Mobility Prediction empowered Proactive Energy Saving Framework for 5G Ultra-Dense HetNets

Hasan Farooq, Ahmad Asghar and Ali Imran
University of Oklahoma, Tulsa, USA 74135
Email: {hasan.farooq, ahmad.asghar, ali.imran}@ou.edu

***Abstract*—Increased network wide energy consumption is a paramount challenge that hinders wide scale ultra-dense networks (UDN) deployments. While several Energy Saving (ES) enhancement schemes have been proposed recently, these schemes have one common tenancy. They operate in reactive mode i.e., to increase ES, cells are switched ON/OFF reactively in response to changing cell loads. Though, significant ES gains have been reported for such ON/OFF schemes, the inherent reactiveness of these ES schemes limits their ability to meet the extremely low latency and high QoS expected from future cellular networks vis-a-vis 5G and beyond. To address this challenge, in this paper we propose a novel user mobility prediction based AUtonomous pROactive eneRgy sAving (AURORA) framework for future UDN. Instead of observing changes in cell loads passively and then reacting to them, AURORA uses past hand over (HO) traces to determine future cell loads. This prediction is then used to proactively schedule small cell sleep cycles. AURORA also incorporates the effect of Cell Individual Offsets (CIOs) for balancing load among cells to ensure QoS while maximizing ES. Extensive system level simulations leveraging realistic SLAW model based mobility traces show that AURORA can achieve significant energy reduction gain without noticeable impact on QoS.**

***Index Terms*—5G, Energy Saving, Mobility Prediction, Proactive SON, Heterogeneous Networks, Sleeping Cells, ON/OFF Small Cells, CIOs.**

I. Introduction

Network densification is currently being treated with a mix of anticipation for its promise of addressing the capacity crunch - and concern for its impact on the energy consumption. This is due to the high aggregated network energy that "always ON" small cells (SCs) are bound to consume in an Ultra Dense Network (UDN). In addition to higher carbon footprint, this translates into higher OPEX. Although SCs have a relatively lower power consumption profile, the always ON approach increases overall network wide energy consumption [1]. As a result, with advent of UDN, the need for ES schemes will be even more compelling. The initial ambivalence about UDN has been replaced by consensus that to avert possible energy crunch in 5G, the 1000x capacity increase must be achieved at a similar or lower power consumption as legacy networks [2]. Energy consumption in cellular systems can be reduced either by optimizing resource allocation such that minimum energy is consumed per bit transmission or by turning OFF underutilized cells during offpeak hours [2]–[5]. To exploit these approaches recently ES has been adopted as a key Self Organizing Network (SON) function by 3GPP [6] and has been extensively studied in literature. The resource allocation optimization can reduce the energy consumption to only a limited degree for a given system throughput target. ES of the cellular systems can be further enhanced significantly by switching under-utilized BSs to sleep mode or turning them OFF entirely during off-peak time [3]–[5], [7]. In this direction of research, several recent works show promising results in terms of potential ES. A detailed survey of current ES schemes can be found in [2]. However, to the best of our knowledge, existing ES approaches fall short of the mark for 5G requirement due to following limitations:

1) **Reactive mode of operation**: Conventional ES SON algorithms are designed to switch OFF/ON cells after detecting network conditions that have already taken effect. However, given the acute dynamics of traffic and cellular environment, by the time triggering conditions are detected and a realistic non-convex NP-hard ES algorithm is solved to produce new network ON/OFF configuration optimal for observed network conditions, the conditions may already change. Thus, the newly determined configuration is likely to be suboptimal before it can be actuated. This problem can particularly exacerbate in 5G.
2) **Difficulty in meeting 5G low latency:** Base Stations require a certain amount of time to wake up from sleep cycle. For a user entering a sleeping cell, this time to wake up will negatively impact latency observed by the users.
3) **SON conflict prone design:** Conventional ES solutions do not take SON conflicts into account. Two SON use cases that become highly relevant to the ES in HetNets are Coverage and Capacity Optimization (CCO) and Load Balancing (LB) [6] because of the overlap among their optimization parameter set: transmission power and cell individual offsets (CIOs). When an ES switches OFF some cells, it may force some users to be associated to neighboring ON cells and overload them thereby conflicting with CCO and LB SON functions. As explicated in [8], such conflict prone ES solution design can actually degrade network's performance instead of improving it.

To address the aforementioned limitations, we propose a novel ES solution called AURORA. AURORA builds on the Big Data empowered SON framework presented in [9]. The key idea is to make emerging cellular systems artificially intelligent and autonomous so that they can anticipate user mobility behavior. This intelligence in turn is then used to

978-1-5386-4727-1/18/$31.00 ©2018 IEEE

formulate a novel ES optimization problem that proactively schedules small cell sleep cycles to divert and focus the right amount of resources when and where needed while satisfying QoS requirements. The contributions of paper can be summarized as follows:

1) We develop and analyze a semi-Markov model based spatio-temporal mobility prediction framework and further propose a novel method to map the next cell spatiotemporal HO information to the estimated future location coordinates based on the idea of Landmarks. This information is then transformed into future cell loads.
2) Based on the intelligence gained from the mobility model i.e., future cell loads, a proactive energy saving optimization problem is formulated to minimize the energy consumption by switching OFF underutilized SCs. Another key novelty is that AURORA leverages CIOs as optimization variables for avoiding overloading scenarios while deciding which cells to switch ON/OFF.
3) We perform a comparative analysis of proposed solution, through multi-tier system level 3GPP compliant rigorous simulations, in Low and High Traffic demand scenarios with the latter comprising of all video users, against several benchmarks. We analyze the impact of cell load thresholds on ES gains and QoS. AURORA achieved 68% and 99% gain in the total network energy reduction for low and high traffic demand scenarios respectively by putting under-utilized SCs in sleep mode with negligible number of unsatisfied users. It is noteworthy that as long as mobility is predictable with 55% or higher accuracy, AURORA continues to yield Energy Reduction Gain.

II. AURORA Framework

A. Semi-Markov based Spatiotemporal Next Cell Prediction

The mobility prediction model developed in this work builds on our recent study validated in real network [10] that exploits following idea: Next cell can be predicted by modelling user transition from one cell to another as a Markov stochastic process and using HO history to estimate state transition probabilities. Discrete Time Markov Chain (DTMC) has been commonly used in the literature for mobility prediction purposes. The reason being that the Markov based scheme can yield more scalable solution as it does not need to store users' past movements. Instead the crux of this information is captured by transition probabilities. However, DTMC is memory less and assumes transition probability is independent of cell sojourn time. Considering these limitations, DTMC model based works only resort to spatial prediction i.e., identification of future cell only, without any information about the time at which handover may take place. However, human mobility exhibits memory property and can be best approximated with power law (heavy tailed) distribution instead of memory less exponential distributions [10]. Fortunately, Semi-Markov is an advanced class of Markov models that allows for arbitrary distributed sojourn times. Few recent works have characterized prediction accuracy performance of Semi-Markov based model for mobility prediction [10], [11]. However, to the best of our knowledge, this study is the first of its kind that presents spatio-temporal mobility prediction model, and a framework to transform that prediction into future cell load estimates. It then uses those load estimates to devise and analyze a proactive and QoS aware energy saving solution.

We begin by modeling user mobility as a Semi-Markov renewal process $\{(X_n, T_n) : n \geq 0\}$ with discrete state space $\mathcal{C} = 1, 2, 3 \ldots, z$ where T_n is the time of n^{th} transition, X_n is the state at n^{th} transition and total of z cells. Each cell is represented by the state of the Semi-Markov process, and a handover from one cell to another is considered as state transition. It is assumed that the process is time-homogeneous during the time period in which the model is built. The associated time-homogeneous Semi-Markov kernel for user 'u' which is the probability of transition to j^{th} cell if user has already spent time t in i^{th} cell is defined as:

$$\psi_{i,j}^{(u)}(t) = Pr(X_{n+1}^{(u)} = j, T_{n+1}^{(u)} - T_n^{(u)} \leq t | X_n^{(u)} = i) \quad (1)$$

$$= p_{i,j}^{(u)} S_{i,j}^{(u)}(t) \quad (2)$$

where

$$p_{i,j}^{(u)} = \lim_{t \to \infty} \psi_{i,j}^{(u)}(t) \quad (3)$$

$$= Pr(X_{n+1}^{(u)} = j | X_n^{(u)} = i),\ p_{i,j}^{(u)} \in P^{(u)} \quad (4)$$

and

$$S_{i,j}^{(u)}(t) = Pr(T_{n+1}^{(u)} - T_n^{(u)} \leq t | X_{n+1}^{(u)} = j, X_n^{(u)} = i) \quad (5)$$

Here $p_{i,j}^{(u)}$ is the probability of handover of user 'u'from cell i to j, $\mathbf{P}^{(u)}$ is the probability transition matrix of the embedded Markov chain of user 'u'given as

$$\mathbf{P}^{(u)} = \begin{bmatrix} p_{1,1}^{(u)} & p_{1,2}^{(u)} & \cdots & p_{1,z}^{(u)} \\ p_{2,1}^{(u)} & p_{2,2}^{(u)} & \cdots & p_{2,z}^{(u)} \\ \vdots & \vdots & \vdots & \vdots \\ p_{z,1}^{(u)} & p_{z,2}^{(u)} & \cdots & p_{z,z}^{(u)} \end{bmatrix} \quad (6)$$

and $S_{i,j}^{(u)}(t)$ is the sojourn time distribution of user 'u'in cell i when next cell is j. The probability that the user 'u' in cell i will leave cell i before or at time t regardless of the next cell is defined as:

$$\Lambda_i^{(u)}(t) = Pr(T_{n+1}^{(u)} - T_n^{(u)} \leq t | X_n^{(u)} = i) \quad (7)$$

$$= \sum_{j=1}^{z} \psi_{i,j}^{(u)}(t) \quad (8)$$

Now the time-homogeneous Semi-Markov process of user 'u' is defined as $X = (X_t,\ t \in \mathbf{R}_{\mathbf{0}}^{+})$ with state transients as:

$$\phi_{i,j}^{(u)}(t) = Pr(X_t^{(u)} = j | X_0^{(u)} = i) \quad (9)$$

$$= (1 - \Lambda_i^{(u)}(t))\delta_{i,j} + \sum_{m=1}^{z} \int_0^t \phi_{m,j}^{(u)}(t - \tau) d\psi_{i,m}^{(u)}(\tau) \quad (10)$$

$$= (1 - \Lambda_i^{(u)}(t))\delta_{i,j} + \sum_{m=1}^{z} \int_0^t \frac{d\psi_{i,m}^{(u)}(\tau)}{d\tau} \phi_{m,j}^{(u)}(t - \tau) d\tau \quad (11)$$

where $\delta_{i,j}$ is the Kronecker function that is only equal to 1 when $i = j$. Integral equations (10) and (11) are Volterra equations of the first and second kind and the integral is the convolution of $\psi_{i,m}^{(u)}(.)$ and $\phi_{m,j}^{(u)}(.)$ i.e., $\psi_{i,m}^{(u)} * \phi_{m,j}^{(u)}$. It gives the probability that user 'u' starting in cell i will be in cell j by t. The first part of the right-hand side is the probability that the user, being in cell i, never leaves cell i until the end of the period t. The second part of the right-hand side of equation accounts for all cases in which the transition from i to j occurs via another cell $m \neq i$ applying the renewal argument. The evolution equation (10) can be re-written for discrete-time homogeneous semi-Markov process as:

$$\phi_{i,j}^{(u)}(k) = h_{i,j}^{(u)}(k) + \sum_{m=1}^{z} \sum_{\tau=1}^{k} \sigma_{i,m}^{(u)}(\tau) \phi_{m,j}^{(u)}(k-\tau) \tag{12}$$

where $h_{i,j}^{(u)}(k) = (1 - \Lambda_i^{(u)}(t))\delta_{i,j}$ and $\sigma_{i,m}^{(u)}(k) = \frac{d\psi_{i,m}^{(u)}(\tau)}{d\tau}$ can be approximated as follows assuming time step is equal to the unit:

$$\sigma_{i,m}^{(u)}(k) = \begin{cases} \psi_{i,m}^{(u)}(1) & , k = 1 \\ \psi_{i,m}^{(u)}(k) - \psi_{i,m}^{(u)}(k-1) & , k > 1 \end{cases} \tag{13}$$

As $\mathbf{P}^{(u)}$ is right stochastic matrix therefore $\psi^{(u)}(k)$ and $\phi^{(u)}(k)$ will also be a right stochastic matrices i.e., $\sum_{j=1}^{z} \psi_{i,j}^{(u)}(k) = \sum_{j=1}^{z} \phi_{i,j}^{(u)}(k) = 1, \forall i, j \in \mathbb{C}$. The $\phi_{i,j}^{(u)}(k)$ gives the probability that the user 'u' is in cell j after k amount of time from the time instant when he/she made transition from somewhere to cell i. However, to predict the location of a user at every k' time steps, we have to estimate the probability $\hat{\phi}_{i,j}^{(u)}(k', s) = P(X_{s+k'}^{(u)} = j | X_0^{(u)} = i, t_{soj} = s)$ i.e., probability that a user is in cell j after k' time given that the current cell is i and user has stayed in cell i for sojourn time $t_{soj} = s$. It can be evaluated as [11]:

$$\hat{\phi}_{i,j}^{(u)}(k', s) = \frac{P(X_{s+k'}^{(u)} = j, t_{soj} = s | X_0^{(u)} = i)}{P(t_{soj} = s | X_0^{(u)} = i)} \tag{14}$$

$$= \frac{h_{i,j}^{(u)}(s+k') + \sum_{m=1}^{z} \sum_{\tau=s+1}^{s+k'} \sigma_{i,m}^{(u)}(\tau) \phi_{m,j}^{(u)}(s+k'-\tau)}{1 - \Lambda_i^{(u)}(s)} \tag{15}$$

Note that for $s = 0 : \hat{\phi}_{i,j}^{(u)}(k', s) = \phi_{i,j}^{(u)}(k)$. Utilizing the past handover history of user 'u' <time, Cell ID>, Probability transition matrix $\mathbf{P}^{(u)}$ and sojourn time distribution matrix $\mathbf{S}^{(u)}$ are initialized as follows [10]:

$$p_{i,j}^{(u)} = \frac{N_{i,j}^{(u)}}{N_i^{(u)}}, S_{i,j}^{(u)}(k) = \frac{N_{i,j,k}^{(u)}}{N_{i,j}^{(u)}} \tag{16}$$

where $N_{i,j}^{(u)}$ is the number of handovers of user 'u'from cell i to j, $N_{i,j,k}^{(u)}$ is the number of handover of user 'u' from cell i to j with sojourn time less than or equal to k and $N_i^{(u)}$ is the total number of handovers of user 'u' from cell i. Whenever there is a handover from cell i to j, it updates $p_{i,j}^{(u)}$ and $S_{i,j}^{(u)}(k)$ and computes $\psi_{i,j}^{(u)}(k)$. Finally $\phi_{i,j}^{(u)}(k)$ and $\hat{\phi}_{i,j}^{(u)}(k', s)$ are computed. The cell with highest probability is chosen as the predicted future destination i.e., $\max_{j \in \mathcal{N}_i} \hat{\phi}_{i,j}^{(u)}(k', s)$ where $\mathcal{N}_i$ is set of all neighboring cells of cell i. In this way, after every k' time steps, the next HO tuple information for each UE $\{\mathcal{C}_N^u, \mathcal{T}_{HO}^u\}$ is generated wherein $\mathcal{C}_N^u$ is next probable cell of user 'u' at time $\mathcal{T}_{HO}^u$.

B. Future Location Estimation

Let the UE's current location coordinates at time instant k be $l_k^u = (x_k^u, y_k^u)$ and the next cell HO tuple information for each UE be $\{\mathcal{C}_N^u, \mathcal{T}_{HO}^u\}$. Next task is to utilize this information for estimating UE's future location coordinates in next time step $k + k'$. Inspired by observation [12] that nodes in a network usually move around a set of well-visited landmarks with landmark trajectory fairly regular, we utilize past mobility logs of UEs to estimate most probable landmarks visited by each UE in each cell. This information is then utilized to estimate direction of trajectory from current location while distance to be travelled in that direction is estimated using next cell HO time $\mathcal{T}_{HO}$. Let the coordinates of most probable landmark for UE 'u' in next cell $\mathcal{C}_N^u$ be $l_{\mathcal{C}_N^u}^{LM} = (x_{\mathcal{C}_N^u}^{LM}, y_{\mathcal{C}_N^u}^{LM})$ then a unit vector $\hat{u}$ originating from current coordinates in direction of $(x_{\mathcal{C}_N^u}^{LM}, y_{\mathcal{C}_N^u}^{LM})$ is given as:

$$\hat{u} = \frac{[l_{\mathcal{C}_N^u}^{LM} - l_k^u]}{||(l_{\mathcal{C}_N^u}^{LM} - l_k^u)||} \tag{17}$$

where $||.||$ is Euclidian norm operator. The future coordinates at time step $k + k'$ can be estimated as:

$$l_{k+k'}^u = l_k^u + \frac{\sqrt{(x_{\mathcal{C}_N^u}^{LM} - x_k^u)^2 - (y_{\mathcal{C}_N^u}^{LM} - y_k^u)^2}}{T_{HO}^u} k' \hat{u} \tag{18}$$

C. Proactive Energy Saving Optimization

The total instantaneous power consumption of a cell can be given by the sum of circuit and the transmit power as:

$$P_c^{\text{total}} = \lambda^c (P_{CT}^{\text{c}} + \eta_c . P_t^{\text{c}}) \tag{19}$$

where P_{CT}^{c} is the constant circuit power which is drawn if BS in cell c is active and is significantly reduced if the BS goes into sleep mode, P_t^{c} is the transmit power of cell c, η_c denotes the load and λ^c is indicator variable that will be 1(0) for ON(OFF) BS in cell c. One way to quantify Energy Savings is to leverage the performance metric criterion of Energy Consumption Ratio (ECR) that for a cell is defined as the amount of energy consumed in Joules per each bit of information that is reliably transmitted in that cell calculated as:

$$ECR_c = \frac{P_c^{\text{total}}}{\sum_{\mathbb{U}_c} \omega_B^u f(\gamma_u^c)} (Joules/bit) \tag{20}$$

where $f(\gamma_u^c)$ is a function that returns achievable spectral efficiency of user 'u' at a given SINR γ_u^c and ω_B^u is the bandwidth assigned to user 'u'. The SINR γ_u^c at an estimated user location $l_{k+k'}^u$ at time step $k + k'$ when associated with

a cell c is defined as the ratio of reference signal received power $P^c_{r,u}$ by user 'u' from cell c to the sum of reference signal received power by user 'u' from all cells i such that $\forall i \in \mathcal{C}/c$, and the noise variable κ:

$$\gamma^c_u(k+k') = \left[\frac{P^c_t G_u G^c_u \delta \alpha (d^c_u)^{-\beta}}{\kappa + \sum\limits_{\forall i \in \mathcal{C}/c} \eta_i P^i_t G_u G^i_u \delta \alpha (d^i_u)^{-\beta}} \right]_{k+k'} \quad (21)$$

where P^c_t is the transmit power of cell c, G_u is the gain of user equipment, G^c_u is the gain of transmitter antenna of the cell c as seen by the user 'u', δ is the shadowing observed by the signal, α is the path loss constant, d^c_u represents the distance of estimated user location of 'u' i.e., $l^u_{k+k'}$ from cell c, β is the pathloss exponent and η_i denotes cell load in a cell i. This way of weighting the interference power received from each cell with its current resource utilization yields a certain coupling of the total interference with different cell utilizations. More loaded cells contribute more interference power than less loaded ones. The time subscript on right hand side of (21) and in rest of the paper indicates that all terms enclosed within $[.]_{k+k'}$ are considered for the next time step $k+k'$. In the scope of this paper, it is assumed that shadowing estimate information for the estimated user location is available with normally distributed error. In practical network, Channel Maps building on the Minimization of Drive Test (MDT) reports recently standardized by 3GPP and Channel Quality Indicator reports collected can be utilized to estimate channel gains in estimated locations. The total load of cell c at time step $k+k'$ will be the fraction of the total resources in the cell required to achieve required rate of all users of a cell given as:

$$\eta_c(k+k') = \left[\frac{1}{N_c} \sum_{\mathcal{U}_c} \frac{\hat{\tau}_u}{\omega_B \log_2(1+\gamma^c_u)} \right]_{k+k'} \quad (22)$$

where ω_B is the bandwidth of one resource block, N_c is the total number of resource blocks in cell c, $\hat{\tau}_u$ is the minimum required rate of the user and $\mathcal{U}_c$ is the number of active users connected to a cell c. It is a virtual load as it is allowed to exceed one to give us a clear indication of how overloaded a cell is. The required rate in the numerator is the minimum bit rate required by the user depending upon the QoS requirements of the services and user subscription level that can be modelled as function of subscriber behavior, subscription level, service request patterns, as well as the applications being used [9]. The set of users connected to cell c is determined by the user association criterion:

$$\mathcal{U}_j := \left\{ \forall u \in \mathcal{U} \mid j = \arg\max_{\forall c \in \mathcal{C}} (P^c_{r,u_{dBm}} + P^c_{CIO_{dB}}) \right\} \quad (23)$$

where $P^c_{r,u_{dBm}}$ is the true reference signal power in dBm received by user 'u' from cell c and $P^c_{CIO_{dB}}$ is the bias parameter (Cell Individual Offset - CIO). This CIO is primarily used to offset lower transmit power of small cells to transfer more load to them. In case some underutilized cells are turned OFF, remaining cells need to have maximum utilization to cater the transferred load from underutilized cells. It is important to highlight here that in case of ES Optimization with guaranteed minimum QoS requirements, it doesn't make sense to look at throughputs, since the UEs either get exactly the constant bit rate or they are unsatisfied. Hence, more appropriate performance metric to analyze is the number of unsatisfied or dropped users "N_{us}" given as [13]:

$$N_{us}(k+k') = \left[\sum_c \max(0, \sum_{\mathcal{U}_c} \mathbb{1}.(1 - \frac{1}{\eta_c})) \right]_{k+k'} \quad (24)$$

Here η_c by definition from (22) is allowed to exceed 1 to give a clear indication how overloaded a cell is. When $\eta_c = 1$, the inner summation in (24) will be zero meaning all users in cell c are satisfied. When $\eta_c = 2$, the inner summation will be equal to half of the number of users of cell c meaning half of the users are satisfied. The unsatisfied users would not be admitted to enter the system, or they would be dropped if they are already active. Now we formulate the general energy consumption minimization problem for time step $k+k'$ as (25-27):

$$\min_{\boldsymbol{\lambda^c}, \boldsymbol{P^c_{CIO}}} \sum_{\mathcal{C}} [ECR_c]_{k+k'} \quad (25)$$

The objective is to optimize the parameters $\boldsymbol{\lambda^c}, \boldsymbol{P^c_{CIO}}$ of SCs ($\mathcal{SC}$) such that energy consumption ratio in all cells is minimized while ensuring coverage reliability and satisfaction of user throughput requirements. The first two constraints define ranges of the parameters while third constraint is to ensure minimum coverage. Here P^c_{th} is the threshold for the minimum received power for user to be considered covered, $\bar{\omega}$ defines the area coverage probability (a QoS KPI) that operator wants to maintain, and 1(.) denotes indicator function. The fourth constraint ensures each users gets the required minimum bit rate depending upon the QoS requirements of the service and user's subscription level. However, this can only happen when the number of resources available in a cell are sufficient to meet user requirement, therefore, this constraint is complemented with a constraint on cell load $\eta_c < \eta_T$ (Load Threshold) with $\eta_T \in (0, 1]$. The formulated combinatorial optimisation problem in (26-27) contains both continuous $\boldsymbol{P^c_{CIO}}$ and binary $\boldsymbol{\lambda^c}$ decision variables. It can be identified as a mixed integer non-linear programming problem (MINLP). The inherent coupling of ON/OFF state vector, CIOs and cell loads indicate it is a large scale non convex optimization problem. As we are dealing with two problem parameters per cell whose effects on the optimization function are not independent, the complexity is expected to grow exponentially with the number of cells. Hence an exhaustive search for the optimal parameters may not be practical for large size network due to high complexity time search that needs to be done in real time. In order to solve the formulated ES problem, we utilized Genetic Algorithm (GA). The reason being it is considered attractive heuristic technique for a multi-variable mixed integer nonlinear programming problems with a large variable count and enormous search space. Due to its random nature, the genetic algorithm significantly improves

$$\min_{\boldsymbol{\lambda^c}, \boldsymbol{P^c_{CIO}}} \sum_{\mathcal{C}} \left[\frac{\lambda^c (P^c_{CT} + \eta_c . P^c_t)}{\sum_{\mathcal{U}_c} \omega^c_u \log_2 (1 + (\frac{P^c_t G_u G^c_u \delta \alpha (d^c_u)^{-\beta}}{\kappa + \sum_{\forall i \in \mathcal{C}/c} \eta_i P^i_t G_u G^i_u \delta \alpha (d^i_u)^{-\beta}}))} \right]_{k+k'} \tag{26}$$

where

$$\mathcal{U}_j := \left\{ \forall u \in \mathcal{U} \mid j = \arg\max_{\forall c \in \mathcal{C}} (P^c_{r,u_{dBm}} + P^c_{CIO_{dB}}) \right\}$$

$$P^c_{CIO.min} \leq P^c_{CIO} \leq P^c_{CIO.max} \forall c \in \mathcal{SC} \tag{27a}$$

$$\lambda^c \in \{0, 1\} \forall c \in \mathcal{SC} \tag{27b}$$

$$\frac{1}{|\mathcal{C}|} \sum_{\mathcal{C}} \frac{1}{|\mathcal{U}_c|} \sum_{\mathcal{U}_c} 1(P^c_{r,u} \geq P^c_{th}) \geq \bar{\omega} \tag{27c}$$

$$\tau_u \geq \hat{\tau}_u \forall u \in \mathcal{U} \tag{27d}$$

$$\eta_c \leq \eta_T \forall c \in \mathcal{C} \tag{27e}$$

chances of finding a global solution especially for highly non-linear objective functions. Consequently based on estimated network state for time step $k + k'$, AURORA Framework devises optimal ON/OFF state array and CIO values for all the SCs ahead of time such that energy consumption ratio of the whole network is minimized. The ON/OFF state array and CIO values remain fixed from k to k'. As in practical network, SCs need some non-zero time in switching their state, therefore, the proposed strategy gives ample time of k' duration for SCs to switch to optimal ON/OFF state.

III. Performance Analysis

In this section, we present results for our proposed AURORA solution. We have benchmarked its performance against four schemes (i): **Nea**r-Optimal Pe**r**formance Bou**n**d (NARN) wherein it is assumed that AURORA estimates future location and channel estimate at that location with 100% accuracy, (ii): **All** Cell **ON** with **Hom**ogeneous **Net**work Settings (AllOn-HomNet) wherein all cells are ON and no CIO is utilized for small cells, (iii) **All** Cell **On** with **Het**erogeneous **Net**work Settings (AllOn-HetNet) wherein all cells are ON and fixed CIO of 10 dB is utilized for all small cells, (iv) Reactive scheme that is simulated by delaying user location information i.e., Optimization with $\eta_T = 1$ is done based on location information of past one minute.

A. Simulation Settings

We generated typical macro and small cell based network and UE distributions leveraging LTE 3GPP standard compliant network topology simulator in MATLAB. The simulation parameters details are given in Table I. We used wrap around model to simulate interference in an infinitely large network thus avoiding boundary effects. To model realistic networks, UEs were distributed non-uniformly in the coverage area such that a fraction of UEs were clustered around randomly located hotspots in each sector. Monte Carlo style simulation evaluations were used to estimate average performance of the proposed framework. The real challenge here was selection of a mobility trace generation model that realistically represents behavior of actual cellular network users. Based on

TABLE I
Network Scenario Settings

System Parameters	Values
Number of Macro Base Stations	7 with 3 Sectors per Base Station
Small Cells per Sector	5
Number of UEs	Mobile: 84, Stationary: 336
LTE System Parameters	Frequency = 2 GHz, Bandwidth = 10 MHz
Macro Cell Tx Parameters	Tx Power = 46 dBm, Tilt = 10^{2^0}
Small Cell Tx Parameters	Tx Power = 30 dBm, CIO = 0 to 10 dB
Base Station Heights	Macro BS = 25m, Small BS = 10m
Area Coverage Probability	100%
Total Simulation Duration	1 hour

an extensive analysis of pros and cons of recently published models, we chose SLAW (Self-similar Least Action Walk) [14]. Contrary to the conventional random walk models where movement at each instant is completely random, chosen randomly from set of allowed speed and angles, SLAW has been shown to be a highly realistic mobility model. It exhibits all the characteristics of real world human mobility i.e., (i) ***truncated power-law flights and pause-times*** (ii) ***heterogeneously bounded mobility areas*** (iii) ***truncated power-law inter- contact times*** and (iv) ***fractal waypoints***. Therefore, the accuracy of AURORA Framework tested using mobility traces generated by SLAW is very likely to represent its true performance in real network. The SLAW mobility model was utilized to generate HO traces of 84 mobile users for one week. Out of which, traces for first six days were utilized to build and train semi-Markov mobility model for each of the 84 UEs. Moreover, additional 336 stationary UEs (80% of total UEs) were deployed to generate additional loading on the network. For traffic demand, we considered two scenarios (i) ***Low Traffic Demand*** comprising of five different uniformly distributed UE traffic requirement profiles corresponding to 24 kbps (voice), 56 kbps (Text Browsing), 128 kbps (Image Browsing), 512 kbps (FTP) and 1024 kbps (video) desired throughputs, (ii) ***High Traffic Demand*** wherein all UEs are video users. Without loss of generality and keeping operational complexity in mind, the prediction interval k' was set as 1 minute in our simulation study.

The semi-Markov model achieved a maximum prediction accuracy of 87.70% having mean accuracy of 81.46%. This high prediction accuracy is in line with our recent published study [10] on benchmarking prediction accuracy of semi-

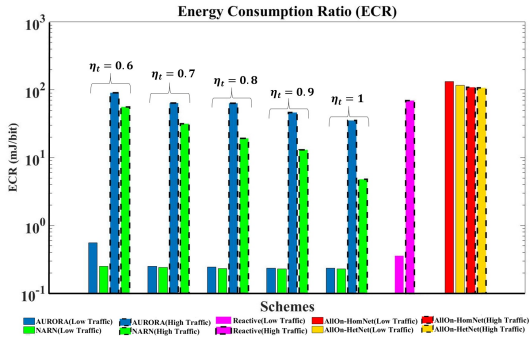

Fig. 1. Energy Consumption Ratio (ECR)

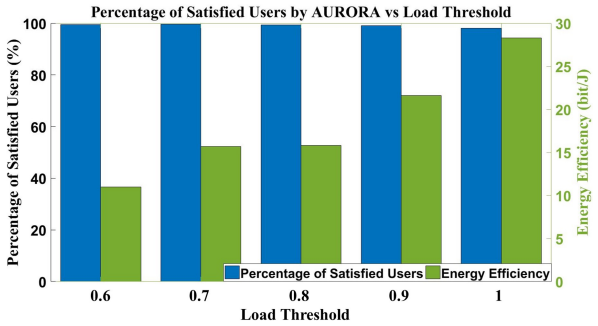

Fig. 2. Percentage of Satisfied Users vs Load Threshold for High Traffic Demand

Markov based mobility prediction model using Real HO measurements collected from live LTE network. A maximum distance error between estimated and actual coordinates was around 33 meters having mean value of around 27.5 meters. One particular reason for high accuracy is that SLAW model is for pedestrian users. Therefore, location of user changes slowly as function of time and thus remains relatively more predictable. With high speed, accuracy is expected to degrade, but then knowledge of street/road layout can be exploited to maintain accuracy. However, this is beyond scope of this paper and will be subject of future study. The Energy Consumption Ratio (ECR) of AURORA and NARN for Low and High Traffic Demands with varying values of Load thresholds η_T along with that of AllOn-HomNet, AllOn-HetNet and state of the art Reactive schemes averaged over 1 hour duration is visualized in Fig. 1. Note that for visualizing ECR ranges for both Traffic Classes in same figure, the y-axis has been plotted in logarithmic scale. The load threshold range is [0.6, 1] since below 0.6 there was no feasible point returned by the P-ES optimization algorithm (27). It is observed that ECR values are higher for high traffic demand scenario as more number of SCs need to be switched ON to cater high load. Moreover AURORA exhibits a linearly decreasing trend with increasing values of η_T. It is significantly much less than the conventional AllOn schemes for all load threshold values. The reason being that for AllOn schemes, all cells are ON at all times that increases energy consumption which is bound to further escalate with densification. At lower η_T values, ECR for AURORA is higher since smaller η_T value compels the AURORA to keep ON larger number of underutilized SCs. For instance at $\eta_T = 0.6$, AURORA switches ON next small cell as soon as the utilization of current ON small cells reach 60%. Thus, on average, large number of SCs will be turned ON for smaller η_T values thus increasing energy consumption. Moreover, with large number of SCs turned ON, there is higher chance that location estimation inaccuracy results in turning ON SCs with very low or no load (i.e., very high ECR - Joules/bit). On the other hand, larger values of η_T enables AURORA to switch OFF large number of SCs. For instance at $\eta_T = 1$, AURORA will switch ON next SC only when the utilization of current ON SCs reaches 100%. As a result ECR is expected to decrease and same trend is observed for NARN. It is interesting to observe that on one hand with increasing value of η_T, less number of SCs are turned ON. Therefore there is less chance of any turned ON SCs with very low or no load. On the other hand, with increasing η_T values, AURORA switches ON smallest possible number of SCs and all of them almost fully utilized with very few resources to spare. As a result inaccuracy in location estimation will result in increased risk of blocking of the UEs (hence increased number of unsatisfied users – see Fig. 2) thereby negatively affecting QoS. However, as the number of fully utilized SCs is a more dominant factor in determining overall ECR as compared to slight increase in the number of unsatisfied users, therefore, overall ECR reduces. The comparison of AURORA with Reactive scheme shows that ECR for Reactive scheme is higher as compared to AURORA. This is because in Reactive scheme, due to delayed user location information, outdated configuration settings that are suboptimal for current instant, are applied to the network. This increases the percentage of unsatisfied users (on average 1.85% with AURORA at $\eta_T = 1$ while 4% with Reactive scheme at high traffic load) and hence higher ECR. Moreover, ECR for AllOn-HomNet is slightly higher as compared to AllOn-HetNet. This is because higher CIO values used in AllOn-HetNet compels SCs to be more utilized and hence reduced ECR as compared to AllOn-HomNet scheme.

The average percentage of satisfied users under AURORA framework vs Load Threshold η_T for high traffic demand scenario is visualized in Fig. 2 on left y-axis while Energy Efficiency $(1/ECR)$ is plotted on right y-axis. It can be observed that at low η_T values, plenty of free resources are available in relatively more number of ON BSs. Hence more users are served with enough resources to meet their minimum QoS requirements. Even with location prediction inaccuracies, the UEs will still have better chance to get enough resources and be satisfied. However, more SCs are turned ON at low η_T with more chance of being underutilized and hence lower Energy Efficiency. As η_T value becomes higher and approaches 1, AURORA returns such an OPC $\boldsymbol{\lambda^c}, \boldsymbol{P^c_{CIO}}$ that results in smallest possible number of switched ON SCs and all of them almost fully utilized with very few resources to spare. Hence a slight location estimation inaccuracy can result in increased risk of blocking and hence decrease in number of satisfied users. Contrary to that, fewer cells turned ON with more utilization improve energy efficiency of the network. It is interesting to observe that for high traffic demand scenario even at $\eta_T = 1$, percentage of satisfied users is above 98%.

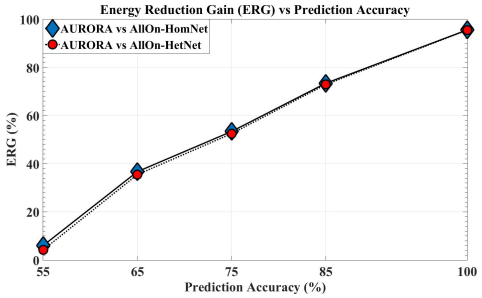

Fig. 3. Energy Reduction Gain vs Prediction Accuracy

It is logical to anticipate that the energy saving gain of AURORA i.e., Energy Reduction Gain (ERG) performance metric given as:

$$ERG = (\frac{ECR_{Benchmark} - ECR_{AURORA}}{ECR_{Benchmark}}) \quad (28)$$

will depend on the accuracy of the underlying mobility prediction model. We analyzed this dependence by generating four set of mobility traces with increasing randomness. As a result, our prediction model trained on these four set of traces exhibited average prediction accuracy of 85%, 75% , 65% and 55% respectively. The average ERG of AURORA for these varying values of Prediction Accuracy against AllOn-HomNet and AllOn-HetNet schemes, averaged over 1 hour duration for high traffic demand scenario, is plotted in Fig. 3. It is observed that as expected the gain of AURORA decreases with decrease in prediction accuracy. However, it is noteworthy that as long as mobility is predictable with 55% or higher accuracy, AURORA continues to yield Energy Reduction Gain. It has been observed, that for accuracy below 55% the AURORA′s gain diminishes and turns into loss. Given that typical human mobility features 93% predictability when averaged over a large real user sample space [15], AURORA is a promising novel proactive ES solution.

IV. Conclusions

This paper has proposed a novel spatiotemporal mobility prediction aware proactive sleep-mode based energy saving optimization algorithm for cracking the future 5G ultra-dense HetNets puzzle. The proposed AURORA framework employs innovative concept of estimating future user locations and leverage that to estimate future cell loads. It then devises energy saving optimization problem for the estimated future network scenario. The majority of the conventional reactive style approaches are expected to solve the formulated energy saving problem dynamically in real-time as network conditions change. However this is close to impossible even when substantial computing power is available. Contrary to that, the innovative proposed approach enables state-of-the-art heuristic techniques like GA to find practically good solutions to the formulated optimization problem predictively ahead of time. This proactiveness make AURORA a key enabler for meeting 5G ambitious latency and QoS requirements. Moreover, AURORA framework also takes into account the interplay among the three intertwined SON functions (ES, CCO and LB) due to the overlap among their primary optimization parameters thereby addressing a key challenge of SON conflicts that traditional ES solutions face. Extensive simulations employing realistic SLAW mobility model indicate that, in best case, AURORA can achieve energy reduction gain of about 68% for high traffic demand scenario in ultra-dense HetNets as compared to Always On approach. Comparative performance analysis with near-optimal performance bound indicate satisfactory robustness of the proposed AURORA solution to location prediction inaccuracy.

Acknowledgement

The core idea of AURORA has won 2017 IEEE International competition for best solution for GREEN ICT [16]. This material is based upon work supported by NSF under Grant Numbers 1619346, 1559483, 1718956 and 1730650.

References

[1] I. Ashraf, F. Boccardi, and L. Ho, "SLEEP mode techniques for small cell deployments," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 72–79, aug 2011.

[2] S. Buzzi, C.-L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, "A Survey of Energy-Efficient Techniques for 5G Networks and Challenges Ahead," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 697–709, apr 2016.

[3] M. Ajmone Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal Energy Savings in Cellular Access Networks," in *2009 IEEE International Conference on Communications Workshops*. IEEE, jun 2009, pp. 1–5.

[4] R. Litjens and L. Jorguseski, "Potential of energy-oriented network optimisation: Switching off over-capacity in off-peak hours," in *21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, sep 2010, pp. 1660–1664.

[5] Z. Niu, "TANGO: traffic-aware network planning and green operation," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 25–29, oct 2011.

[6] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions (3GPP TR 36.902 version 9.2.0 Release 9)," Tech. Rep., 2010.

[7] F. Z. Kaddour, E. Vivier, L. Mroueh, M. Pischella, and P. Martins, "Green Opportunistic and Efficient Resource Block Allocation Algorithm for LTE Uplink Networks," pp. 4537–4550, 2015.

[8] H. Y. Lateef, A. Imran, and A. Abu-dayya, "A framework for classification of Self-Organising network conflicts and coordination algorithms," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, sep 2013, pp. 2898–2903.

[9] A. Imran and A. Zoha, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Network*, vol. 28, no. 6, pp. 27–33, nov 2014.

[10] H. Farooq and A. Imran, "Spatiotemporal Mobility Prediction in Proactive Self-Organizing Cellular Networks," *IEEE Communications Letters*, vol. 21, no. 2, pp. 370–373, feb 2017.

[11] J.-K. Lee and J. C. Hou, "Modeling Steady-state and Transient Behaviors of User Mobility: Formulation, Analysis, and Application," in *Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. MobiHoc '06. New York, NY, USA: ACM, 2006, pp. 85–96.

[12] J. Ghosh, S. J. Philip, and C. Qiao, "Sociological Orbit Aware Location Approximation and Routing (Solar) in DTN," State Univ. of New York at Buffalo, Tech. Rep., 2005.

[13] I. Viering, M. Dottling, and A. Lobinger, "A Mathematical Perspective of Self-Optimizing Wireless Networks," in *2009 IEEE International Conference on Communications*. IEEE, jun 2009, pp. 1–6.

[14] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "SLAW: A New Mobility Model for Human Walks," in *IEEE INFOCOM 2009 - The 28th Conference on Computer Communications*. IEEE, apr 2009, pp. 855–863.

[15] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[16] [Online]. Available: http://greenict.ieee.org/summit/gtict-summit-2017/green-ict-competition-young-professionals