A Base Station Sleeping Strategy in Heterogeneous Cellular Networks Based on User Traffic Prediction

Xinyu Wang, Bingchen Lyu, Chao Guo, Jiahe Xu, and Moshe Zukerman, *Life Fellow, IEEE*

***Abstract*—Real-time traffic in a cellular network varies over time and often shows tidal patterns, such as the day/night traffic pattern. With this characteristic, we can reduce the energy consumption of a cellular network by consolidating workloads spreading over the entire network to fewer Base Stations (BSs). In this work, we propose a BS sleeping strategy for a two-tier Heterogeneous Cellular Network (HeCN) that consists of Macro Base Stations (MaBS) and Micro Base Stations (MiBS). We first use a Bidirectional Long Short-Term Memory (BLSTM) neural network to predict the future traffic of each user. Based on the predicted traffic, our proposed BS sleeping strategy switches user connections from underutilized MiBSs to other BSs, then switches off the idle MiBSs. The MaBSs are never switched off. All user connections have predefined Signal-to-Interference-plus-Noise Ratio thresholds, and we ensure that each user's service quality, which is related to the user's traffic demand rate, is not degraded when switching user connections. We demonstrate the effectiveness and superiority of our proposed strategy over four other baselines through extensive numerical simulations, where our proposed strategy substantially outperforms the four baselines in different scenarios.**

***Index Terms*—Heterogeneous cellular networks, traffic forecasting, energy saving, SINR, quality of service, BLSTM.**

I. Introduction

MOBILE internet services, such as mobile payments, online maps, and real-time videos, have become integral to people's daily lives. With the widespread use of smartphones, tablets, wearable devices, and Internet of Things (IoT) applications, the number of devices connected to the network has significantly increased. In early 2022, the number of IoT devices reached 14.4 billion, an increase of 18% compared to 2021 [1]. This growth in connected devices has led to a substantial increase in data traffic. By the end of 2021, the global mobile data traffic had reached around 84 EB per month, and it is estimated to increase by about 4.2 times, reaching 368 EB per month in 2027 [2].

To keep up with this explosive growth in data traffic, the construction and deployment of fourth-generation and fifth-generation wireless networks have been promoted and popularized, increasing the number of Base Stations (BSs) and the capacity and coverage of mobile cellular networks. However, as Information and Communications Technologies (ICTs) account for 1.8%-2.8% of global greenhouse gas emissions [3], and BSs consume more than 80% of the energy in mobile cellular networks [4], energy efficiency (EE) has become a crucial concern. Optimal BS switching on/off and BS sleeping strategies [5, 6] can significantly improve the EE of mobile cellular networks, reducing greenhouse gas emissions and leading to considerable cost savings. Therefore, such strategies and solutions have become critical to the design of future mobile cellular networks.

However, the unprecedented number of mobile device connections in the network and the diversification of data service types have rendered traditional Homogeneous Cellular Networks (HoCNs) inadequate to meet the complex requirements of future networks. HoCNs consist of BSs with similar working modes and transmit patterns [7]. To address this issue, Heterogeneous Cellular Networks (HeCNs) have been introduced, which can achieve flexible and low-cost deployment by adding Micro Base Stations (MiBSs), such as pico and femto BSs, with different working modes and transmit patterns within the coverage of Macro Base Stations (MaBSs). This approach increases data transmission rates and complements the coverage area, as shown in Fig. 1. The MiBSs have different carrier frequencies than MaBSs and do not cause interference to MaBSs in the same coverage area.

In HeCNs, MiBSs can be shut down or switched to sleep mode at appropriate times, providing a significant opportunity to reduce the excessive energy consumption of BSs. This attribute makes HeCNs more amenable to BS sleeping than the earlier, more rigid HoCNs because, in a HeCN, a user of a shutdown MiBS can connect to a MaBS that covers their location. In contrast, in a HoCN, if a BS shuts down, its users may not obtain the required Quality of Service (QoS) by connecting to active, far-away BSs. The flexibility of HeCNs enables efficient network management and better utilization of network resources, making them a promising solution for future mobile networks.

Researchers have focused on analyzing and predicting mobile cellular network data traffic to find the best BS sleeping schedule. Studies such as [8, 9] have found that traffic in mobile cellular networks exhibits strong self-similarity and follows general regularity in spatio-temporal distribution. This predictability of mobile cellular network traffic has made it a topic of exploration for researchers. Additionally, there is a significant difference in the traffic pattern between day and

The work described in this paper was supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA) and by City University of Hong Kong under Projects 7020007 and 9610544. *(Corresponding author: Chao Guo.)*

Xinyu Wang, Jiahe Xu, and Moshe Zukerman are with the Department of Electrical Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China. (e-mail: xywang47-c@my.cityu.edu.hk; jiahexu3-c@my.cityu.edu.hk, m.zu@cityu.edu.hk).

Chao Guo is with the Centre for Intelligent Multidimensional Data Analysis Limited, Hong Kong SAR, China. He was with the City University of Hong Kong, Kowloon, Hong Kong SAR, China. (e-mail:chaoguo6-c@my.cityu.edu.hk).

Bingchen Lyu was with the City University of Hong Kong, Hong Kong. He is now with Shanghai Kuanyu Digital Technology Co., Ltd., Shanghai, 200090, China. (e-mail: bingchenlv@gmail.com).

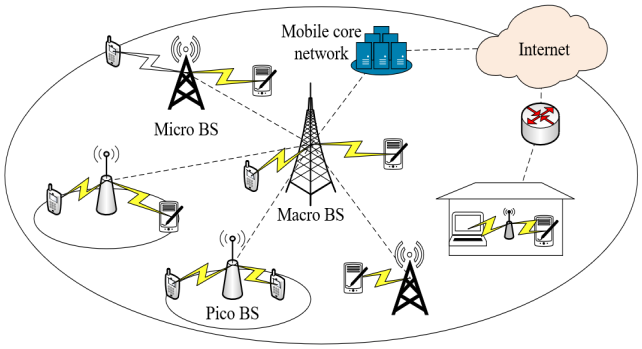


Fig. 1: Architecture of a HeCN.

night for every working day [8, 10]. Therefore, analyzing the historical traffic data for a specific period in the same area can effectively predict the network traffic demand in the area for a certain period in the future based on the spatio-temporal distribution of users' mobile data traffic.

Prediction methodologies for the traffic of mobile cellular networks that enable BS sleeping scheduling in different periods have been studied under specific scenarios. In [11–13], the authors proposed several BS sleeping strategies to reduce energy consumption based on predicted traffic, but they were all for HoCNs. Lin *et al.* [14] predicted data traffic in HeCNs and proposed BS sleeping strategies based on the predicted traffic without considering the differences between the traffic demands of different users in the network and the effect of the locations of the different BSs on the performance of sleeping strategies. Donevski *et al.* [15] proposed methods to find the optimal BS sleeping times in HeCNs based on traffic prediction using neural networks but did not specify user reallocation. Specifically, for the association between users and BSs, since the transmission power of MaBSs in HeCNs is much greater than that of MiBSs, users will prefer to choose MaBSs that can provide a better QoS without particular intervention, causing MiBSs not to be fully utilized or even empty-loaded. Therefore, a BS sleeping strategy must consider the effect of different BS deployment methods on the performance of sleeping strategies and control user choices to achieve sufficient utilization of MiBSs in HeCNs during the sleeping periods. These essential enhancements to the BS sleeping strategy will lead to load balancing among BSs and energy saving.

In this paper, our main contributions to developing BS sleeping strategies in HeCNs are as follows.

- We use a Bidirectional Long Short-Term Memory (BLSTM) to predict the traffic demand rates of each user in the HeCN rather than predicting the traffic of the whole network or individual BSs as done in prior literature. We use real historical traffic data sets for training and testing the prediction model. The predicted traffic is then used as the input for our designed sleeping strategy.
- Based on stochastic geometry, we abstract the deployment of BSs in HeCNs to a random point process. Specifically, we consider a two-tier HeCN consisting of MaBSs and MiBSs and use a Poisson Point Process (PPP) and a Matérn Hard-Core Point Process (MHCPP) to model the positions of BSs, respectively. We then evaluate the energy-saving performance of the two-tier HeCN under different BS deployment methods, namely PPP and MHCPP.
- Given the predicted user traffic demand and the deployment results of BSs, we design a BS sleeping strategy, which we refer to as *Minimum Load Sleep First* (MiLSF) strategy, for the two-tier HeCN. In this strategy, we try to reallocate users of MiBSs in the network to other BSs and turn the idle MiBSs into sleep mode to minimize the overall energy consumption of the whole network. We consider each user connection to have a predefined Signal-to-Interference-plus-Noise Ratio (SINR). We guarantee each user's QoS requirement, which is related to the user's demand traffic rate, by ensuring the SINR does not violate the given threshold when switching the connection from one BS to another. Our goal is to optimize the selection of the set of MiBSs that are switched to sleep mode at the beginning of a low-load period and then are switched back to active mode once the low-load period ends. In practice, the low-load period is normally at night. This design avoids frequent BS switching changes, which are problematic because of initialization requirements during switching from sleep to active mode [16]. Notice that MaBSs in the network will not be switched to sleep mode, and our MiLSF strategy is implemented to not degrade all users' QoSs after reallocating the users. We demonstrate the effectiveness and superiority of MiLSF over four other baseline strategies through extensive numerical simulations.

The MiLSF strategy effectively reduces energy consumption and operating expenses for the HeCN. This, in turn, helps to reduce carbon emissions, making it an environmentally friendly solution. When combined with AI techniques to predict future traffic, the MiLSF strategy can help ensure QoS by proactively avoiding violations of the SINR threshold provided to each user. This proactive approach helps to prevent potential client loss and revenue loss for mobile carriers, thus making it a highly desirable solution for improving network performance. The remainder of this paper is organized as follows. In Section II, we discuss other related work for developing green HeCNs, traffic prediction, and BS sleeping strategies in HeCNs. In Section III, a description of the system models is provided, which includes the BS development model, SINR model, BS power consumption model, and BLSTM traffic prediction model. In Section IV, the energy-saving optimization problem in the HeCN is formulated. In Section V, we propose the MiLSF in the HeCN based on traffic prediction. Section VI introduces the parameter settings and the simulation scenarios and presents extensive numerical results that demonstrate the superiority of our proposed MiLSF strategy over four other baseline strategies. Finally, in Section VII, we draw our conclusions.

II. Related Work

A. Green HeCNs

The topic of developing green HeCNs has been widely studied over the past decades. Wildemeersch *et al.* [17] demonstrated that deploying MiBSs in MaBSs can improve user satisfaction, network capacity, network coverage, and reliability of HeCNs. Ghosh *et al.* [18] discussed new problems encountered in HeCNs compared to HoCNs, including multi-layer network modeling, SINR analysis, interference management between different types of BSs, and network maintenance. Elsawy *et al.* [19] introduced a new interference coordination management mechanism in HeCNs to solve the problems caused by BS power differences, distribution randomness of BSs, and frequent switching between users and BSs. Guo *et al.* [20] presented a theoretical framework for optimizing BS deployment in order to maximize network frequency efficiency and mitigate interference caused by a high density of MiBSs. In [21], Cheng *et al.* used the system throughput as the performance index and found the best placement positions of various MiBSs. Yang *et al.* [22] studied the problem of energy-saving BS deployment under probabilistic line-of-sight and non-line-of-sight transmissions. Mlika *et al.* [23] focused on the issue of BS operation and user association to develop energy-efficient HeCNs to meet the minimum rate requirements. In [24], Ak *et al.* used a PPP model to decide the locations of BSs and studied the impact of different network parameters on energy consumption. Chen *et al.* [25] analyzed the effects of various BS deployment topologies on network throughput and reliability. In [26–28], the BS deployment strategies with maximum reliability for HeCNs in real-world metropolitan scenarios were studied. However, the above studies did not consider the movement of users and the variability of the traffic load generated by these users over time.

B. Traffic prediction

Traffic prediction is usually realized by analyzing the network traffic pattern and extracting traffic changes' characteristics. Since cellular network traffic is usually presented as time series, time series modeling methods can describe and predict changes in cellular network traffic [29]. Xu *et al.* [30] decomposed the traffic of BSs into periodic and random components and used time-series predictions to predict the BS traffic. In [31], Sultan *et al.* proposed a method to remove abnormal data to predict traffic better. Zhang *et al.* [32] analyzed the mobile traffic data by three aspects: periodic flow components, directional flow components, and additional elements that contain the effects of noise and anomalous events. Xu *et al.* [33] proposed a decomposition model consisting of traffic trend, seasonal, and holiday components and built a sub-model of the holiday component based on out-of-direction residuals and seasonal components.

Moreover, the emergence of many neural network methods provides new ideas for modeling and predicting cellular network traffic. Guo *et al.* [34] used a deep neural network model to learn the characteristics of historical data for prediction. LSTM is a particular Recurrent Neural Network (RNN) that successfully addresses the gradient disappearance problem [35] and easily learns long-term dependent information. Specifically, LSTM feeds the previous step's output to the current step's input layer, a dynamic feedback connection that models dependencies in a time series. In [36], Trinh *et al.* studied the effectiveness of RNNs for traffic forecasting and used LSTM to successfully predict the traffic for distant timeslots, showing the applicability and accuracy of LSTM in mobile cellular network traffic prediction. Azari *et al.* [37] compared the user traffic prediction performance of the statistical learning-based forecasting method named autoregressive integrated moving average (ARIMA) and LSTM in various scenarios and then demonstrated the superiority of LSTM over ARIMA. However, existing publications focus more on the traffic prediction of BSs or the entire network without considering individual users' traffic demands.

C. BS sleeping techniques

In HeCNs, considering the high flexibility of MiBSs and the ease of switching on and off, sleep strategies are usually applied to MiBSs to reduce the overall energy consumption of the networks. Wu *et al.* [38] improved the EE of a HeCN through cooperation among MiBSs. In [39], the influence of the sleeping ratio of femto BSs on EE was studied to obtain the optimal sleeping rate of femto BSs with maximum EE. In [40], Li *et al.* comprehensively studied the EE-related issues in femtocell networks and proposed a fixed-time sleeping strategy for femto BSs to save energy consumption in HeCNs.

BS sleeping strategies must ensure that the active BSs can bear the network's traffic load when reducing energy consumption. The dynamic change of the network traffic load is an essential reference basis for formulating the sleeping strategies of the BSs. In [41], Wyner proposed using sleep mode for BSs during the low traffic load to reduce system energy consumption. In [42], Saker *et al.* associated the user with the MaBSs and woke up the sleeping MiBSs only when the traffic load was huge to make up for service defects. Considering the load of BSs, Wu *et al.* [43] proposed a BS sleeping strategy based on E/M/1 and E/M/n queuing models. In [44], Wu *et al.* modeled every sleeping BS as an M/G/1/K queue and studied the tradeoff between the grade of service and power consumption under various BS sleeping strategies. The neural networks were used to predict the future traffic load of BSs in [11, 12, 14], and then the corresponding BS sleeping strategies were formulated according to the predicted traffic load. In [45], Wu *et al.* provided a detailed survey presenting facts and figures about BS sleeping strategies and the potential issues and future research in green cellular networks.

The sleeping of BSs needs to ensure the QoS of users and user associations with the remaining active BSs. In [46], some MaBSs were switched to sleep mode, and the users associated with the sleeping MaBSs were reallocated to surrounding MiBSs without reducing the QoS of users. In [47], the energy consumption was reduced by sleeping some MiBSs that are not associated with users. In [48], Soh *et al.* reduced energy consumption while guaranteeing all network requests by introducing transmission power adaptation and BS sleeping

technique. Liu *et al.* [16] considered the extra energy consumption and initialization delay caused by frequent switches of BSs and performed BS sleeping without reducing the QoS of users. Zhang *et al.* [49] proposed a BS sleeping strategy for 5G small-cell networks that also included energy harvesting and matched the dynamics of arriving energy with all the users' demands. Alqasir *et al.* [50] considered a HeCN where the power source of MiBSs comprises harvested energy and a grid power source. Then, Alqasir minimized the energy consumption of the HeCN by optimizing the transmitted power and activation/deactivation of the MiBSs. In [51], a user association and scheduling algorithm was proposed to help reduce energy consumption, where users connected to sleeping BSs can be rapidly reconnected to active BSs according to multiple parameter access criteria.

In summary, research on BS sleeping strategies mainly considers the traffic load of BSs, resource allocation, and user association. Researchers use greedy algorithms, heuristic algorithms, or game theory to obtain energy-saving sleeping strategies to maintain user satisfaction. To the best of our knowledge, most of the current BS sleeping strategies are based on traffic prediction of BSs rather than individual users with different traffic demands in the network. The traffic prediction for different users can help formulate BS sleeping strategies that accurately meet the needs of different users. Meanwhile, the impact of different deployment methods of MaBSs and MiBSs in heterogeneous networks on user association and energy consumption has not been studied in depth. In this paper, based on the BLSTM predicted traffic data of every user with different traffic patterns in the network, our proposed MiLSF selectively sleeps some MiBSs and considers the user association problem, thereby reducing the energy consumption in the network without reducing the QoS of every user.

III. System Model

The introduction of different types of BSs in HeCNs has led to the development of multi-layer network structures, which comprise multiple types of BSs. This paper considers a two-tier HeCN consisting of MaBSs and MiBSs. In the following, we describe the details of the two-tier HeCN and the BLSTM-based model for user traffic prediction.

A. BS deployment

Compared to the traditional cellular topologies where MaBSs use a hexagonal grid model and the MiBSs are deployed within the MaBSs' coverage in a specific way, the stochastic geometry theory provides an effective and tractable way to study the HeCNs' performance from a statistical perspective [19, 52]. Specifically, PPP is a tractable and popular point process due to its independence [53]. The PPP can be used to model the deployment process of the BSs of cellular networks, while MHCPP [54], which is based on PPP but avoids points being too close, can better reflect the BSs' deployment in actuality and describe the network more practically. This paper considers both a PPP and an MHCPP in a two-dimensional Euclidean plane to deploy the BSs.

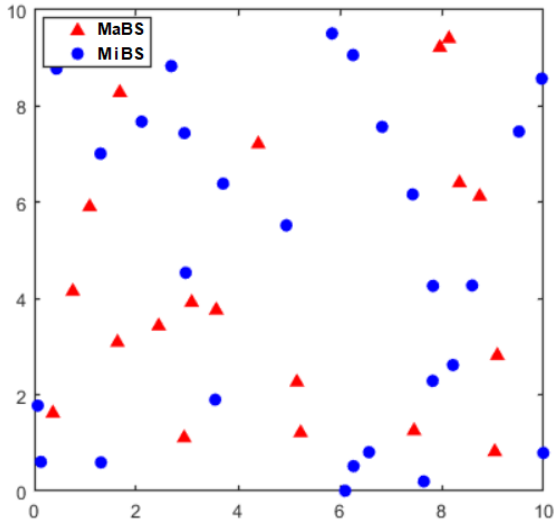
- PPP. The deployment process of BSs $\Phi = \{B_i; i = 1, 2, ...\}$ in a two-dimensional Euclidean plane $\mathbb{R}^2$ can be modeled as a PPP if the number of the BSs in any two-dimensional compact set $\mathcal{D} \subset \mathbb{R}^2$ is a Poisson random variable, where B_i is the i-th BS. Specifically, we deploy the MaBSs and MiBSs by applying different intensity values to a PPP. Notice that the deployments of MaBSs and MiBSs are independent of each other.
- MHCPP. The MHCPP is based on the PPP but avoids interference between very close BSs by removing the points that coexist within a predefined non-negative distance (also known as the hard-core parameter r_h), namely $\| B_i - B_j \| \geq r_h, \forall B_i, B_j \in \Phi, i \neq j, r_h \geq 0$. Specifically, for any two BSs B_i and B_j separated by less than r_h, we remove the BS with a smaller subscript, which is B_i if $i < j$.

The realization of a PPP and an MHCPP for the BSs' deployment is illustrated in Fig. 2. The deployment processes for MaBSs and MiBSs are independent of each other. We used different intensities for MaBS and MiBS because of MaBS's higher transmitting power and larger coverage area, which resulted in fewer than MiBS in the actual deployment. We can observe from Fig. 2(a) that BSs of the same type may exist within a very close range, resulting in considerable signal interference. This situation is improved in Fig. 2(b) by removing BSs of the same type that are too close together. It should be noted that, for ease of exploration, we only consider the signal interference among the same type BSs. This is because, as mentioned above, MaBS and MiBS are characterized by entirely different carrier frequencies, so the interference among different types of BSs can be ignored compared with the interference among the same type of BSs.

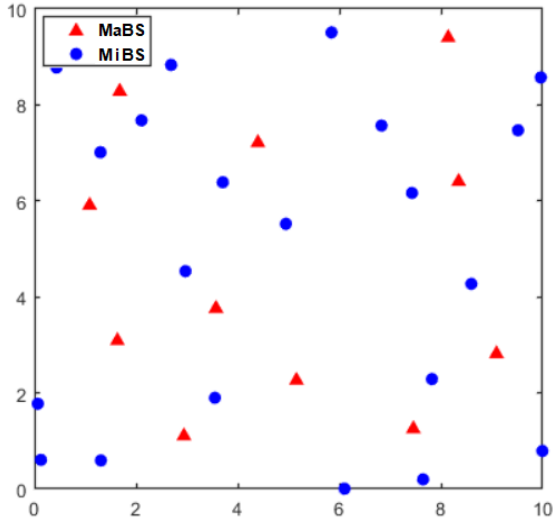
It should be noted that the hard-core constraint in the MHCPP model may potentially adversely affect its scalability when applied to large or high-density point patterns. As the number of BSs increases, finding feasible configurations that satisfy the hard-core constraint becomes more challenging, potentially limiting the model's scalability in practical scenarios of highly dense networks in 6G architectures. Consequently, it would be beneficial to conduct a further study that includes an asymptotic analysis of the scalability of MHCPP in relation to the number of BSs. However, such an analysis falls beyond the scope of this paper and should be considered as a valuable direction for future research.

Nevertheless, it is worth mentioning that for the actual scenarios of 5G or future 6G high-density BS networks, the increase in carrier frequency and the shortening of the wavelength of the BSs hasten the attenuation during signal propagation. Compared to 4G BSs, the coverage of 5G and 6G BSs is significantly reduced [55]. These characteristics suggest that while increasing the BS density in the MHCPP model, we could employ a smaller hard-core constraint distance, thereby lending the MHCPP model scalability in the high-density network practical scenarios of 5G or even 6G architecture.

Another limitation of the MHCPP model is that it assumes a specific form of spatial dependence characterized by a correlation structure [54, 56], which may not fully capture the complex spatial dependencies present in real-



(a) PPP in a 10km×10km area with MaBS intensity 0.2 points/km^2 and MiBS intensity 0.4 points/km^2, respectively.



(b) MHCPP based on the PPP in (a) with the hard-core parameter r_h set as 1km and 2km for MaBS and MiBS, respectively.

Fig. 2: BS deployment realized by PPP and MHCPP.

world scenarios. To address this limitation, alternative models, such as Voronoi Tessellation [57] and network-based models, offer more flexibility in capturing complex spatial dependence patterns. Voronoi tessellation divides space into regions around each base station, ensuring that any point within each region is closer to its corresponding base station than any other [57]. This model effectively captures the spatial interaction between base stations and provides a more accurate representation of coverage areas and interference patterns. It takes into account the proximity of users to individual base stations and enables the analysis of signal strength and interference levels. Network-based models explicitly consider the underlying network structure and connectivity between base stations. Graph-based models, such as random geometric graphs or random connection models, capture the spatial proximity of base stations and their connectivity patterns. These models are particularly useful for modeling wireless networks with interference, signal propagation, and connectivity constraints.

B. SINR model

In this paper, we assume all homogeneous BSs have the same characteristics. The MaBSs and MiBSs are characterized by different deployment intensities, transmit powers, number of contained transmit antennas, carrier frequencies, and bandwidths.

We consider K users, M MaBSs, and N MiBSs in the two-tier HeCN, where

- the transmit powers of a single transmit antenna in MaBSs and MiBSs are represented by p_1 and p_2;
- the carrier frequencies of MaBSs and MiBSs are characterized by f_1^c and f_2^c;
- the bandwidths of MaBSs and MiBSs are denoted as w_1 and w_2.

For clarity of presentation, we define sets $\Psi_M = \{1, 2, ..., M\}$ and $\Psi_S = \{M+1, M+2, ..., M+N\}$, and the MaBSs and MiBSs in the two-tier HeCN are tagged as $B_i, i \in \Psi_M$ and $B_i, i \in \Psi_S$, respectively. Let $\Psi_K = \{1, 2, ..., K\}$ be the set of users in the two-tier HeCN. Since MiBSs and MaBSs have different characters (e.g., they use different carrier frequencies and different bandwidths), we introduce a binary indicator $\theta(x)$ to clarify it, which indicates which tier a BS belongs to, defined as:

$$\theta(x) = \begin{cases} 1, & \text{if } x \in \Psi_M, \\ 2, & \text{if } x \in \Psi_S. \end{cases} \tag{1}$$

Consider user $k \in \Psi_K$ who is allocated to BS $i \in \Psi_M \cup \Psi_S$, its path loss (dB) is obtained by

$$L_{i,k} = 20\log\left(\frac{4\pi f^c_{\theta(i)}}{c}\right) + 10\beta\log\left(d_{i,k}\right), \quad \forall\ k \in \Psi_K,\ i \in \Psi_M \cup \Psi_S, \tag{2}$$

where c is the speed of light, β is the pass loss exponent, $f^c_{\theta(i)}$ is the carrier frequency of BS i, and $d_{i,k}$ is the distance between BS i and user k. Then the received SINR from BS i to user k is

$$S_{i,k} = \frac{p_{\theta(i)}\rho_{i,k}L^{-1}_{i,k}}{\sum_{j\in\Psi_M\cup\Psi_S\setminus\{i\}} p_{\theta(j)}\rho_{j,k}L^{-1}_{j,k} + \eta_0 w_{\theta(i)}}, \quad \forall\ k \in \Psi_K,\ i \in \Psi_M \cup \Psi_S, \tag{3}$$

where $p_{\theta(i)}$ and $w_{\theta(i)}$ are the transmit power and bandwidth of BS i, respectively. η_0 denotes the noise spectral density. $\rho_{i,k}$ is the small-scale fading between user k and BS i where we consider Rayleigh fading that follows an exponential distribution [58]. In this context, user k will be blocked from BS i if the received SINR is lower than threshold γ_0. The numerator in (3) (i.e., $p_{\theta(i)} \cdot \rho_{i,k} \cdot L^{-1}_{i,k}$) is the signal power received by user k from its associated BS i. The denominator consists of both interference power and white noise power. The interference power (i.e., $\sum_{j\in[M+N]\setminus\{i\}} p_{\theta(j)} \cdot \rho_{j,k} \cdot L^{-1}_{j,k}$) is the signal power received by user k from other BSs, i.e., the BSs except BS i.

Different interference mitigation strategies have been used in a HeCN in order to improve the SINR performance, such as the beamforming technique [59]. Instead of a single antenna, an antenna array is used for signal transmitting or receiving in beamforming. The signals received by multiple antennas have different phases (and amplitudes) due to the different signal paths. The received signals are adjusted to be coherent by introducing appropriate delays such that different signals can enhance each other, achieving a channel gain much more than a single antenna can realize. In addition, other techniques, such as power control and resource allocation, can also be used to mitigate interferences and improve the quality of signals. However, our core contribution to this paper is to present a BS sleeping strategy to save energy with guaranteed quality services. We consider a simple and general SINR model but do not delve into detailed interference management techniques. These mitigation techniques can be combined with our work by providing a different SINR threshold to our proposed algorithm.

C. BS power consumption model

For a user k connected to BS i with a SINR no less than the threshold γ_0, namely $S_{i,k} \geq \gamma_0$, according to Shannon–Hartley theorem [60], the required bandwidth of BS i by user k at moment t is

$$b_{i,k}(t) = \frac{r_k(t)}{\log_2(1 + S_{i,k})}, \tag{4}$$

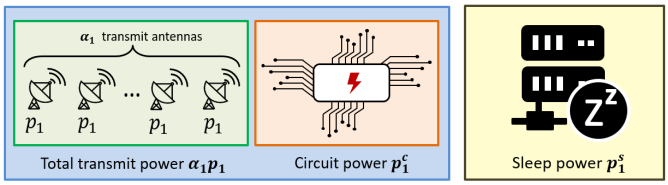
where $r_k(t)$ is the demand rate of user k at moment t. (4) describes the relationship between the required bandwidth resources and the demand traffic rate of users. Notice that the $S_{i,k}$ in (4) is expressed as a linear power ratio, not as logarithmic decibels. Let $\Psi_K^i(t)$ be the set of the users connected to BS i at the moment t. Then, the load of BS i at the moment t is obtained by

$$\mu_i(t) = \sum_{k \in \Psi_K^i(t)} \frac{b_{i,k}(t)}{w_{\theta(i)}}. \tag{5}$$

Notice that $\mu_i(t) \in [0, 1]$ since any BS is not allowed to be overloaded. For ease of exploration, we consider two modes for a BS: *sleep mode* and *active mode*. We introduce the following definitions.

- We define P_1^s and P_2^s as the sleep powers of MaBSs and MiBSs, respectively. We consider the sleep powers invariant with time and load because the BSs in sleep state will not bear any traffic load.
- For active mode, we define P_1^a and P_2^a as the active powers of MaBSs and MiBSs. The power of an active BS can be regarded as two parts, transmit power and circuit power [61], where the former is related to the load of BSs, and the latter contains the necessary circuit component energy consumption, which can be regarded as a constant. Specifically, we define p_1^c and p_2^c as the circuit powers of MaBSs and MiBSs. We assume every MaBS and MiBS consists of multiple transmit antennas, and the numbers of transmit antennas of every MaBS and MiBS are α_1 and α_2, respectively. In this context, we use the linear approximation model of [61] to represent the total power of BS i at moment t as follows.

$$P_{\theta(i)}^a(t) = \alpha_{\theta(i)} p_{\theta(i)} \mu_i(t) + p_{\theta(i)}^c. \tag{6}$$



(a) Active mode. (b) Sleep mode.

Fig. 3: Block diagram of the power consumption of a MaBS.

D. BLSTM model

The traffic patterns of users in cellular networks change over time [30]. RNN has gained significant attention in recent years [62] for handling sequence data as input and performing recursion in the direction of the evolution of the sequence. In RNNs, all the nodes (which are artificial neurons) are connected to allow the output of some nodes to impact the subsequent input to the same nodes. However, RNNs usually cannot handle long dependencies in reality, even with carefully chosen parameters, for very long sequences [35].

LSTM is an RNN designed to handle the long-term dependence problem that standard RNNs cannot handle [35]. LSTM lets the model learn how to forget the previous hidden state and update the current state, making it well-suited to capturing long-term temporal dependencies in sequences. The LSTM architecture includes forget, input, and output gate mechanisms that allow the model to forget or retain information selectively. Fig. 4 illustrates the differences between a standard RNN and a standard LSTM regarding their repeating models. At time t, $\boldsymbol{x_t}$, $\boldsymbol{c_t}$, $\tilde{\boldsymbol{c}}_t$, $\boldsymbol{h_t}$, $\boldsymbol{f_t}$, $\boldsymbol{i_t}$, and $\boldsymbol{o_t}$ represent the input, internal state, candidate state, hidden state, forget gate output, input gate output, and output gate output, respectively. The gate structure includes *sigmoid* and *tanh* activation functions, denoted by $\sigma(\cdot)$ and $\tau(\cdot)$, respectively. The function called *sigmoid* compresses the input between 1 and 0, while the function called *tanh* compresses the input between 1 and -1. This feature allows LSTM to forget information by multiplying zero and memorize part or all of the information by multiplying a non-zero number.

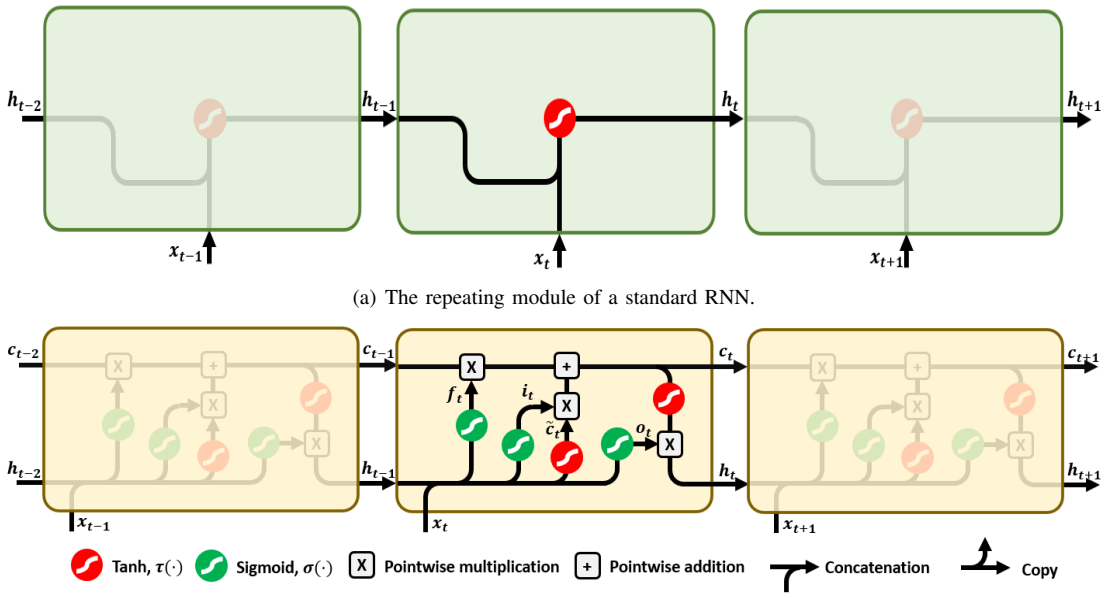
For the repeating model described in Fig. 4(b) at time t, LSTM introduces an internal state $\boldsymbol{c_t} \in \mathbb{R}^D$ to linearly transfer the cyclic information and to output information non-linearly to the hidden state $\boldsymbol{h_t} \in \mathbb{R}^D$. The internal state $\boldsymbol{c_t}$ and hidden state $\boldsymbol{h_t}$ are obtained by

$$\boldsymbol{c_t} = \boldsymbol{f_t} \odot \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \odot \tilde{\boldsymbol{c}}_t, \tag{7}$$

$$\boldsymbol{h_t} = \boldsymbol{o_t} \odot \tau(\boldsymbol{c_t}), \tag{8}$$

where $\odot$ is the pointwise multiplication operation, $\tilde{\boldsymbol{c}}_t \in \mathbb{R}^D$ is the candidate state obtained by

$$\tilde{\boldsymbol{c}}_t = \tau(\boldsymbol{W_c x_t} + \boldsymbol{U_c h_{t-1}} + \boldsymbol{v_c}). \tag{9}$$



(a) The repeating module of a standard RNN.

(b) The repeating module of a standard LSTM.

Fig. 4: Differences between RNN and LSTM.

For the outputs of the three gates, we have

- the forget gate output $\boldsymbol{f_t} \in [0,1]^D$, which controls how much information needs to be forgotten for the last cell state c_{t-1},

$$\boldsymbol{f_t} = \sigma(\boldsymbol{W_f x_t} + \boldsymbol{U_f h_{t-1}} + \boldsymbol{v_f}), \tag{10}$$

- the input gate output $\boldsymbol{i_t} \in [0,1]^D$, which decides how much information needs to be memorized for a candidate state $\tilde{\boldsymbol{c}}_t$,

$$\boldsymbol{i_t} = \sigma(\boldsymbol{W_i x_t} + \boldsymbol{U_i h_{t-1}} + \boldsymbol{v_i}), \tag{11}$$

- and the output gate output $\boldsymbol{o_t} \in [0,1]^D$, which decides how much information the internal state $\boldsymbol{c_t}$ needs to output to the hidden state $\boldsymbol{h_t}$,

$$\boldsymbol{o_t} = \sigma(\boldsymbol{W_o x_t} + \boldsymbol{U_o h_{t-1}} + \boldsymbol{v_o}), \tag{12}$$

where the $\boldsymbol{W}_*$, $\boldsymbol{U}_*$, and $\boldsymbol{v}_*$, $* \in \{i, f, o, c\}$ are the learnable network parameters. Concisely, we describe (9)-(12) as

$$\begin{bmatrix} \tilde{\boldsymbol{c}}_t \\ \boldsymbol{o_t} \\ \boldsymbol{i_t} \\ \boldsymbol{f_t} \end{bmatrix} = \begin{bmatrix} \tau(\cdot) \\ \sigma(\cdot) \\ \sigma(\cdot) \\ \sigma(\cdot) \end{bmatrix} \left(\boldsymbol{W} \begin{bmatrix} \boldsymbol{x_t} \\ \boldsymbol{h_{t-1}} \end{bmatrix} + \boldsymbol{v} \right), \tag{13}$$

where $\boldsymbol{x_t} \in \mathbb{R}^F$ is the input at time t, $\boldsymbol{W} \in \mathbb{R}^{4D \times (D+F)}$ and $\boldsymbol{v} \in \mathbb{R}^{4D}$ are the network parameters of LSTM. (13) together with (7)-(8) comprehensively describe a repeating model of a standard LSTM in Fig. 4(b).

Although the standard LSTM described above is widely used in literature, not all LSTMs are the same. Different publications may use slightly different versions of the LSTM network, and the output at time t may be related to both past and future information. In our paper, we enhance the LSTM network's ability to predict user traffic data more accurately using a Bidirectional LSTM (BLSTM). We add a specific network layer that transmits information in reverse order of time, allowing the BLSTM to process both past and future context while making predictions. This modification improves the BLSTM's ability to capture temporal dependencies in the data and produces more accurate predictions than the standard LSTM. See Fig. 5. The two layers in Fig. 5 have different information transmission directions, where Layer 1 is in order of time, and Layer 2 is reversed in time. The internal state outputs ($\boldsymbol{c}_t^{(1)}$, $\boldsymbol{c}_t^{(2)}$) and hidden state outputs ($\boldsymbol{h}_t^{(1)}$, $\boldsymbol{h}_t^{(2)}$) of the two layers are obtained in the same vain of $\boldsymbol{c_t}$ and $\boldsymbol{h_t}$ of the standard LSTM by (7), (8), and (13). Then, we can obtain the internal state output and hidden state output of BLSTM at time t by

$$\boldsymbol{h}_t^{\boldsymbol{B}} = \boldsymbol{h}_t^{(1)} \oplus \boldsymbol{h}_t^{(2)}, \tag{14}$$

$$\boldsymbol{c}_t^{\boldsymbol{B}} = \boldsymbol{c}_t^{(1)} \oplus \boldsymbol{c}_t^{(2)}, \tag{15}$$

where $\oplus$ is the concatenation operation.

The computational complexity of the BLSTM model depends on factors such as the input sequence length, the number of layers in the model, and the number of LSTM units in each layer. The time complexity of the BLSTM model can be high for long input sequences and models with many layers and LSTM units. It is worth mentioning that there are techniques to mitigate this complexity. For example, truncated backpropagation through time (TBPTT) can be employed during training to limit the number of time steps considered, reducing the overall computational burden. Readers are referred to [63, 64] for more details on this.

The BLSTM model also has some disadvantages compared to other types of RNNs, such as computational complexity, higher memory requirements, overfitting risks, and sensitivity to hyperparameters. However, considering that traffic predic-

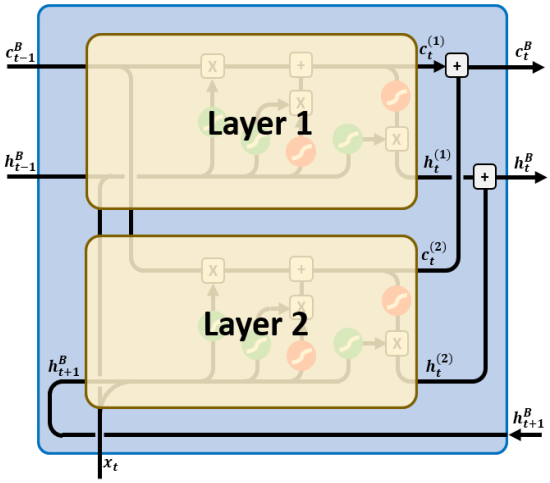


Fig. 5: Architecture of the BLSTM.

tion accuracy is very important in our problem, which may have a large influence on the performance of our proposed sleeping strategy, we chose the BLSTM model, which performs better than the RNN for our dataset. Our comparison of the accuracy of the BLSTM and the RNN models on our dataset shows the superiority of the BLSTM model, which we have presented in Section VI.A.

IV. Problem Formulation

We aim to reduce the total energy consumption of all BSs in the two-tier HeCN by selectively placing MiBSs in sleep mode during periods of low network load. To achieve this, we impose three constraints: first, we ensure that all users' SINRs remain above the threshold γ_0; second, we ensure that there is no degradation in the QoS for each user, which is closely related to the demand traffic rates; and third, we ensure that the load of each BS does not exceed 1.

Define an *action vector* $\boldsymbol{a_t^\phi} = (a_i^\phi(t), i \in \Psi_M \cup \Psi_S)$ as a state vector associated with strategy ϕ that

- if $a_i^\phi(t) = 1$, then BS i at time t is in active mode;
- otherwise, BS i at time t is in sleep mode.

In this context, the set of all active BSs and sleeping BSs are $\mathcal{I}^{\text{active}}(t) = \{i \mid a_i^\phi(t) = 1\}$ and $\mathcal{I}^{\text{sleep}}(t) = \{i \mid a_i^\phi(t) = 0\}$, respectively. Let $R_k = \max_{t\in[T_1,T_2]} r_k(t), k \in \Psi_K$ be the set of maximum demand rates of users within low-load time period $[T_1, T_2]$. To mathematically define the feasibility of strategy ϕ during low-load time period $[T_1, T_2]$, we introduce the following constraints for the action variables:

$$\begin{cases} S_{i,k} \geqslant \gamma_0 \\ \mu_i(t) + \frac{R_k}{w_{\theta(i)} \log_2(1+S_{i,k})} \leq 1 \end{cases}, \quad \forall k \in \Psi_K, t \in [T_1, T_2], \exists i \in \mathcal{I}^{\text{active}}(t). \tag{16}$$

(16) ensures that every user in the network has access to at least one active BS that can maintain an SINR above the threshold γ_0 while also being capable of accommodating the user's maximum traffic demand rate during the low-load period $[T_1, T_2]$.

The different traffic demands of each user connected to an active BS jointly constitute a load of the BS. Since any BS is not allowed to be overloaded, we have

$$i \in \mathcal{I}^{\text{active}}(t), 0 \leq \mu_i(t) \leq 1, \ \forall t \in [T_1, T_2]. \tag{17}$$

Recall that we only consider the sleeping operations on MiBSs due to the high flexibility of MiBSs and the ease of switching on and off. This can be addressed by

$$\mathcal{I}^{\text{sleep}}(t) \cap \Psi_M = \emptyset, \ \forall t \in [T_1, T_2]. \tag{18}$$

We aim to minimize the energy consumption of the BSs in the two-tier HeCN during a low-load time period $[T_1, T_2]$. Precisely, let

$$\mathbb{P}(t) = \sum_{i\in\Psi_M\cup\Psi_S} \left(a_i^\phi(t) P^a_{\theta(i)}(t) + \left(1 - a_i^\phi(t)\right) P^s_{\theta(i)}\right), \quad t \in [T_1, T_2], \tag{19}$$

be the total power of all BSs in the two-tier HeCN at time $t \in [T_1, T_2]$, then our optimization problem is

$$\min_{\phi} \int_{T_1}^{T_2} \mathbb{P}(t) dt, \tag{20}$$

subject to (16), (17), and (18). Let Φ represent the set of all strategies constrained by (16), (17), and (18).

V. MiLSF Strategy

This section describes the MiLSF strategy based on the user traffic demands $r_k(t)$ predicted by BLSTM and the load of BSs $\mu_i(t)$. Notice that we only consider the sleeping operation on MiBSs in this paper due to the high flexibility and ease of operation. Furthermore, the frequent switching of the state of BSs will cause additional network delay and extra energy consumption for BS initialization. Therefore, in this paper, for the low-load period $[T_1, T_2]$, we only consider one sleep operation for the MiBSs, which occurs at the start time of the low-load period, namely T_1. Before the T_1, we assume all M MaBSs and N MiBSs are in active mode. As the low-load period is typically during the night period when human activity is low, the mobility of users is also very low. For simplicity, we assume that mobile terminals or users are stationary during this period. However, due to the unpredictable nature of human activity, this assumption may not be practical for all users. Therefore, further research is needed to understand the impact of mobility, which we plan to investigate in future studies.

The detailed implementation process of our MiLSF strategy is listed as follows.

1) **Sorting**. We sort the active MiBSs based on their load situation at time T_1 from low to high.
2) **Sleeping**. Using the sorting order from **Step (1)**, we evaluate each MiBS to determine whether it can be switched to sleep mode. If an active BS exists, including MaBSs and MiBSs, that can meet the maximum traffic demand rate during the low-load period $[T_1, T_2]$ of every user currently connected to the MiBS, then the MiBS can be switched to sleep mode, and we move on to **Step (3)**.

Otherwise, we will continue evaluating the next MiBS until we find one that can be switched to sleep mode.

3) **User reallocation**. We select active BSs, including MaBSs and MiBSs, that can meet the maximum traffic demand rate during the low-load period $[T_1, T_2]$ of every user connected to the MiBS which will be switched to sleep mode in **Step (2)**. We then reassign the users to the selected active BSs. We preferentially reallocate each user to the MaBS with the highest SINR among the selected BSs. If no qualified MaBS exists among the selected BSs, we reallocate the user to the MiBS with the highest load among the selected BSs.
4) **Repeating**. We repeat **Steps (1)** and **(2)** until no more MiBSs can be switched to sleep mode.

In **Step (3)**, since our goal is to have as many MiBSs as possible in sleep mode, to save energy without degrading user service quality and sleeping any MaBS, priority is given to the active MaBS with the highest SINR among the selected BSs. Since all the MaBSs have the same power consumption characteristics, assigning users to the MaBS with the highest SINR can obtain the best spectral efficiency.

Suppose there is no active MaBS among the selected BSs in **Step (3)**. In this case, we reallocate users to the MiBS with the maximum load, which helps speed up the execution process and reduce the complexity since we always try to sleep the minimum load MiBS first in each cycle of the above execution.

We propose the pseudo-code of our MiLSF strategy ϕ^* in Algorithm 1. The tie-breaking rules in Line 8 and Line 20 of Algorithm 1 can be arbitrary if *argmax* returns more than one argument, and *card*$(\cdot)$ returns the cardinality of a set.

VI. Numerical Results

In this section, we numerically demonstrate the effectiveness of our MiLSF strategy by comparing it with the following four other baseline strategies in different scenarios.

1) *Randomly Sleep* (RS). We randomly select active MiBSs (see **Step (2)** in Section V) to try to have them in sleep mode until no remaining MiBS can be switched to sleep mode. The user reallocation process of users in RS is the same as that in MiLSF (see **Step (3)** in Section V).
2) *Randomly Reallocate Users* (RRU). The selection process of active MiBSs is the same as the selection process of active MiBSs in MiLSF (see **Step (2)** in Section V), but under RRU, the users will be randomly reallocated to active BSs that meet the SINR threshold and can guarantee the users' traffic demand rates, which is different from **Step (3)** in MiLSF.
3) *Closest User Reallocation* (CUR) [65]. The CUR strategy always reallocates a user to the closest available BSs (including both MaBSs and MiBSs) that can guarantee the maximum traffic demand rate for this user during the low-load period $[T_1, T_2]$ with an SINR larger than the threshold, while the selection of MiBSs follows the same principle of our MiLSF strategy (see **Step (2)** in Section V).
4) *Closest Base Station Sleep First* (CBSSF) [66]. In the CBSSF strategy, the MiBS closer to the MaBSs is

Algorithm 1: MiLSF strategy ϕ^* in a two-tier HeCN.

Input: $R_k, k \in \Psi_K; P_o^s, p_o^c, p_o, \alpha_o, f_o^c, w_o, o \in \{1,2\}; \Psi_K^i(T_1), \mu_i(T_1), i \in \Psi_M \cup \Psi_S$
Output: $\boldsymbol{a}_{\boldsymbol{T_1}}^{\boldsymbol{\phi^*}}$

1 sort N active MiBSs $(B_i, i \in \Psi_S)$ according to the loads in ascending order;
2 **foreach** $i \in \Psi_S$ **do**
3 $k \leftarrow 1$;
4 $\xi_k \leftarrow 0$;
5 **while** $k \le card\big(\Psi_K^i(T_1)\big)$ **do**
6 $\Gamma_{\text{MaBS}} \leftarrow \Psi_M \setminus \{m \mid \mu_m(T_1) = 1 \text{ or } S_{m,k} < \gamma_0\}$;
7 **while** $\Gamma_{MaBS} \neq \emptyset$ & $\xi_k = 0$ **do**
8 $j_* \leftarrow \underset{j \in \Gamma_{\text{MaBS}}}{\arg\max}\ S_{j,k}$;
9 $\Delta\mu_{j_*,k} \leftarrow \dfrac{R_k}{w_1 \log_2(1 + S_{j,k})}$;
10 **if** $\Delta\mu_{j_*,k} + \mu_{j_*} > 1$ **then**
11 $\Gamma_{\text{MaBS}} \leftarrow \Gamma_{\text{MaBS}} \setminus \{j_*\}$;
12 **else**
13 $\mu_{j_*}(T_1) \leftarrow \mu_{j_*}(T_1) + \Delta\mu_{j_*,k}$;
14 $\xi_k \leftarrow 1$;
15 **end**
16 **end**
17 **if** $\xi_k = 0$ **then**
18 $\Gamma_{\text{MiBS}} \leftarrow \Psi_S \setminus \{n \mid \mu_n = 1 \text{ or } S_{n,k} < \gamma_0\}$;
19 **while** $\Gamma_{MiBS} \neq \emptyset$ & $\xi_k = 0$ **do**
20 $j_* \leftarrow \underset{j \in \Gamma_{\text{MiBS}}}{\arg\max}\ \mu_{j,k}$;
21 $\Delta\mu_{j_*,k} \leftarrow \dfrac{R_k}{w_2 \log_2(1 + S_{j,k})}$;
22 **if** $\Delta\mu_{j_*,k} + \mu_{j_*} > 1$ **then**
23 $\Gamma_{\text{MiBS}} \leftarrow \Gamma_{\text{MiBS}} \setminus \{j_*\}$;
24 **else**
25 $\mu_{j_*}(T_1) \leftarrow \mu_{j_*}(T_1) + \Delta\mu_{j_*,k}$;
26 $\xi_k \leftarrow 1$;
27 **end**
28 **end**
29 **end**
30 **if** $\xi_k = 1$ **then**
31 $k \leftarrow k + 1$;
32 **end**
33 **end**
34 **if** $\prod_{k \in \mathcal{K}_i(T_1)} \xi_k = 1$ **then**
35 $a_i^{\phi^*}(T_1) \leftarrow 0$;
36 **else**
37 $a_i^{\phi^*}(T_1) \leftarrow 1$;
38 **end**
39 $i \leftarrow i + 1$;
40 **end**

prioritized to consider sleep, while the user reallocation process is the same as that in MiLSF (see **Step (3)** in Section V).

In Section VI-A, we introduce the traffic data source and demonstrate the prediction of user traffic demand rates using BLSTM. In Section VI-B, we demonstrate the superiority of our proposed MiLSF strategy over RS and RRU in different scenarios. We obtained the 95% confidence intervals based on the *Student's t-distribution* for the results on energy-saving percentages of MiLSF, RS, and RRU relative to the non-sleeping strategy, all within 2% of the observed mean.

A. User traffic data prediction by BLSTM

We obtain user traffic data from [67], which analyzed the week-long traffic data of medium-sized Chinese cities with large populations. To predict each user's traffic demand for the next day, we use their historical traffic demand rates over the past week for the modeling training. Specifically, we split each user's hourly average traffic rate dataset for eight consecutive days into two parts: the first seven days for training and the eighth day for testing.

We use the BLSTM neural network, which consists of an input layer for receiving user traffic data sequences, two LSTM layers with 500 hidden units each, a fully connected layer with an output size matching the input size, and a regression layer that computes the half-mean-squared-error loss for regression tasks [68]. We use the RMSProp as the optimizer when training our model, and the parameters, including the optimization algorithm, learning rate, and batch size, are set as default values of this optimizer.

To train the BLSTM neural network on the above-mentioned real-world user traffic data from [67], we observe that user traffic exhibits a daily periodic characteristic. Thus, we designed a special training approach that utilizes this characteristic. Firstly, we trained the model on the user traffic of the first two days and used it to predict the traffic for the third day. We then recorded the errors between the predicted and actual traffic and modified the model's parameters (such as gate parameters) to minimize the errors after modification. With the modified model, we predicted the user traffic of the fourth day based on the traffic of the last two days (i.e., the second and third days), recorded the errors, and modified the model's parameters. We repeated these steps until we used all the traffic data for seven days, which marks the end of one round of training. We repeated the above round until the average error in this round was lower than a given threshold, thus completing the model training.

To assess the performance of our BLSTM neural network, we utilize two commonly used metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Specifically, MAE is obtained by the average of the absolute prediction errors, while RMSE is computed as the square root of the average of the squared prediction errors.

We employ 20% hold-out validation to assess the performance of the BLSTM, RNN, and ARIMA models. Specifically, we reserve 20% of the user traffic data from the reference [67] for validation, while the remaining 80% is utilized for training. Table I presents the hold-out validation results of these three models on the validation set, where BLSTM shows great superiority for user traffic prediction over RNN and ARIMA with the lowest MAE and RMSE.

TABLE I: 20% hold-out validation of the BLSTM, RNN, and ARIMA models.

	BLSTM	RNN	ARIMA
MAE	62.23	79.81	232.46
RMSE	91.67	144.09	398.77

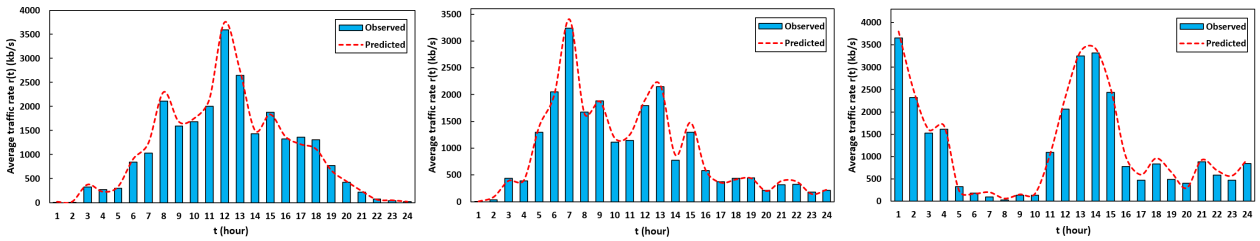
In Figs. 6, 7, and 8, we present the hourly traffic rate predictions for three users with distinct traffic patterns, as predicted by the BLSTM, RNN, and ARIMA models, respectively. The results presented by each subfigure are based on the performance metrics MAE and RMSE. These figures serve to illustrate the prediction accuracy of each model for different traffic patterns.

It can be seen that the prediction results of ARIMA differ the most from the actual values, which can be observed both from the MAE and RMSE metrics. This is because ARIMA is a classic linear model, assuming that time series data has a linear relationship, but actual data may contain nonlinear relationships, thus may cause inaccurate prediction results for nonlinear time series [69–71]. The analysis of nonlinear time series requires using corresponding nonlinear models or methods to model and predict data, such as RNN and LSTM in neural network models.

We observe that the MAE of the prediction results for the three users is similar between RNN and BLSTM (BLSTM has a slight advantage over RNN), indicating comparable average prediction accuracy for both methods. However, BLSTM significantly outperforms RNN in predicting peak rates, as reflected by the much lower RMSE of BLSTM for the three users in Fig. 6 and Fig. 7. Therefore, we use BLSTM to predict the future traffic rates of every user in the two-tier HeCN. Based on the predicted user traffic rates from BLSTM, we implement our MiLSF strategy and demonstrate its superiority over RS and RRU strategies in different scenarios.

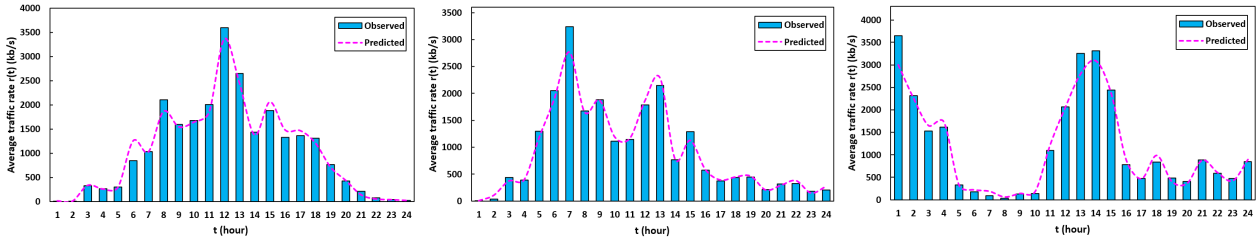
B. Effectiveness of MiLSF

Here we provide a comparison between our proposed MiLSF strategy with the four baseline strategies, RS, RRU, CUR, and CBSSF in various scenarios based on predicted data of user traffic rates by BLSTM. We use the simulation parameter values in Table II unless otherwise specified. Although different users have varying traffic demands at different times, most users typically have lower traffic demands at night, resulting in a lower overall network load during this period. This provides an opportunity for BS sleeping strategies to save energy and therefore, by considering night-time, we can also demonstrate the superiority of MiLSF over the four baseline strategies. Additionally, we assume that users do not move during the low-load period at night, as user mobility is significantly reduced during this time. We consider the *low-load period* to be between 10:00 p.m. and 6:00 a.m. the following day as the sleeping period. That is, a selected set of MiBSs



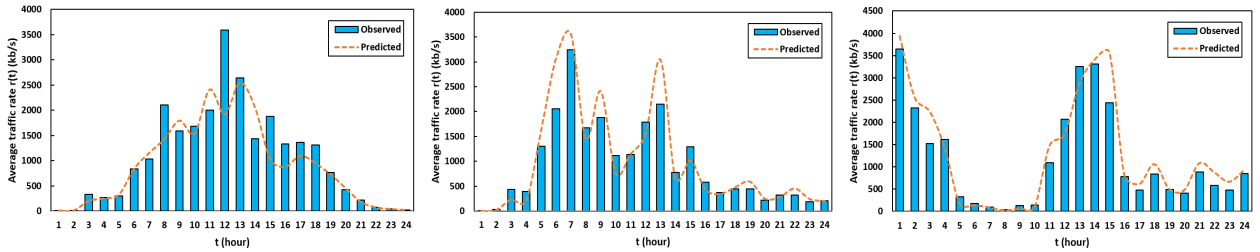
(a) User 1, MAE = 75.56, RMSE = 98.32. (b) User 2, MAE = 55.78, RMSE = 75.17. (c) User 3, MAE = 99.26, RMSE = 115.23.

Fig. 6: The hourly average traffic rates prediction by BLSTM for three users.



(a) User 1, MAE = 87.89, RMSE = 135.52. (b) User 2, MAE = 68.07, RMSE = 118.93. (c) User 3, MAE = 104.99, RMSE = 179.61.

Fig. 7: The hourly average traffic rates prediction by RNN for three users.



(a) User 1, MAE = 257.931, RMSE = 460.085. (b) User 2, MAE = 221.24, RMSE = 339.779. (c) User 3, MAE = 221.132, RMSE = 332.755.

Fig. 8: The hourly average traffic rates prediction by ARIMA for three users.

enter sleep mode at 10:00 p.m. and enter active mode at 6:00 a.m., and we aim to select these MiBSs optimally. We use PPP to determine user locations. We randomly allocate each user to a BS that meets the SINR threshold and the user's traffic demand rate for initialization. We then implement MiLSF, RS, RRU, CUR, and CBSSF strategies during the low-load period and compare their energy-saving performances relative to a non-sleeping strategy. Our numerical results demonstrate the effectiveness and superiority of MiLSF over the four baseline strategies.

1) Scenario I: In this scenario, we compare the energy-saving performance of MiLSF with the four baseline strategies under two different BS deployment methods. Specifically, for each strategy, we use two different deployment processes, PPP and MHCPP, to determine the locations of all BSs. It is important to note that the deployment processes of MaBSs and MiBSs are independent and have different intensities, as shown in Table II. By comparing the energy-saving performance of each strategy under these different deployment methods, we can gain insight into the effectiveness of each strategy under different BS deployment methods.

In Fig. 9, we observe the effect of different BS deployment

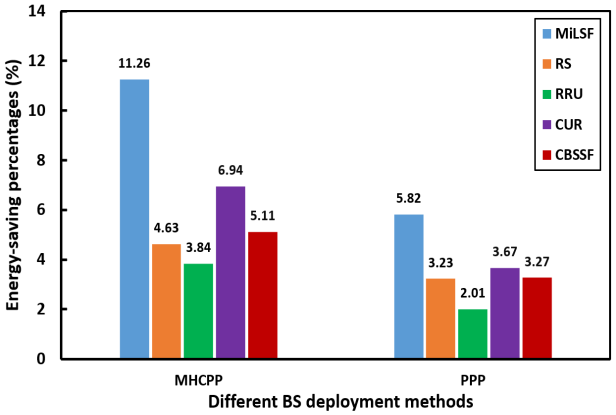


Fig. 9: Comparison of energy-saving performance between MiLSF and the four baseline strategies under both PPP and MHCPP conditions.

methods on the energy-saving performance of BS sleeping strategies. The results show that MiLSF outperforms all the four baseline strategies under both PPP and MHCPP deployments. However, using MHCPP instead of PPP for BS

TABLE II: Simulation parameters.

Description	Value
The two-tier HeCN area	10×10 km^2
Intensity of MaBSs for PPP deployment, λ_m^{PPP}	2/km^2
Intensity of MiBSs for PPP deployment, λ_s^{PPP}	4/km^2
Intensity of MaBSs for MHCPP deployment, λ_m^{MHCPP}	2/km^2
Intensity of MiBSs for MHCPP deployment, $\lambda_s^{\text{MHCPPP}}$	4/km^2
Hard-core parameter for MaBSs, $r_{h,m}$	2 km
Hard-core parameter for MiBSs, $r_{h,s}$	1 km
Intensity of users in the HeCN for PPP deployment, λ_u	25/km^2
Transmit power of a single antenna in MaBSs, p_1	8 W
Transmit power of a single antenna in MiBSs, p_2	3 W
Number of transmit antennas in a MaBS, α_1	6
Number of transmit antennas in a MiBS, α_2	2
Carrier frequency of MaBSs, f_1^c	2.4 GHz
Carrier frequency of MiBSs, f_2^c	20 GHz
Bandwidth of a MaBS, w_1	20 MHz
Bandwidth of a MiBS, w_2	50 MHz
Circuit power of MaBSs, p_1^c	120 W
Circuit power of MiBSs, p_2^c	10 W
Sleep power of MaBSs, p_1^s	8 W
Sleep power of MiBSs, p_2^s	2 W
Path loss exponent, β	3.7
SINR threshold (dB), γ_0	-6

deployment significantly increases the energy-saving percentage for the same strategy. For example, the MiLSF strategy achieves an energy-saving percentage of 5.82% under PPP deployment, but it achieves an energy-saving percentage of 11.26% under MHCPP deployment, almost twice that under PPP. The disadvantage of PPP is due to the presence of the same-type BSs that are too close to each other, resulting in significant signal interference. This reduces the number of selectable BSs for users in **Step (3)**, resulting in a significant reduction in the number of MiBSs that can eventually be switched to sleep mode and a reduction in energy-saving percentages. Therefore, we use MHCPP for BS deployment in Scenarios II, III, and IV.

2) Scenario II: In this scenario, we observe the effect of user loads on the energy-saving performance of MiLSF the four baseline strategies by varying the user intensity λ_u in Fig. 10.

Fig. 10 demonstrates the superiority of MiLSF over all the four baseline strategies, with MiLSF achieving higher energy-saving percentages under different numbers of users, particularly when the number of users is moderate. This is because, unlike RS, MiLSF always selects the MiBSs with the lowest load and reallocates connected users to other BSs. Sleeping the lowest-load MiBS is easier than high-load MiBSs, as users connected to the lowest-load MiBS are more likely to be "satisfied" and assigned to other BSs. Prioritizing user allocation of lowest-load MiBS to other BSs also reduces the likelihood of problems during subsequent sleep operations of other MiBSs. Additionally, MiLSF prioritizes user allocation to the MaBS with the maximum SINR, leading to higher energy-saving percentages than RRU, where users are reallocated to a random active BS that may not use full use of the MBS's bandwidth resource.

In the CBSSF strategy, MiBSs that are closest to the MaBSs are always prioritized for sleep consideration. Consequently, users previously associated with these MiBSs are most likely to be reallocated to the nearest corresponding MaBSs. This characteristic results in an overemphasis on MaBSs during the user reallocation process, triggering premature load saturation in the MaBSs. Concurrently, MiBSs that cannot transition to sleep mode (if one user solely relies on a particular MiBS and can not be allocated to other BSs) remain underutilized, operating in a low-load working state without the option to enter sleep mode. This situation reduces the total number of MiBSs that can ultimately be switched to sleep mode, significantly diminishing the energy-saving percentage compared to that achieved by the MiLSF strategy.

As for the CUR strategy, it simply reallocates users to the BSs that are closest to them and can guarantee the maximum user traffic demand with an SINR larger than the threshold. However, it overlooks the load distribution of each BS across the entire network and the different SINR situations that various BSs can offer. Unlike our MiLSF strategy, which fully considers the load information of all BSs in the network and

the optimal SINR throughout its implementation, the CUR strategy does not achieve as effective global optimization in energy saving across the network as MiLSF does.

Furthermore, Fig. 10 shows that the energy-saving percentages of five strategies are similar when there are either no users or a large number of users (close to 400). This is because, under all the strategies, all the MiBSs sleep when the number of users is small, or remain in active mode when the number of users is large, resulting in similar energy-saving performance. Additionally, the more users in the network, the lower the energy-saving percentage of all five strategies, as fewer MiBSs can be switched to sleep mode.

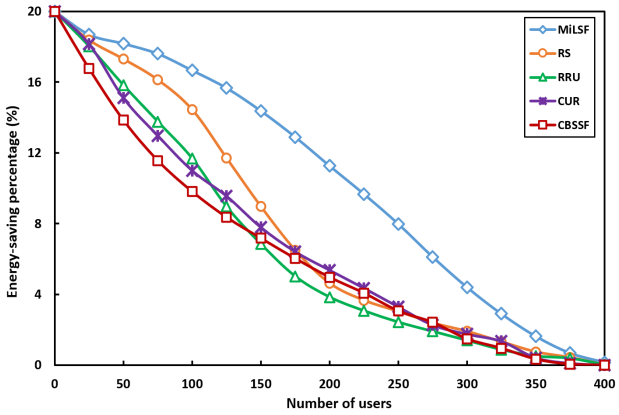


Fig. 10: Performance comparison between MiLSF and the four baseline strategies across different user numbers.

3) Scenario III: In this scenario, we observe the effect of the SINR threshold value γ_0 on the energy-saving performance of MiLSF and the four baselines in Fig. 11. The results show that MiLSF outperforms all four baselines under different SINR thresholds. Additionally, the energy-saving percentages of all five strategies gradually decrease as the SINR threshold increases, as the number of BSs that can be selected by each user for reallocation decreases, reducing the possibility of successfully reallocating all users of each MiBS and resulting in fewer MiBS that can be switched to sleep mode.

Moreover, the most significant difference in energy-saving percentages between MiLSF and the four baseline strategies occurs when the SINR threshold is moderate. When the SINR threshold is too large (i.e., $\gamma_0 = 0$ dB), very few active BSs can be selected to meet the SINR threshold for user reallocation in **Step (3)**, making it challenging for all five strategies to reallocate users successfully. As a result, the difference in energy-saving percentages is small. Conversely, when the SINR threshold is too small (i.e., $\gamma_0 = -12$ dB), almost all active BSs can become qualified active BSs for user reallocation in **Step (3)**, making it easier for all five strategies to reallocate all users connected to a MiBS and achieve higher energy-saving percentages.

When the SINR threshold is moderate, the user reallocation in **Step (3)** cannot be too "arbitrary", as there may be only a few (or even one) qualified active BSs for each user. In this case, MiLSF prioritizes the MaBS with the highest SINR, maximizing the probability of satisfying all users' demand traffic rates and sleeping as many MiBSs as possible. Therefore, when the SINR threshold is moderate, MiLSF has a significant energy-saving advantage versus RS, RRU, CUR, and CBSSF.

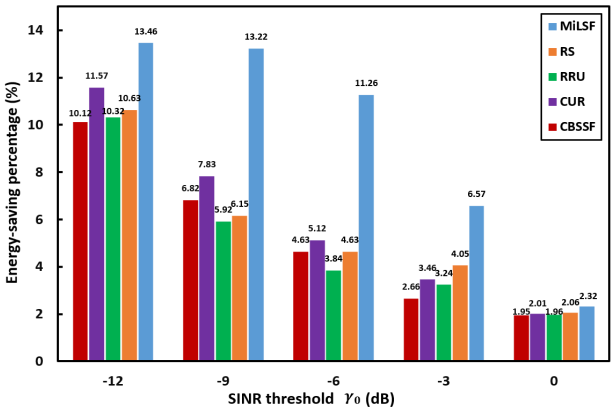


Fig. 11: Performance comparison between MiLSF and the four baseline strategies across different SINR thresholds γ_0.

4) Scenario IV: In this scenario, we observe the relationship between the number of sleeping MiBSs and the energy-saving percentages for five strategies in Fig. 12. The results demonstrate a linearly positive correlation between the number of sleeping MiBSs and the energy-saving percentage. Notice that there is only a small difference in the energy-saving percentage among the five strategies with the same number of sleeping MiBSs. Because under MiLSF and the four baseline strategies, different MiBSs may sleep, resulting in different user reallocation situations (users may be reallocated to different MaBSs or MiBSs). This small difference in the energy-saving performances between MiLSF and the four baselines shows that if we only consider different BS sleeping strategies or user reallocations without switching more MiBSs to sleep mode, it is challenging to reduce energy consumption significantly. Therefore, a BS sleeping strategy that can have more MiBSs in sleep mode is always preferred. In this case, the results in Fig. 9, Fig. 10, and Fig. 11 illustrate that, given the network parameters, our MiLSF strategy can always achieve a higher energy-saving percentage than the four baseline strategies by sleeping more MiBSs, demonstrating its effectiveness and superiority over RS, RRU, CUR, and CBSSF.

VII. Conclusions

We have proposed a new BS sleeping strategy named MiLSF for a two-tier HeCN, based on predicted traffic demand rates for each user. Underutilized MiBSs are switched to sleep mode, and users are reallocated to other active BSs that meet the SINR threshold and can guarantee the users' maximum demand traffic rates during the low-load period. We have used BLSTM to predict each user's future traffic demand rates based on historical traffic data over the past week, achieving higher prediction accuracy than RNN, particularly for peak rates. We have implemented our MiLSF strategy during a low-load period at night to explore its energy-saving performance in different scenarios. Specifically, our MiLSF strategy has always tried to have the least loaded MiBS in

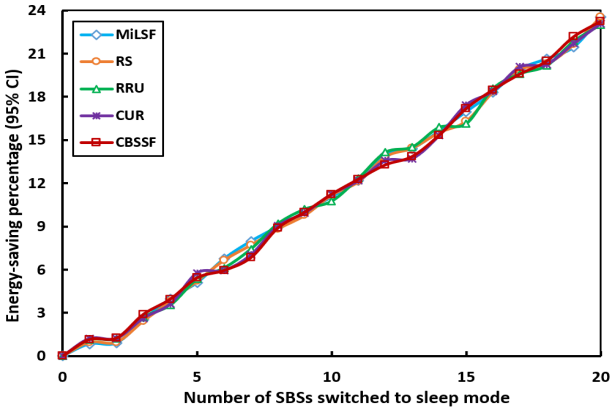


Fig. 12: Performance comparison between MiLSF and the four baseline strategies across different numbers of sleeping MiBSs.

sleep mode and reallocate users to the MaBS with the highest SINR that can guarantee their traffic demand rates. Through extensive numerical simulations, we have demonstrated the effectiveness and superiority of the MiLSF strategy over the four baselines, RS, RRU, CUR, and CBSSF. Furthermore, although we only consider one sleep operation for a two-tier HeCN, our MiLSF strategy can be extended to any multi-layer HeCNs for multiple sleep operations during various low-load periods.

While our work focuses on improving energy efficiency while guaranteeing the QoS of each service with a predefined SINR ratio, we acknowledge that load balancing is an important aspect to consider in network design. Load balancing can help evenly distribute network traffic and improve the quality of service for end-users. While we have ensured the QoS of each deployed service, we recognize that combining our strategy with load-balancing techniques could further improve the overall performance and efficiency of the network. Therefore, we propose to investigate load-balancing mechanisms or algorithms in future work to complement our energy-efficient MiLSF strategy.

REFERENCES

[1] M. Hasan, "State of IoT 2022: Number of connected IoT devices growing 18% to 14.4 billion globally," IoT Analytics, Zirkusweg 2, 20359 Hamburg, Germany, Tech. Rep., May 2022 (accessed on March 10, 2023). [Online]. Available: https://iot-analytics.com/number-connected-iot-devices/

[2] R. Möller, "Ericsson Mobility Report," Ericsson, SE-164 80 Stockholm, Sweden, Tech. Rep., June 2022 (accessed on March 10, 2023). [Online]. Available: https://www.ericsson.com/49d3a0/assets/local/reports-papers/mobility-report/documents/2022/ericsson-mobility-report-june-2022.pdf

[3] C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair, and A. Friday, "The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations," *Patterns*, vol. 2, no. 9, p. 100340, 2021.

[4] T. Yang, F. Heliot, and C. H. Foh, "A survey of green scheduling schemes for homogeneous and heterogeneous cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 11, pp. 175–181, 2015.

[5] J. Wu, S. Zhou, and Z. Niu, "Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks," *IEEE Trans. Wirel. Commun.*, vol. 12, no. 8, pp. 4196–4209, 2013.

[6] Z. Niu, X. Guo, S. Zhou, and P. R. Kumar, "Characterizing energy-delay tradeoff in hyper-cellular networks with base station sleeping control," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 4, pp. 641–650, 2015.

[7] S. Salim, C. H. Oey, and S. Moh, "Are heterogeneous cellular networks superior to homogeneous ones?" in *Proc. 9th IFIP Int. Conf. Netw. Parallel Comput. (NPC)*. Springer, 2012, pp. 61–68.

[8] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. 30th IEEE Conf. Comput. Commun. (INFOCOM)*, 2011, pp. 882–890.

[9] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area," in *Proc. 13th IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, 2015, pp. 230–235.

[10] Y. Wang, M. Faloutsos, and H. Zang, "On the usage patterns of multimodal communication: Countries and evolution," in *Proc. 32nd IEEE Conf. Comput. Commun. (INFOCOM)*, 2013, pp. 3135–3140.

[11] Q. Wu, X. Chen, Z. Zhou, L. Chen, and J. Zhang, "Deep reinforcement learning with spatio-temporal traffic forecasting for data-driven base station sleep control," *IEEE/ACM Trans. Netw.*, vol. 29, no. 2, pp. 935–948, 2021.

[12] G. Jang, N. Kim, T. Ha, C. Lee, and S. Cho, "Base station switching and sleep mode optimization with LSTM-based user prediction," *IEEE Access*, vol. 8, pp. 222 711–222 723, 2020.

[13] R. Shinkuma, N. Kishi, K. Ota, M. Dong, T. Sato, and E. Oki, "Smarter base station sleeping for greener cellular networks," *IEEE Netw.*, vol. 35, no. 6, pp. 98–103, 2021.

[14] J. Lin, Y. Chen, H. Zheng, M. Ding, P. Cheng, and L. Hanzo, "A data-driven base station sleeping strategy based on traffic prediction," *IEEE Trans. Netw. Sci. Eng.*, 2021.

[15] I. Donevski, G. Vallero, and M. A. Marsan, "Neural networks for cellular base station switching," in *Proc. 38th IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2019, pp. 738–743.

[16] C. Liu, Y. Wan, L. Tian, Y. Zhou, and J. Shi, "Base station sleeping control with energy-stability tradeoff in centralized radio access networks," in *Proc. 34th IEEE Glob. Commun. Conf. (GLOBECOM)*, 2015, pp. 1–6.

[17] M. Wildemeersch, T. Q. Quek, C. H. Slump, and A. Rabbachin, "Cognitive small cell networks: Energy efficiency and trade-offs," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 4016–4029, 2013.

[18] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal,

M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews, P. Xia, H. S. Jo, H. S. Dhillon, and T. D. Novlan, "Heterogeneous cellular networks: From theory to practice," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54–64, 2012.

[19] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Commun. Surv. Tutor.*, vol. 15, no. 3, pp. 996–1019, 2013.

[20] W. Guo, S. Wang, X. Chu, J. Zhang, J. Chen, and H. Song, "Automated small-cell deployment for heterogeneous cellular networks," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 46–53, 2013.

[21] H. T. Cheng, A. Callard, G. Senarath, H. Zhang, and P. Zhu, "Step-wise optimal low power node deployment in LTE heterogeneous networks," in *Proc. 76th IEEE Veh. Technol. Conf. (VTC Fall)*, 2012, pp. 1–4.

[22] B. Yang, G. Mao, X. Ge, M. Ding, and X. Yang, "On the energy-efficient deployment for ultra-dense heterogeneous networks with NLoS and LoS transmissions," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 2, pp. 369–384, 2018.

[23] Z. Mlika, E. Driouch, and W. Ajib, "Energy-efficient base station operation and association in HetNets: Complexity and algorithms," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 4, pp. 2690–2702, 2018.

[24] S. Ak, H. Inaltekin, and H. V. Poor, "A tractable framework for the analysis of dense heterogeneous cellular networks," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 3151–3171, 2018.

[25] C. S. Chen, V. M. Nguyen, and L. Thomas, "On small cell network deployment: A comparative study of random and grid topologies," in *Proc. 76th IEEE Veh. Technol. Conf. (VTC Fall)*, 2012, pp. 1–5.

[26] C. Coletti, P. Mogensen, and R. Irmer, "Deployment of LTE in-band relay and micro base stations in a realistic metropolitan scenario," in *Proc. 74th IEEE Veh. Technol. Conf. (VTC Fall)*, 2011, pp. 1–5.

[27] C. Coletti, L. Hu, N. Huan, I. Z. Kovács, B. Vejlgaard, R. Irmer, and N. Scully, "Heterogeneous deployment to meet traffic demand in a realistic LTE urban scenario," in *Proc. 76th IEEE Veh. Technol. Conf. (VTC Fall)*, 2012, pp. 1–5.

[28] L. Hu, I. Z. Kovács, P. Mogensen, O. Klein, and W. Stormer, "Optimal new site deployment algorithm for heterogeneous cellular networks," in *Proc. 74th IEEE Veh. Technol. Conf. (VTC Fall)*, 2011, pp. 1–5.

[29] A. Cao, Y. Qiao, K. Sun, H. Zhang, and J. Yang, "Network traffic analysis and prediction of Hotspot in cellular network," in *Proc. 6th IEEE Int. Conf. Netw. Infrastruct. Digit. Content (IC-NIDC)*, 2018, pp. 452–456.

[30] F. Xu, Y. Lin, J. Huang, D. Wu, H. Shi, J. Song, and Y. Li, "Big data driven mobile traffic understanding and forecasting: A time series approach," *IEEE Trans. Serv. Comput.*, vol. 9, no. 5, pp. 796–805, 2016.

[31] K. Sultan, H. Ali, and Z. Zhang, "Call detail records driven anomaly detection and traffic prediction in mobile cellular networks," *IEEE Access*, vol. 6, pp. 41 728–41 737, 2018.

[32] M. Zhang, H. Fu, Y. Li, and S. Chen, "Understanding urban dynamics from massive mobile traffic data," *IEEE Trans. Big Data*, vol. 5, no. 2, pp. 266–278, 2017.

[33] M. Xu, Q. Wang, and Q. Lin, "Hybrid holiday traffic predictions in cellular networks," in *Proc. 16th IEEE/IFIP Netw. Oper. Manage. Symp. (NOMS)*, 2018, pp. 1–6.

[34] J. Guo, C. Yang, and I. Chih-Lin, "Exploiting future radio resources with end-to-end prediction by deep learning," *IEEE Access*, vol. 6, pp. 75 729–75 747, 2018.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] H. D. Trinh, L. Giupponi, and P. Dini, "Mobile traffic prediction from raw data using LSTM networks," in *Proc. 29th IEEE Annu. Int. Symp. Pers., Indoor Mob. Radio Commun. (PIMRC)*, 2018, pp. 1827–1832.

[37] A. Azari, F. Salehi, P. Papapetrou, and C. Cavdar, "Energy and resource efficiency by user traffic prediction and classification in cellular networks," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 2, pp. 1082–1095, 2021.

[38] S. Wu, Z. Zeng, and H. Xia, "Coalition-based sleep mode and power allocation for energy efficiency in dense small cell networks," *IET Commun.*, vol. 11, no. 11, pp. 1662–1670, 2017.

[39] J. Kim, W. S. Jeon, and D. G. Jeong, "Effect of base station-sleeping ratio on energy efficiency in densely deployed femtocell networks," *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 641–644, 2015.

[40] Y. Li, H. Celebi, M. Daneshmand, C. Wang, and W. Zhao, "Energy-efficient femtocell networks: Challenges and opportunities," *IEEE Wirel. Commun.*, vol. 20, no. 6, pp. 99–105, 2013.

[41] A. D. Wyner, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1713–1727, 1994.

[42] L. Saker, S.-E. Elayoubi, R. Combes, and T. Chahed, "Optimal control of wake up mechanisms of femtocells in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 664–672, 2012.

[43] G. Wu, G. Feng, and S. Qin, "Cooperative sleep-mode and performance modeling for heterogeneous mobile network," in *Proc. IEEE Wirel. Commun. Netw. Conf. Workshops (WCNCW WKSHPS)*, 2013, pp. 6–11.

[44] J. Wu, E. W. Wong, Y.-C. Chan, and M. Zukerman, "Power consumption and GoS tradeoff in cellular mobile networks with base station sleeping and related performance studies," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 4, pp. 1024–1036, 2020.

[45] J. Wu, Y. Zhang, M. Zukerman, and E. K.-N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 2, pp. 803–826, 2015.

[46] P. Ren and M. Tao, "A decentralized sleep mechanism in heterogeneous cellular networks with QoS constraints," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 509–512, 2014.

[47] J. Kim, W. S. Jeon, and D. G. Jeong, "Base-station sleep management in open-access femtocell networks," *IEEE Trans. on Veh. Technol.*, vol. 65, no. 5, pp. 3786–3791, 2015.
[48] Y. S. Soh, T. Q. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, 2013.
[49] H. Zhang, Y. Wang, H. Ji, and X. Li, "A sleeping mechanism for cache-enabled small cell networks with energy harvesting function," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 2, pp. 497–505, 2020.
[50] A. M. Alqasir and A. E. Kamal, "Cooperative small cell hetnets with dynamic sleeping and energy harvesting," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 3, pp. 774–782, 2020.
[51] X. Li and C. Zhang, "Semi-dynamic Markov approximation-based base station sleep with user association for heterogeneous networks," *IET Commun.*, vol. 17, no. 6, pp. 704–711, 2023.
[52] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of k-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, 2012.
[53] T.-C. Hou and V. Li, "Transmission range control in multihop packet radio networks," *IEEE Trans. Commun.*, vol. 34, no. 1, pp. 38–44, 1986.
[54] B. Matérn, *Spatial variation*. Springer Science & Business Media, 2013, vol. 36.
[55] M. E. Leinonen, M. Jokinen, N. Tervo, O. Kursu, and A. Pärssinen, "Radio interoperability in 5G and 6G multiradio base station," in *Proc. 92nd IEEE Veh. Technol. Conf. (VTC2020-Fall)*, 2020, pp. 1–5.
[56] A. Baddeley, I. Bárány, and R. Schneider, "Spatial point processes and their applications," *Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 13–18, 2004*, pp. 1–75, 2007.
[57] B. Boots, K. Sugihara, S. N. Chiu, and A. Okabe, *Spatial tessellations: concepts and applications of Voronoi diagrams*. John Wiley & Sons, 2009.
[58] B. Sklar, "Rayleigh fading channels in mobile digital communication systems. I. Characterization," *IEEE Commun. Mag.*, vol. 35, no. 7, pp. 90–100, 1997.
[59] F. Rashid-Farrokhi, K. R. Liu, and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1437–1450, 1998.
[60] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Technol. J.*, vol. 27, no. 3, pp. 379–423, 1948.
[61] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wirel. Commun.*, vol. 18, no. 5, pp. 40–49, 2011.
[62] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.
[63] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. on Neural Net.*, vol. 5, no. 2, pp. 157–166, 1994.
[64] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.
[65] J. Hu, W. Heng, G. Zhang, and C. Meng, "Base station sleeping mechanism based on traffic prediction in heterogeneous networks," in *2015 International Telecommunication Networks and Applications Conference (ITNAC)*. IEEE, 2015, pp. 83–87.
[66] L. Li and W. Meng, "Collaborative base station sleeping solution design in heterogeneous cellular network," in *Proc. 27th Asia Pacific Conference on Communications (APCC)*. IEEE, 2022, pp. 231–235.
[67] X. Chen, Y. Jin, S. Qiang, W. Hu, and K. Jiang, "Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale," in *Proc. 25th IEEE Int. Conf. Commun. (ICC)*, 2015, pp. 3585–3591.
[68] MathWorks, "Regression output layer - MATLAB regressionLayer," 2017 (accessed on March 10, 2023). [Online]. Available: https://www.mathworks.com/help/deeplearning/ref/regressionlayer.html
[69] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparison of ARIMA and LSTM in forecasting time series," in *Proc. 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 1394–1401.
[70] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparative analysis of forecasting financial time series using ARIMA, LSTM, and BiLSTM," *arXiv preprint arXiv:1911.09512*, 2019.
[71] Y. Peter T, Y. Li, and G. Pius K, "A comparison between ARIMA, LSTM, and GRU for time series forecasting," in *Proc. 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, 2019, pp. 49–55.

Xinyu Wang received the B.Eng. degree in Electronic Science and Technology from Zhejiang University, Hangzhou, and the M.Sc. degree in Electronic Information Engineering from City University of Hong Kong, in 2016 and 2018 respectively. He is currently a Ph.D. candidate with the Department of Electrical Engineering, City University of Hong Kong, China. His current research interests are path planning and resource allocation in telecommunication networks.

Bingchen Lyu received his B.Eng. degree in Network Engineering from University of Jinan, China, and his M.Sc. degree in Electronic Information Engineering from City University of Hong Kong, in 2019 and 2020, respectively. He is currently engaged in data center infrastructure construction projects in the Internet industry.

Chao Guo received his Ph.D. degree from the Department of Electrical Engineering, City University of Hong Kong, Hong Kong in 2023. He is now a Postdoctoral Fellow at the Centre for Intelligent Multidimensional Data Analyses Limited (CIMDA), Hong Kong Science Park, Hong Kong. His research interests include the areas of data center resource scheduling and network optimization, cable path planning, and wireless networking.

Jiahe Xu received the B.Eng. degree in Electronics Information Science and Technology at the Harbin Institute of Technology and the M.Sc. degree at the City University of Hong Kong, where she is currently pursuing the Ph.D. degree with the Department of Electrical Engineering. Her research interests include Network Slicing.

Moshe Zukerman (M'87–SM'91–F'07–LF'20) received the B.Sc. degree in industrial engineering and management, an M.Sc. degree in operations research from the Technion – Israel Institute of Technology, Haifa, Israel, and a Ph.D. degree in engineering from University of California, Los Angeles, in 1985. He was an independent consultant with the IRI Corporation and a Postdoctoral Fellow with the University of California, Los Angeles, in 1985– 1986. During 1986–1997, he was with Telstra Research Laboratories (TRL), first as a Research Engineer and, during 1988–1997, as a Project Leader. He also taught and supervised graduate students at Monash University in 1990–2001. During 1997-2008, he was with The University of Melbourne, Victoria, Australia. Since December 2008 he has been with the Electronic (now Electrical) Engineering Department of City University of Hong Kong (CityU) as a Chair Professor of Information Engineering and a team leader. Between 2020 - 2022, he also served as the Acting Chief Information Officer of CityU. He has over 300 publications in scientific journals and conference proceedings. He has served on various editorial boards such as Computer Networks, IEEE Communications Magazine, IEEE Journal of Selected Areas in Communications, IEEE/ACM Transactions on Networking and Computer Communications.