Joint Traffic Prediction and Base Station Sleeping for Energy Saving in Cellular Networks

Yuchao Zhu and Shaowei Wang
School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
E-mail: dz20230030@smail.nju.edu.cn, wangsw@nju.edu.cn

ICC 2021 - IEEE International Conference on Communications | 978-1-7281-7122-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICC42927.2021.9500442

***Abstract*—Densely deployed base station (BS) network is one of the important technologies for 5G and beyond mobile communication system, which improves the system throughput by deploying a large number of BSs in the service area. However, such a mobile network has to deal with the consequent power issue since the energy consumption of the BSs generally accounts for a substantial part of the whole system. In this paper, we propose an intelligent BS sleeping scheme to reduce the system energy consumption as much as possible with reasonable signaling overhead while guaranteeing the quality of experience of users. First, we introduce a long short term memory learning method to forecast the traffic distribution in the service area, by which we can determine when the BS sleeping operation is triggered; second, we develop an efficient three-step procedure to determine which of the BSs would sleep or be wakened. Experiment results show that our proposed traffic prediction method works quite well in practical scenarios. The prediction error is not more than 10%, and the energy consumption decreases more than 40% in average for a commercial mobile network with 20 BSs.**

***Index Terms*—Base station sleeping, long short-term memory, traffic prediction.**

I. Introduction

Densely deployed base station (BS) network is a promising architecture in 5G and beyond mobile networks to enhance system capacity [1], as discussed extensively in heterogeneous network (HetNet) and cloud radio access network (C-RAN). The former is a mixture of both macrocells and different kinds of small cells to achieve seamless coverage and high-speed transmissions [2]. The latter enables efficient network deployment with low complexity by separating the baseband signal processing and the radio transceivers with the centralized baseband units and remote radio heads [3]. On one hand, it is a reasonable way to deploy dense network to meet the explosive future traffic demand. On the other hand, the system energy consumption will multiply increase since the BSs in active mode account for nearly 80% of the network energy consumption [4]. Energy efficiency has long been a problem of great interest nowadays and reducing the energy consumption under the condition of ensuring different key performance indicators (KPIs) is urgent from both environmental and economic perspectives [5], [6].

Quality of experience (QoE) of users is one of the most important KPIs for mobile service providers, the prerequisite of which is generally to guarantee the transmission rates of users at any time. As a result, commercial mobile networks are usually designed to meet the peak rates demand of the served users, i.e., the number of BSs deployed in the network depends solely on the statistical peak traffic distribution [7]. On the other hand, the traffic demand distribution for a given mobile network always fluctuates throughout a day due to the population migration [8]. For instance, the crowd in central business district is thick in the daytime, however, few people live there at night. If the BSs are active 24 hours a day in such an area, a large amount of energy would be wasted since the active BSs have been powered even though they serve few users in the night. Naturally, turning off the BSs serving few users is a reasonable way to save energy from the service provider perspective. The potential of BS sleeping for energy-saving in dense coverage areas is validated in [9], which is based on the dataset from an operational network. The BS would be turned off if its traffic load is below a predefined threshold, and numerical results indicate that 17% of the BSs have low traffic 50-99% of the time. In [10], BS sleeping and user handover are jointly considered in the HetNet, where an approximation algorithm is introduced to deal with the user association for a given set of BSs. In [11], the authors propose a distributed BS switching on/off scheme, which tries to sleep BSs one by one that minimally affect the network by introducing a network-impact parameter. In [12], an effective local search BS selection algorithm is proposed to minimize power consumption in the C-RAN. In [13], the BS sleeping scheme is designed in a load-balancing way by introducing a fairness index, where the trade-off between system energy saving and load balance among BSs is achieved in C-RAN. In [14], a joint BS clustering and sleeping strategy is presented, which uses the queue length information to capture the mismatch between the traffic demand and the achievable data rate.

The key point of BS sleeping is to determine not only which BSs should be turned on/off but also when these operations should be triggered. The aforementioned schemes pay little attention to the latter, for which the knowledge of real-time traffic demand is required. To get the information of the fluctuating traffic loads of BSs, accurate and robust traffic prediction plays an important role. In [15], auto-regressive integrated moving average (ARIMA) and its derivatives are used as common methods for time series analysis and prediction. The

This work was partially supported by the National Natural Science Foundation of China (61931023, U1936202, 61801208).

978-1-7281-7122-7/21/$31.00 ©2021 IEEE

investigations in [16]–[18] indicate that, with the development of big data and machine learning (ML), the traffic distribution can be analyzed and predicted extensively.

It is natural to combine traffic prediction with BS sleeping so as to trigger the BS sleeping procedure according to the real time traffic demand. The authors in [19] propose a hybrid traffic prediction model based on linear regression and sort the BSs according to the coverage to make the switching off order more reasonable. Simulation results show that at least 14% of BSs can be turned off without impacting the QoE. Different from [19] which just considers statistical models due to the concern of complexity, neural networks are investigated in [20] for traffic load estimation, based on which BS would be switched off if its load is under a threshold, and its load is migrated to a macro BS. In [21], the authors employ a *K*-means clustering algorithm to divide BSs into different categories to train a model for each cluster, and apply a classification method to predict the states of BSs according to the traffic load. The proposed strategy can predict the idle periods in advance so as to switch off partial BSs while guaranteeing the service quality.

In a nutshell, the key issue discussed above is to match the BS sleeping strategy with fluctuating traffic profiles since frequent handover would deteriorate the QoE of users [11]. Besides the traffic prediction, it is also important to determine the predicting interval as well as switching frequency for practical mobile networks. In this paper, we aim to find an appropriate interval to turn off the BSs with slight loads to minimize the power consumption while satisfying a series of network KPIs, such as bandwidth budget, spectral efficiency requirement and especially, traffic demand. We introduce an intelligent long short-term memory (LSTM) network to deal with the multiple time steps BS traffic prediction problem based on the raw data collecting from signal measuring instrument, where we also consider the influence of prediction interval. With the prediction results, we formulate the BS sleeping problem into a standard convex optimization task and develop a three-step local search algorithm to find the promising solutions. The performance of our proposed scheme is demonstrated by experimental results.

The rest of this paper is organized as follows. In Section II, network model is illustrated, as well as the optimization problem formulation. The LSTM for traffic prediction and the BS sleeping algorithm is given in Section III. In Section IV, experiment results are given with discussions. Finally, we conclude this work in Section V.

II. Network Model And Problem Formulation

A. Traffic Model

We consider a region $\mathcal{D} \in \mathbf{R}^2$ served by a densely deployed mobile network with N BSs. The service area is discretized into K traffic demand areas (TDAs) as shown in Fig. 1, and each TDA contains multiple users with different rate demands. The traffic demand of TDAs can be abstractly equal using a load balancing method proposed in [22]. Denote $\mathcal{N} = \{1, 2, ..., N\}$ and $\mathcal{K} = \{1, 2, ..., K\}$ as the set of BSs and TDAs, respectively.

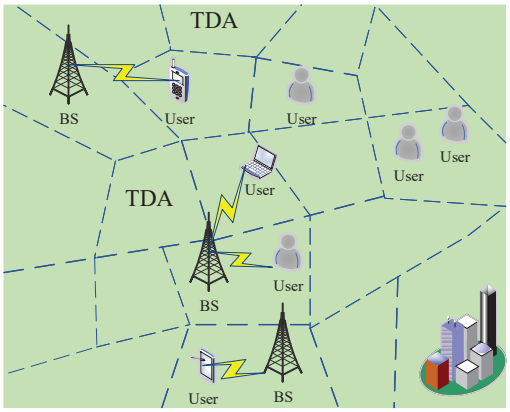


Fig. 1. The network model.

Let $x \in \mathcal{D}$ be a location in the service area, and the density of rate requirement is denoted by $\Phi(x)$. The total traffic requirement in the area is R_D. $\mathcal{D}_k$ and R_k are the service region of TDA k and its traffic demand, respectively. We have

$$\iint_{\mathcal{D}_k} \Phi(x) \mathrm{d}\sigma = R_k. \tag{1}$$

Denote $h_{k,n}$ as the channel gain between BS n and TDA k, $b_{k,n}$ and $p_{k,n}$ are represented as the bandwidth and power that BS n allocated to TDA k, respectively. The transmission rate between BS n and TDA k can be calculated as

$$r_{k,n} = b_{k,n} log_2[1 + \frac{p_{k,n}|h_{k,n}|^2}{b_{k,n}(N_0 + I_{k,n})}], \tag{2}$$

where N_0 is noise power, and $I_{k,n}$ represents the interference introduced by the active BSs with unit bandwidth. Since the inter-cell interference can be eliminated by signal processing techniques, we set $I_{k,n} \approx 0$ for simplifying analysis. So $p_{k,n}$ can be written as:

$$p_{k,n} = \frac{N_0 b_{k,n}}{|h_{k,n}|^2} \cdot (2^{r_{k,n}/b_{k,n}} - 1). \tag{3}$$

Assume that $\mathcal{K}_n$ is the set of TDAs served by BS n, the total rate requirement of $\mathcal{D}$ at time t can be calculated as

$$R_D = \sum_{k \in \mathcal{K}} R_k \quad = \sum_{n \in \mathcal{N}} \sum_{k \in \mathcal{K}_n} r_{k,n} = \sum_{n \in \mathcal{N}} r_n, \tag{4}$$

where r_n is denoted as the traffic load of BS n.

B. Power Model

The total power consumption of the dense network can be written as follows:

$$P_{total} = \sum_{n \in \mathcal{N}} P_f^n + \sum_{n \in \mathcal{N}} P_t^n, \tag{5}$$

where P_f^n and P_t^n is the fixed power and transmission power of BS n, respectively. Also note that $P_t^n = \sum_{k \in \mathcal{K}_n} p_{k,n}$.

Binary variable x_n represents the state of BS $n \in \mathcal{N}$:

$$x_n = \begin{cases} 1, & \text{BS } n \text{ is active,} \\ 0, & \text{BS } n \text{ is inactive.} \end{cases} \tag{6}$$

Note that, P_f^n is influenced by the state of BS, but can be ignored when BS is inactive, i.e., although the BS still consumes powers to guarantee the wake up from sleep mode, it is negligible as compared with the power when BS is active [4]. Then the total power consumption can be transformed into:

$$P_{total} = \sum_{n\in\mathcal{N}} P_f^n x_n + \sum_{n\in\mathcal{N}} \frac{x_n}{\eta_n} \sum_{k\in\mathcal{K}} p_{k,n}, \tag{7}$$

where η_n is the power amplifier efficiency factor.

C. Problem Formulation

Our goal is to select a series of BSs to minimize the total power consumption of the network under practical constraints. Define p_n^{max} and b_n^{max} as the maximum transmission power and the available bandwidth for BS $n \in \mathcal{N}$, respectively. To simplify the notations, we collect the variables x_n's, $b_{k,n}$'s and $p_{k,n}$'s into vectors $\vec{x}$, $\vec{b}$, and $\vec{p}$, respectively. Define $\mathbf{X} = \{\vec{x} | x_n \in \{0,1\}\}$, the optimization problem can be mathematically formulated as follows:

$$\begin{aligned} \underset{\vec{x},\vec{b},\vec{p}}{\text{minimize}} \quad & P_{total} \\ s.t. \quad & C_1: \textstyle\sum_{k\in\mathcal{K}} p_{k,n} \le x_n p_n^{max}, \forall n \in \mathcal{N}, \\ & C_2: \textstyle\sum_{k\in\mathcal{K}} b_{k,n} \le x_n b_n^{max}, \forall n \in \mathcal{N}, \\ & C_3: \textstyle\sum_{n\in\mathcal{N}} r_{k,n} \ge R_k, \forall k \in \mathcal{K}, \\ & C_4: p_{k,n} \ge \triangle_{k,n} b_{k,n}, \forall k \in \mathcal{K}, \forall n \in \mathcal{N}, \\ & C_5: \vec{x} \in \mathbf{X}, \quad \vec{b} \in \mathbb{R}_+^{KN}, \quad \vec{p} \in \mathbb{R}_+^{KN}. \end{aligned} \tag{8}$$

In (8), C_1 and C_2 are the maximum transmission power constrains and available bandwidth budgets for the selected BSs, C_3 ensures the traffic demand of TDAs, C_5 is intuitive. C_4 is the spectral efficiency requirement of TDA k, which is defined as $S_{k,n} = \frac{r_{k,n}}{b_{k,n}} \ge S_k^{min}$, and can be rewritten as follows according to (2):

$$log_2(1 + \frac{p_{k,n}|h_{k,n}|^2}{b_{k,n}N_0}) \ge S_k^{min}. \tag{9}$$

Let $\triangle_{k,n} = (2^{S_k^{min}} - 1)N_0/|h_{k,n}|^2$, we can get C_4.

The traffic demand distribution of the service area usually keeps changing throughout a day due to the population migration. Solving the optimization problem define by (8) requires knowledge of the future rate requirement on each TDA (i.e., R_k), which needs to be predicted. According to equation (4), the problem of spatial and temporal flow distribution prediction can be transformed into a time series prediction problem for a single BS. We can use time series prediction methods, such as ARIMA and LSTM, to handle the problem after obtaining the historical traffic data of BSs.

Assume that the collected data set is $\mathcal{R}_{BS} = \{R_{B_1}, R_{B_2}, ..., R_{B_N}\}$, where $R_{B_n} = \{r_n^1, r_n^2, ..., r_n^t\}$ $(n \in \{1, 2, ...N\})$ stores the traffic data of BS n with granularity δ (in hour, $0 < \delta < 1$). The traffic load for the next δ interval could be predicted when applying one-step prediction. Multi-step prediction is required when it comes to BS sleeping, since the granularity of traffic data is generally much smaller than the allowable interval for switching BS on/off. When the next z time steps need to be predicted, the prediction interval is denoted as $T_{pred} = \delta z$, and the total traffic demand of the region in duration $[t, t + T_{pred}]$ is given by:

$$R_D = \sum_{n\in\mathcal{N}} \frac{\sum_{j=t+1}^{j=t+z} r_n^j}{z}, \tag{10}$$

here we take the average of the prediction values as the rate demand at a given time interval to achieve a trade-off between switch frequency and real-time traffic demand.

III. Our Proposed Algorithm

Equation (8) is a mixed integer programming problem, which is NP-hard in general. We use LSTM network to handle the traffic prediction problem at first as well as offer a reference for choosing an appropriate time interval for BS sleeping, and then a local search algorithm is designed to deal with (8). The BS sleeping procedure is triggered based on the traffic prediction result at a given time interval.

A. Cellular traffic prediction

Recurrent neural network (RNN) is commonly used for time series forecasting. RNNs are networks with loops in them, allowing information to persist, thus predicting future information with the knowledge of previous data. LSTM network is a special type of RNNs, where cell states and different gates are introduced to avoid vanishing-gradient problem in RNNs. As depicted in Fig. 2, a standard LSTM unit has three gates interacting with each other to learn long-term dependencies, namely, input gate, forget gate, and output gate. The output of

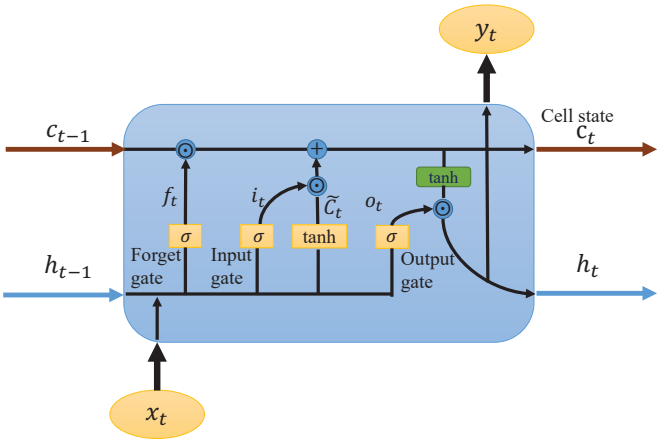


Fig. 2. Standard LSTM hidden unit.

these gates can be written as follows:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t \odot tanh(C_t). \end{aligned} \tag{11}$$

In (11), $\sigma(\cdot)$ is sigmoid function, $\odot$ stands for Hadamard product, W and b are the weight matrix and the bias of the gates, respectively.

As mentioned before, we obtained the T-hours traffic data of BS $n \in \mathcal{N}$ in a service area $\mathcal{D}$ from the signal measuring instrument, it is denoted as R_{B_n}. The number of historical data obtained is $m = T/\delta$ for each BS $n \in \mathcal{N}$, and the traffic load data of all the BSs is stored in set $\mathcal{R}_{BS}$. For BS n,

Algorithm 1 Traffic prediction algorithm by LSTM

1: Initializations: $\mathcal{N}$, $\mathcal{R}_{BS}$, δ, T, m, z $(0<z<m/2, z \in \mathbb{Z})$, $\mathcal{P}_{BS} = \{P_{B_1}, P_{B_2}, ...P_{B_N}\}, P_{B_i} = 0$ $(i \in \{1, 2, ...N\})$.
2: **for** $n = 1 : N$; **do**
3: $T_x \leftarrow \{r_n^1, ...r_n^{m-z}\}$, $T_y \leftarrow \{r_n^2, ..., r_n^{m-z+1}\}$;
4: $V_x \leftarrow \{r_n^{m-z+1}, ..., r_n^{m-1}\}$, $V_y \leftarrow \{r_n^{m-z+2}, ..., r_n^m\}$;
5: INPUT:
Training set $\mathcal{T}_n \leftarrow (T_x, T_y)$,
Validation set $\mathcal{V}_n \leftarrow (V_x, V_y)$.
6: OUTPUT:
$P_{B_n} = \{r_n^{m+1}, r_n^{m+2}, ..., r_n^{m+z}\}$,
Calculate $NRMSE$ using (12).
7: **end for**
8: **return** R_D using (10).

we divide its data set into training set and validation set to train the LSTM network, the prediction procedure is given in Algorithm 1. Here we choose the normalized root mean square error (NRMSE) as the metric of accuracy, which is defined as:

$$NRMSE = \frac{1}{y_{max} - y_{min}} \sqrt{\frac{1}{z} \sum_{i=1}^{z} \left| y_i - y_i^{'} \right|^2}, \tag{12}$$

where z is the number of predicted time steps, y_i and $y_i^{'}$ are the observation value and prediction value at the time i, respectively. y_{max} and y_{min} are the maximum and minimum of y, respectively.

B. Base station sleeping

We need to deal with the power and bandwidth allocation problem at first, where a capacity margin for emergency situations is taken into consideration. That is, given a subset of BSs, the rate requirement of TDA k, which is defined as $R_k^{'}$ (e.g., $R_k^{'} = 1.1R_k$), needs to be satisfied. It can be mathematically formulated as follows:

$$\begin{aligned} \text{find} \quad & \vec{b}, \vec{p} \\ s.t. \quad & \textstyle\sum_{n \in \mathcal{N}} r_{k,n} = R_k^{'}, \forall k \in \mathcal{K}, \\ & C_1, C_2, C_4 \ in \ (8). \end{aligned} \tag{13}$$

If a feasible solution to (13) exists, we can claim that the selected BS set $\mathcal{N}_s = \{n | x_n = 1\}$ can meet the TDAs' rate requirements with the power and bandwidth budgets. (13) can be solved by a standard convex optimization algorithm [12].

Equation (8) has the similar form with (13), and it can also be solved in the same way. Define $P(\mathcal{N}_s)$ as the optimal solution to (8), we propose a three-step local search algorithm to find the subset of selected BSs. Starting with a feasible solution, such as $\mathcal{N}_s = \mathcal{N}$, we search the optimal subset of BSs as follows:

Open: Activate BS $n \notin \mathcal{N}_s$, update the system power consumption, if $P(\mathcal{N}_s \cup \{n\}) < P(\mathcal{N}_s)$, add BS n to $\mathcal{N}_s$: $\mathcal{N}_s \leftarrow \mathcal{N}_s \cup \{n\}$.

Close: Switch BS $n \in \mathcal{N}_s$ to sleep mode, update the system power consumption, if $P(\mathcal{N}_s \cup \{n\}) < P(\mathcal{N}_s)$, remove BS n from $\mathcal{N}_s$: $\mathcal{N}_s \leftarrow \mathcal{N}_s \setminus \{n\}$.

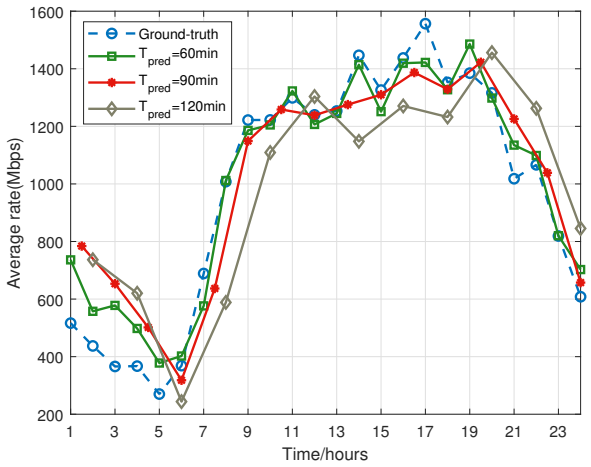
Exchange: Activate BS $n^{'} \notin \mathcal{N}_s$ and turn off BS $n \in \mathcal{N}_s$ at the same time, update the system power consumption, if $P(\mathcal{N}_s \setminus \{n\} \cup \{n^{'}\}) < P(\mathcal{N}_s)$, make the exchange: $\mathcal{N}_s \leftarrow \mathcal{N}_s \setminus \{n\} \cup \{n^{'}\}$.

IV. Experiment Results

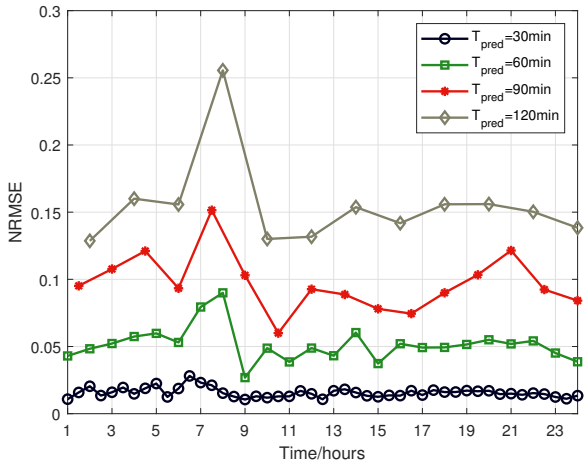
We consider an area served by a commercial mobile network, where 20 BSs are uniformly distributed. The number of TDAs is $K = 100$ and the required spectral efficiency of each TDA is set as 0.1 bit/Hz. For each BS, the maximum transmission power and available bandwidth are 1 W and 100 MHz, respectively. The power consumption P_f^n is 3.1 W. The efficiency of radio frequency power amplifier η_n is 25%. The path loss (in dB) between BS and TDA can be calculated as $140.7 + 36.7\log_{10}(D)$, where D (in km) is the distance between the BS and the center of TDA. The standard deviation of lognormal shadowing is 10 dB and the noise PSD is -184 dBm/Hz.

The LSTM neural network has two hidden layers considering the trade-off between computational cost and complexity [23], and the hidden units of each layer are 128. The initial learning rate is set as 0.005. The maximum number of iterations is 200 and the learning rate after 150 epochs is shrunk by multiplying a factor of 0.2 to avoid divergence. Adam optimization is used to update the network weights [24]. We use the collected 24-hours mobile traffic data from BSs in service area to predict the future rate requirement of the region in a given time interval. The granularity of historical data is $\delta = 10$ min. The training sets and the test sets are divided according to the required prediction interval.

First, we evaluate the performance of the LSTM network for multi-step prediction with different prediction interval T_{pred}. As can be seen in Fig. 3, the longer the prediction interval is, the lower the prediction accuracy will be. As the average of predicted values is taken to represent the rate requirement for period $[t, t + T_{pred}]$ in the service region, the data will be too sparse to depict the trend of the traffic if the predicted time steps are too long. On the other hand, the precision will decrease due to the accumulative error of the LSTM network. Fig. 3(b) shows more clearly that the



(a) Prediction of 24 hours.



(b) Average NRMSE of prediction.

Fig. 3. Traffic prediction and NRMSE throughout a day with different T_{pred}.

accuracy is roughly inversely proportional to the prediction interval. Taking signaling overhead into consideration, we can make a conclusion that the appropriate prediction interval is $T_{pred} = 60$ min under our scenario, since it can achieve a balance between prediction error and switching frequency.

The performance of ARIMA and LSTM in the case of $T_{pred} = 60$ min is shown in Fig. 4. ARIMA model is a widely used time series prediction method based on statistic models, which is defined by three parameters: the auto-regressive term, the differentiation term and the moving average term. In this work, they are set as 3, 1, 3, respectively. The performance of the two methods is similar on the whole since the errors are weakened by the operation of taking the average of predicted values as the hourly traffic demand. The accuracy of LSTM outperforms ARIMA about 2%. However, since the ARIMA model requires that the data is difference-stationary, LSTM is more suitable for cellular traffic prediction in practice.

Finally, we evaluate the power saving throughout a day

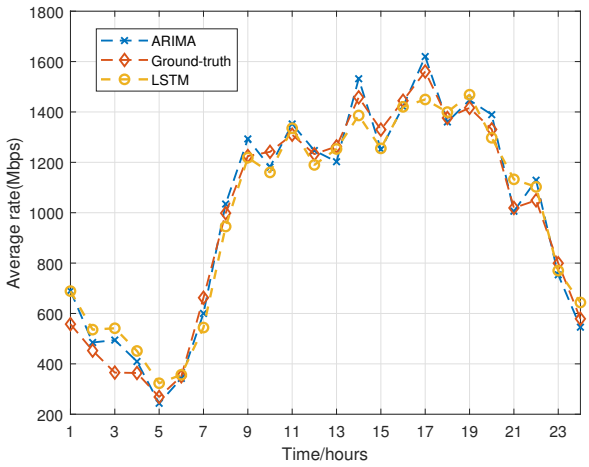


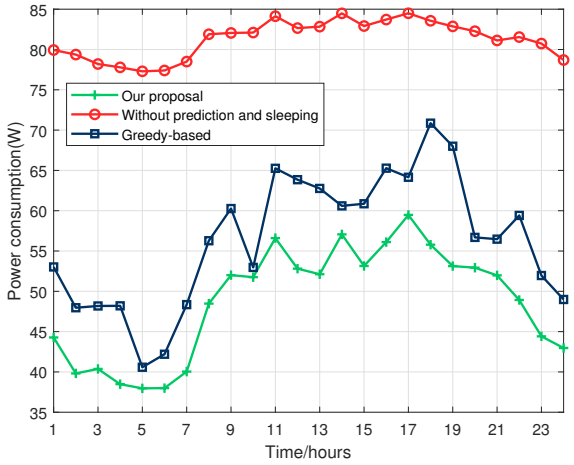
Fig. 4. Hourly prediction result, $T_{pred} = 60$ min.

with our proposal. The BS sleeping procedure is triggered at the beginning of every hour. The traffic demand margin is set as 10% since the prediction error is generally within 10%. We compare our proposed traffic prediction based BS sleeping scheme with the following ones: greedy-based BS sleeping [25] and no BS sleeping. For the former, all the BSs are active at first, then the BS that could yield the largest power saving is turned off in each local search based on the traffic prediction result. For the latter, which is used as a benchmark, all the BSs are always active and the bandwidth and power allocation procedure is performed based on the algorithm discussed in [12] according to the real-time traffic distribution in every hour.

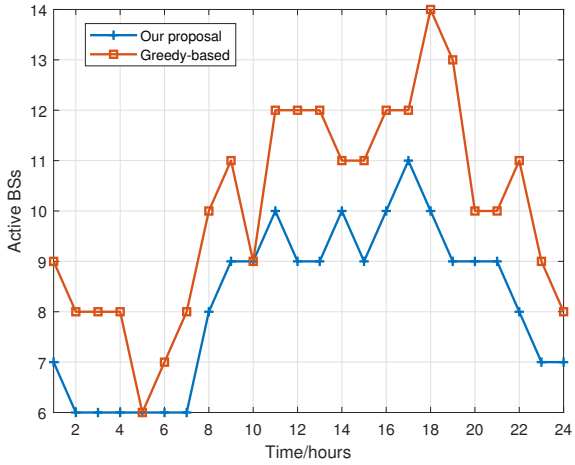
As seen from Fig. 5(a), the power consumption varies with traffic demand. Both of the BS sleeping mechanisms verify that the potential of power saving is significant, especially when the load is relatively low (e.g., 1:00-7:00). Compared with no traffic prediction and BS sleeping, about 40.2% and 30.8% power consumption can be saved on average by our scheme and greedy-based scheme, respectively. Specifically, up to 50% power can be saved at 6:00 and at least 30% power saving can be achieved at 17:00 by our proposal. Moreover, Fig. 5(a) suggests that our proposal outperforms the greedy-based method, especially when the traffic demand is relatively high. For example, the gap is larger than 21% at 19:00, which validates that our proposal can reduce the power consumption efficiently while meeting the varying rate requirements. The number of active BSs throughout a day is given in Fig. 5(b). Our proposed method needs fewer active BSs to meet the traffic demand than greedy-based selection scheme at most of the time. It only needs to keep 6 and 11 BSs active during the lowest and highest load hours, respectively, as can be found in Fig. 5(b).

V. Conclusions

In this paper, we studied the traffic loads based BS sleeping problem in dense network, where we took prediction interval



(a) Total power consumption.



(b) Number of active BSs.

Fig. 5. Total power consumption and active number of BSs throughout a day.

and switching frequency into account. We transformed the complex spatial-temporal traffic prediction into time series prediction for BSs. The problem was then solved by an effective LSTM prediction network and an "open/close/exchange" BS selection algorithm, where a bandwidth and power allocation problem was handled to guarantee the feasibility. Experiment results reveal that the LSTM network can predict the traffic for multiple time steps ahead with high accuracy, based on which we can further implement BS sleeping with a trade-off between signaling overhead and real-time traffic demand. The results show that our proposed joint traffic prediction and BS sleeping scheme can save up to 50% energy when the traffic demand is relatively low throughout a day, which is a promising solution to green network.

REFERENCES

[1] X. You *et al.*, "AI for 5G: research directions and paradigms," *Sci. China Inf. Sci.*, vol. 62, no. 2, pp. 1–13, Oct. 2019.

[2] X. Sun and S. Wang, "Resource allocation scheme for energy saving in heterogeneous networks," *IEEE Trans. Wirel. Commun*, vol. 14, no. 8, pp. 4407–4416, Aug. 2015.

[3] J. Liu *et al.*, "Graph-based framework for flexible baseband function splitting and placement in C-RAN," in *Proc. IEEE ICC'15*, London, UK, Jun. 2015.

[4] C. Liu, B. Natarajan, and H. Xia, "Small cell base station sleep strategies for energy efficiency," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1652–1661, Mar. 2016.

[5] S. Buzzi *et al.*, "A survey of energy-efficient techniques for 5G networks and challenges ahead," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 697–709, Apr. 2016.

[6] Q. Shen, Z. Ma, and S. Wang, "Deploying C-RAN in cellular radio networks: An efficient way to meet future traffic demands," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7887–7891, Aug. 2018.

[7] Y. S. Soh *et al.*, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.

[8] F. Xu *et al.*, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Trans. Netw.*, vol. 25, no. 2, pp. 1147–1161, Apr. 2017.

[9] S. Samulevičius *et al.*, "Energy savings in mobile broadband network based on load predictions: Opportunities and potentials," in *Proc. IEEE VTC Spring'12*, Yokohama, Japan, May 2012.

[10] X. Lin and S. Wang, "Joint user association and base station switching on/off for green heterogeneous cellular networks," in *Proc. IEEE ICC'17*, Paris, France, May 2017.

[11] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. Wirel. Commun.*, vol. 12, no. 5, pp. 2126–2136, May 2013.

[12] W. Zhao and S. Wang, "Traffic density-based RRH selection for power saving in C-RAN," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3157–3167, Dec. 2016.

[13] X. Lin and S. Wang, "Efficient remote radio head switching scheme in cloud radio access network: A load balancing perspective," in *Proc. IEEE INFOCOM'17*, Atlanta, GA, May 2017.

[14] J. Kim, H. W. Lee, and S. Chong, "Traffic-aware energy-saving base station sleeping and clustering in cooperative networks," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 2, pp. 1173–1186, Feb. 2018.

[15] J. Wang, "A process level network traffic prediction algorithm based on ARIMA model in smart substation," in *Proc. IEEE ICSPCC'13*, KunMing, China, Aug. 2013.

[16] F. Xu *et al.*, "Big data driven mobile traffic understanding and forecasting: A time series approach," *IEEE Trans. Serv. Comput.*, vol. 9, no. 5, pp. 796–805, Sept.-Oct. 2016.

[17] J. Wang *et al.*, "Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach," in *Proc. IEEE INFOCOM'17*, Atlanta, GA, May 2017.

[18] A. Azari *et al.*, "Cellular traffic prediction and classification: A comparative evaluation of lstm and arima," in *Proc. Springer DS'19*, Split, Croatia, Oct. 2019, pp. 129–144.

[19] S. Dawoud *et al.*, "Optimizing the power consumption of mobile networks based on traffic prediction," in *Proc. IEEE COMPSAC'14*, Vasteras, Sweden, Jul. 2014.

[20] I. Donevski, G. Vallero, and M. A. Marsan, "Neural networks for cellular base station switching," in *Proc. IEEE INFOCOM WKSHPS'19*, Paris, France, Apr. 2019.

[21] D. Sesto-Castilla *et al.*, "Use of machine learning for energy efficiency in present and future mobile networks," in *Proc. IEEE WCNC'19*, Marrakesh, Morocco, Apr. 2019.

[22] C. Ran, S. Wang, and C. Wang, "Cellular networks planning: A workload balancing perspective," *Comput. Netw.*, vol. 84, pp. 64–75, Jun. 2015.

[23] G. Vallero *et al.*, "Greener RAN operation through machine learning," *IEEE Trans. Netw. Serv. Manag.*, vol. 16, no. 3, pp. 896–908, Sept. 2019.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green Cloud-RAN," *IEEE Trans. Wirel. Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.