki}^2 - \bar{\rho}_{ki}\vartheta_{ki}^2}, \qquad (7)$$

then $\mathcal{P}_{ki} \leq \rho_{ki}$.

Proof: By substituting (5) and (6) into (4) and some rearrangement of the terms, we obtain the desired result in (7). ∎

Corollary 1: If $\Lambda_{ki} \leq 0$ in (7), the target RDP ρ_{ki} cannot be met for any arrivals in the (k,i)th BS during the next time horizon T.

C. Effect of Spatially Varying Cell Sizes

In order to further elaborate how the finite-horizon RDP of BSs behaves in a spatiotemporal multitier network, we can evaluate an averaged network level result where the effect of spatially varying cell sizes of BSs within each tier is incorporated.

Lemma 2: Let $0 \leq p_k(n) \leq 1$ denote the conditional RDP calculated either from (3) or (4) for $N_k = n$ associated UE in a typical[1] BS of the kth tier during T. Specifically, $p_k(n) \coloneqq \mathcal{P}_{ki}$ conditioned that $\lambda_{ki} \coloneqq \lambda(n) = n/T$, and assuming the other involved BS parameters are fixed within each tier. Let $\mathcal{A}_k$ denote the probability that a typical UE is associated with the kth tier[2]. Then, the RDP of a typical BS in the kth tier of the defined multitier network model is given as

$$\mathcal{P}_k = \frac{\omega^\omega}{\Gamma(\omega)} \sum_{n=0}^{\infty} \frac{p_k(n)\Gamma(n+\omega)\xi_k^n\,(\omega + \xi_k)^{-n-\omega}}{\Gamma(n+1)} \qquad (8)$$

where $\xi_k = \frac{\theta T \mathcal{A}_k}{\phi_k}$, $\Gamma(\cdot)$ is the gamma function, and $\omega = 7/2$.

Proof: From the law of large numbers, for a Poisson process that is conditioned of having $n(t)$ arrivals in t seconds, the mean arrival rate is given as $\lim_{t\to\infty} n(t)/t$ for which $\lambda(n) = n/t$ serves as an approximation for finite t [23]. The tier level RDP $\mathcal{P}_k$ is found by deconditioning $p_k(n)$ on discrete random variable n as $\mathcal{P}_k = \sum_{n=0}^{\infty} p_k(n)\Pr(N_k = n)$ where $\Pr(N_k = n) = \int_0^\infty \Pr(N_k = n | X_k = x) f_{X_k}(x)\mathrm{d}x$ is the probability mass function of the number of UE N_k associated with a typical BS of the kth tier, $\Pr(N_k = n | X_k = x)$ denotes the Poisson distributed number of UE conditioned by the random kth tier cell area X_k, and $f_{X_k}(x)$ is the probability density function of X_k. We adopt the Poisson-Voronoi cell area evaluation approach presented in [4], [24] which leads to $\Pr(N_k = n) = \frac{\omega^\omega \Gamma(n+\omega+z)}{n!\Gamma(\omega)} \left(\frac{\theta_s \mathcal{A}_k}{\phi_k}\right)^n \left(\omega + \frac{\theta_s \mathcal{A}_k}{\phi_k}\right)^{-n-\omega-z}$ where θ_s is the mean number of UE in a given network area and $z \in \{0,1\}$ with $z = 1$ if the typical UE under investigation is assumed to be separated from n other UE, and $z = 0$, otherwise. In our system, we have $z = 0$ and $\theta_s = \theta T$, after which (8) is finally obtained. ∎

Proposition 2: Given the tier level RDPs from (8), the network level finite-horizon RDP is given as

$$\mathcal{P} = \frac{\omega^\omega}{\Gamma(\omega)\sum_k \phi_k} \sum_{k=1}^{K} \sum_{n=0}^{\infty} \frac{\phi_k p_k(n)\Gamma(n+\omega)\xi_k^n\,(\omega + \xi_k)^{-n-\omega}}{\Gamma(n+1)}. \qquad (9)$$

Proof: Using the law of total probability, which describes the average probability from several distinct processes, we obtain $\mathcal{P} = \sum_k \mathcal{P}_k \mathcal{T}_k$ where $\mathcal{T}_k$ is the probability that a typical BS is located in the kth tier and $\mathcal{P}_k$ is obtained from Lemma 2. By applying the superposition theorem for the Poisson point processes Φ_k, it is straightforward to show that $\mathcal{T}_k = \phi_k / \sum_k \phi_k$, which leads to (9). ∎

Remark 1: The right tail of $\Pr(N_k = n)$ is found to diminish rapidly already for $n > 4\xi_k$ so that $\sum_{n<4\xi_k} \Pr(N_k = n) \approx 1$. Therefore, the infinite sums in (8) and (9) are well approximated with $n < 4\xi_k$, $\forall k$.

IV. Proposed Hybrid User Association Approach

In this section, we present a novel HUA method which dynamically evaluates the association bias values of the fixed DUA approach according to a CCU-optimized load vector and enables tier-based RDP provisioning over finite control cycles.

[1]The term typical indicates that a BS is randomly selected.

[2]For the biased UA to be discussed in Section IV, $\mathcal{A}_k$ is presented in [1].

A. Decentralized User Association

We first outline the standard DUA approach, based on maximum average received power (RP) and fixed preassigned biasing, also known as the cell range extension method [1]. Specifically, in the DUA approach, the jth UE is associated with the (k,i)th BS that satisfies

$$\max_{k,i} \epsilon_k P^{\text{rx}}_{kij} \tag{10}$$

where ϵ_k is the fixed bias factor for the kth tier BSs and P^{rx}_{kij} is the average RP at the jth UE received from the (k,i)th BS which can be measured using a pilot signal transmitted by the BS. The unbiased DUA approach, where $\epsilon_k = 1, \forall k$ in (10), is the optimal UA for maximizing the SINR coverage if all BSs are transmitting [1]. For light traffic load scenarios, there is also a marginal improvement on SINR coverage via biasing [25]. However, the primary objective of the biasing is to balance the network load so that a UE is not blocked due to lack of RBs in the BSs [1], [26].

B. Hybrid User Association

The basic idea of the HUA approach is to split the UA process into two control loops, namely the slow CCU loop for updating the bias values ϵ_k in (10) and the fast DUA loop to perform the final UA decisions, see Fig. 1. The primary focus of the proposed HUA is on relating the expected cycle T of the CCU intervention with the maximum allowable RDP. The CCU control cycle T, which is decided and broadcasted by the CCU, is the time that the DUA must wait until the new bias values are activated. The proper selection of T is affected by the expected changes in the statistical system parameters presented in Sections II and III that affect the RB depletion. By ensuring a proper minimum value for the CCU control cycle T, the ping-pong problem (cf. [9]) can be mitigated. On the other hand, to avoid unnecessary bias updates and to effectively adapt the CCU control cycle, the CCU can employ a simple incipient congestion indicator as $I_{ki} = 1$, if $\lambda_{ki} > \Lambda_{ki}$ and $I_{ki} = 0$, otherwise. Specifically, if $I_{ki} = 0, \forall k, i$, a bias update is not needed for the prevailing control cycle. It is also obvious that if $I_{ki} = 1, \forall k, i$, a bias update cannot help significantly as the whole network is congested. The DUA decision loop cycle, on the other hand, is proportional to the mean UE interarrival time $\lambda_{\text{net}}^{-1}$. We make a reasonable assumption that $T \gg \lambda_{\text{net}}^{-1}$.

The association bias values ϵ_k in the CCU are found in two steps by incorporating results from the queueing theory (7) and stochastic geometry [1]. The first step evaluates the UE arrival rates per cell that satisfy the arrival rate constraint (7) while leading to a minimum number of offloadings with respect to the unbiased DUA. In other words, the step directs a UE to access a BS with maximum RP when there are enough RBs left to support a maximum allowed RDP. In the second step, the bias factors are found using the arrival rates from the first step. A pseudocode of the HUA approach is given in Fig. 2.

STEP 1: Let $N_k(\boldsymbol{\epsilon})$ be the average number of UE nodes associated to a typical BS of the kth tier during T and for the given bias vector $\boldsymbol{\epsilon} = \{\epsilon_k\}, k = 1, 2, \ldots, K$, where $\boldsymbol{\epsilon} = \mathbf{1}$ denotes the unbiased DUA. The desired UE arrival rates $\boldsymbol{\lambda}^* = \{\lambda^*_{ki}\}, i \in \mathcal{B}_k, k = 1, 2, \ldots, K$, among the set of new candidate arrival rates $\boldsymbol{\lambda}' = \{\lambda'_{ki}\}$, are found in the CCU by solving the following constrained least squares optimization problem

$$\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}'} \{\sum_{k,i} [N_k(\mathbf{1}) - \lambda'_{ki} T]^2\} \tag{11}$$

s.t.

$$0 \le \lambda'_{ki} \le \Lambda_{ki}, \quad \forall k, i \tag{12}$$

$$\sum_{k,i} \lambda'_{ki} = \lambda_{\text{net}}. \tag{13}$$

STEP 2: Using the arrival rates from STEP 1, the corresponding nonnegative bias factors $\boldsymbol{\epsilon}^* = \{\epsilon^*_k\}, k = 1, 2, \ldots, K$, are found by minimizing the sum of squared errors as

$$\boldsymbol{\epsilon}^* = \arg\min_{\boldsymbol{\epsilon}} \{\sum_{k,i} [N_k(\boldsymbol{\epsilon}) - \lambda^*_{ki} T]^2\}. \tag{14}$$

The above minimization problems can be solved using available least squares numerical methods [27]. There are a few options to evaluate $N_k(\boldsymbol{\epsilon})$, including a stochastic geometry based approach [1] and an RP-based approach [26]. In this

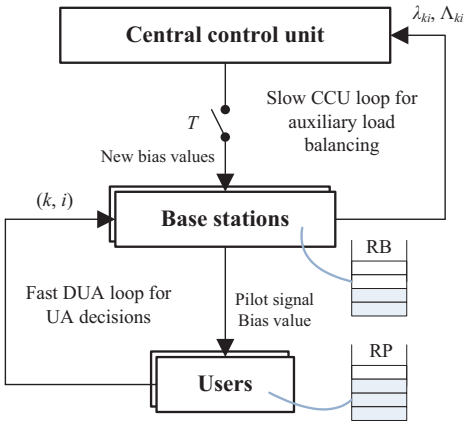

Fig. 1. Outline of the proposed proactive HUA approach.

1: Initialization: $\epsilon_k = 1, \forall k$
2: **for** each CCU control cycle with duration T **do**
3: Observe λ_{ki} and Λ_{ki} in each BS and inform the CCU
4: **if** $I_{ki} = 1$ for any k, i **then**
5: Trigger the CCU to obtain new bias values
6: Update $\lambda_{\text{net}} = \sum_{k,i} \lambda_{ki}$
7: Calculate new arrival rates $\boldsymbol{\lambda}^*$ using STEP 1
8: Calculate new bias values $\boldsymbol{\epsilon}^*$ using STEP 2
9: Deliver new bias values to BSs
10: **end if**
11: Perform the DUA with the updated biases using (10)
12: **end for**

Fig. 2. Pseudocode for the proposed proactive HUA approach.

work, we apply the former method from [1] as

$$N_k(\boldsymbol{\epsilon}) = 2\pi\theta T \times \int_0^\infty r\exp\{-\pi\sum_l \phi_l[P_l\epsilon_l/(P_k\epsilon_k)]^{2/\eta_l} r^{2\eta_k/\eta_l}\}\mathrm{d}r \overset{\{\eta_l\}=\eta}{=} \frac{\theta T}{\phi_k + \sum_{l,l\neq k}\phi_l[P_l\epsilon_l/(P_k\epsilon_k)]^{2/\eta}}. \quad (15)$$

V. Numerical Results

In the following numerical examples, we focus on a two-tier network with $K = 2$ in order to obtain presentable plots of tier-specific results for different UA methods. All connection requests are associated without any UE blocking prior potential resource depletions. The selected network area is $\mathcal{L} = \pi r^2$ with radius $r = 1000$ m. The tier-based transmission powers of the BSs are set to $\{P_1, P_2\} = \{53, 33\}$ dBm. It is assumed that there are $N_{\text{rb}} = 400$ RBs to be allocated with W_{rb} = 180 kHz and cell frequency reuse factor of one. The SINR outage threshold is set to 0 dB and Rayleigh fading models the random fluctuation of the channel. Other parameter values in the SINR model include $L_0 = -38.5$ dB, N_0 = -174 dBm/Hz, and $\{\eta_1, \eta_2\} = \{3, 3.5\}$. In order to demonstrate the inherent load balancing effects, we assign different initial load conditions for each tier as $\{b_{1i}, b_{2i}\} = \{400, 100\}$. The tier-dependent UE interdeparture times are assumed to be exponential with rates $\{\mu_{1i}, \mu_{2i}\} = \{0.1, 0.4\}$, and variance $\{\sigma_{1i}^2, \sigma_{2i}^2\} = \{100, 6.25\}$. To demonstrate the random RB batch process, we use three UE service classes, each class requiring $\{Q_1, Q_2, Q_3\} = \{1, 5, 10\}$ RBs per flow arrival.

In Fig. 3, we first illustrate the RDP analysis methods at the network level against the Monte Carlo simulation results with $\phi_1 = 1/\mathcal{L}$ and $T = 10$ s. Although (4) is slightly more inaccurate than (3), it yet provides an upper bound for (3) and, therefore, retains the RDP below the target threshold. It is also seen how the RDP increases with increasing UE flow density θ while the RDP decreases with the increasing BS density ϕ_2. In Fig. 4, we then compare the simulated tier-level RDPs of the DUA and HUA as a function of θ with two different control

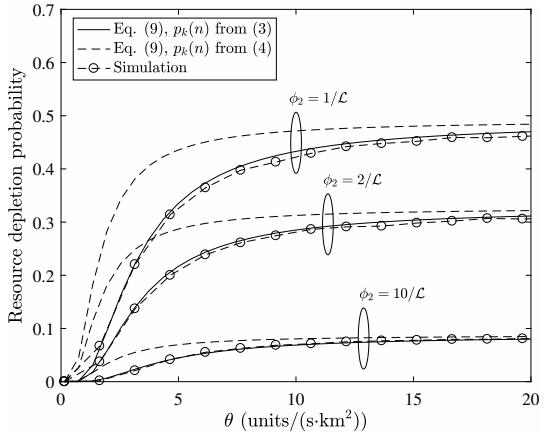

Fig. 3. Assessment of RDP evaluation methods.

cycles $T = 10$ s and $T = 60$. For the DUA, typical fixed bias factors $\{\epsilon_1, \epsilon_2\} = \{0, 10\}$ dB are used and we set $\{\phi_1, \phi_2\} = \{1/\mathcal{L}, 10/\mathcal{L}\}$. The target RDP per tier is set to be $\{\rho_{1i}, \rho_{2i}\} = \{0.1, 0.01\}$. It is seen that the DUA clearly exceeds the target maximum RDP by allowing too many UE to associate with the tier-1 BSs while the HUA approach better supports the RDP. The demonstration shows that load balancing is effective if the network is unbalanced but not fully congested and there is a sufficiently short CCU cycle (cf. the case with $T = 10$ s). If the network is fully congested or the CCU cycle is too long (cf. the case with $T = 60$ s), it is clear that not all UE can be served at the time and their connection should be blocked in order to support the target RDP for the remaining users.

The probability of UE association is then evaluated in Fig. 5 to further reveal the differences between the DUA with the preassigned biases and HUA with the RDP-aware biases. It is seen that as the value of θ is increased, the association probability per tier remains constant for the DUA whereas the HUA has the ability to modify the association probability between tiers to avoid exceeding the target RDP.

Finally, the probability of SINR outage is studied in Fig. 6. It is seen that the biasing increases the SINR outage probability, as the users are no longer allowed to associate with the BS offering the maximum received power. However, as seen from Fig. 6, the HUA approach can reduce the SINR outages compared to the DUA by avoiding unnecessary offloadings in the case the given RDP is satisfied. In order to illustrate the differences between the DUA and HUA approaches, we have assumed the full cell frequency reuse which naturally leads to relatively high SINR outage probability. The interference problem can be mitigated by orthogonal spectrum allocation between tiers or using other existing intercell interference coordination approaches (cf. [26]).

VI. Conclusions

In this paper, we have presented a novel HUA method which dynamically evaluates the association bias values of the DUA by using a combined mathematical framework from stochastic geometry and finite-horizon queueing theory. The main benefit of the approach is to enable proactive resource depletion probability provisioning over finite CCU control cycles. The proposed HUA approach uniquely abstracts a number of BS-specific parameters into a single parameter of maximum arrival rate target that is communicated to the CCU. Consequently, the CCU does not require instantaneous UE-specific information which would result in an extensive amount of control information. We further demonstrated the inherent trade-off of the load balancing which leads to a potential decrease of the SINR coverage for the offloaded users. In the future work, this trade-off will be further investigated by using more sophisticated intercell interference avoidance methods.

References

[1] H.-S. Jo, Y. Sang, P. Xia, and J. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.

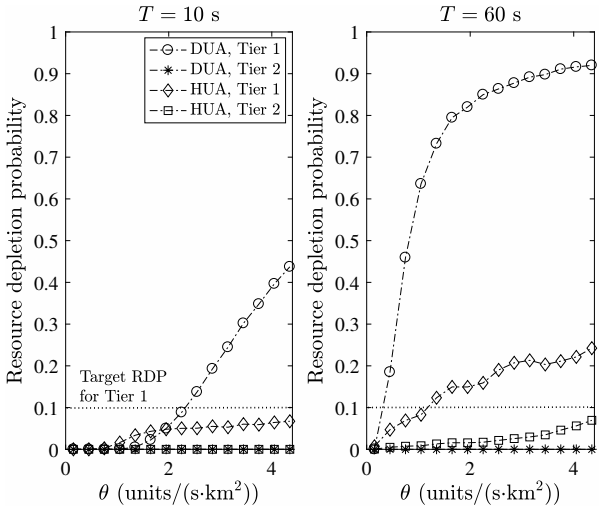

Fig. 4. Comparison of RDP for different UA methods.

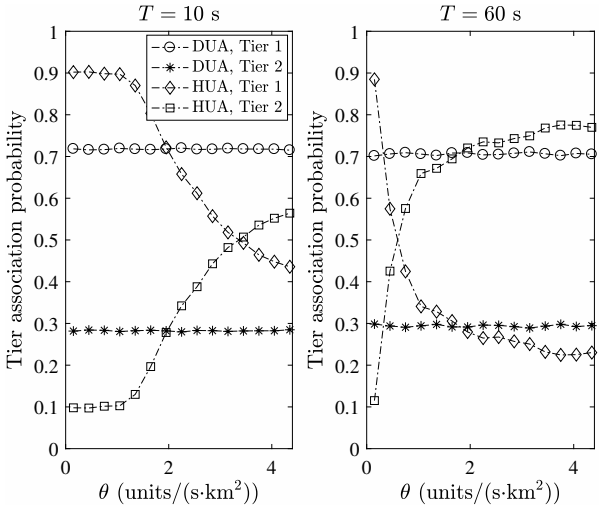

Fig. 5. Comparison of tier association probability for different UA methods.

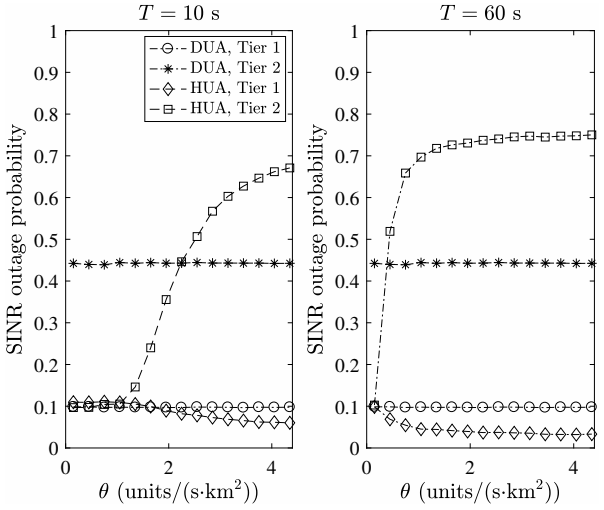

Fig. 6. Comparison of SINR outage probability for different UA methods.

[2] T. Chen, H. Zhang, X. Chen, and O. Tirkkonen, "SoftMobile: Control evolution for future heterogeneous mobile networks," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 70–78, Dec. 2014.

[3] A. Mämmelä, J. Riekki, A. Kotelba, and A. Anttonen, "Multidisciplinary and historical perspectives for developing intelligent and resource-efficient systems," *IEEE Access*, to be published.

[4] S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[5] D. Liu *et al.*, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys & Tut.*, vol. 18, no. 2, pp. 1018–1044, 2nd quarter 2016.

[6] Q. Ye *et al.*, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[7] H. Boostanimehr and V. Bhargava, "Unified and distributed QoS-driven cell association algorithms," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1650–1662, Mar. 2015.

[8] M. Haddad, S. Elayoubi, E. Altman, and Z. Altman, "A hybrid approach for radio resource management in heterogeneous cognitive networks," *IEEE J. Sel. Areas in Commun.*, vol. 29, no. 4, pp. 831–842, Apr. 2011.

[9] H. Du *et al.*, "A load fairness aware cell association for centralized heterogeneous networks," in *Proc. Int. Conf. Commun.*, London, UK, 2015, pp. 2178–2183.

[10] X. Luo *et al.*, "An adaptive measured-based preassignment scheme with connection-level QoS support for mobile networks," *IEEE Trans. Wireless Commun.*, vol. 1, no. 3, pp. 521–530, Jul. 2002.

[11] O. Tonguz and E. Yanmaz, "The mathematical theory of dynamic load balancing in cellular networks," *IEEE Trans. Mobile Comp.*, vol. 7, no. 12, pp. 1504–1518, Dec. 2008.

[12] W. Bao and B. Liang, "Near-optimal spectrum allocation in multi-tier cellular networks with random inelastic traffic," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Process.*, Florence, Italy, 2014, pp. 855–859.

[13] P. Mankar, B. Sahu, and S. Pathak, "Evaluation of blocking probability for downlink in Poisson networks," *IEEE Wireless Commun. Lett.*, vol. 4, no. 6, pp. 625–628, Dec. 2015.

[14] A. Shojaeifard *et al.*, "On the design of irregular HetNets with flow-level traffic dynamics," in *Proc. IEEE Veh. Tech. Conf.*, Montreal, Canada, 2016, pp. 1–5.

[15] K. Stamatiou and M. Haenggi, "Traffic management in random cellular networks," in *Proc. Inf. Theory and Applications*, San Diego, CA, USA, 2014, pp. 1–5.

[16] 3GPP standardization, "Radio frequency (RF) system scenarios (Release 14)," TR 36.942, v14.0.0, 2017.

[17] H. Klessig, D. Ohmann, A. Fehske, and G. Fettweis, "A performance evaluation framework for interference-coupled cellular data networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 938–950, Feb. 2016.

[18] H. Kim, G. Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE Trans. Networking*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

[19] I. Viering, M. Dottling, and A. Lobinger, "A mathematical perspective of self-optimizing wireless networks," in *Proc. Int. Conf. Commun.*, Dresden, Germany, 2009, pp. 1–6.

[20] A. Duda, "Diffusion approximation for time-dependent queueing systems," *IEEE Sel. Areas Commun.*, vol. 4, no. 6, pp. 905–918, Sep. 1986.

[21] A. Anttonen and A. Mämmelä, "Interruption probability of wireless video streaming with limited video lengths," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1176–1180, Jun. 2014.

[22] S. Chiamsiri and M. Leonard, "A diffusion approximation for bulk queues," *Management Science*, vol. 27, no. 10, pp. 1188–1199, Oct. 1981.

[23] U. Bhat, *An Introduction to Queueing Theory*. Boston, MA, USA: Springer, 2008.

[24] S. Yu and S.-L. Kim, "Downlink capacity and base station density in cellular networks," in *Proc. IEEE WiOpt*, Tsukuba, Japan, 2014, pp. 119–124.

[25] H. Tang, J. Peng, P. Hong, and K. Xue, "Offloading performance of range expansion in picocell networks," *IEEE Wireless Commun. Lett.*, vol. 2, no. 5, pp. 511–514, Oct. 2013.

[26] S. Deb, J. Miernik, and J. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets," *IEEE Trans. Networking*, vol. 22, no. 1, pp. 137–150, Feb. 2014.

[27] Å. Björck, *Numerical Methods for Least Squares Problems*. Philadelphia, PA, USA: Society for Industrial and Applied Math., 1996.