Advanced Sleep Modes with Context-Aware Cellular Traffic Prediction for Network Energy Saving

Wankyu Choe*, Hyunwoo Park*, Minsoo Jeong†, and Sunwoo Kim*
*Department of Electronic Engineering, Hanyang University
†Intelligent RAN Research Section, Electronics and Telecommunications Research Institute (ETRI)
{choewg, stark95, remero}@hanyang.ac.kr*, qwjms@etri.re.kr†

Abstract

This paper proposes a network energy-saving method that leverages advanced sleep modes (ASMs) in conjunction with context-aware cellular traffic prediction. Cellular traffic is predicted using a convolutional long short-term memory network with cross-domain datasets to capture complex spatiotemporal patterns. Thereafter, the optimal ASM parameters are selected by exploiting the predicted traffic and applied to a base station sleeping strategy. Simulation results demonstrate the effectiveness of the proposed method toward sustainable radio access network operation.

***Index Terms*—Network energy saving, advanced sleep modes, cellular traffic prediction**

I. Introduction

In fifth-generation (5G) networks, as cellular traffic demand continues to grow and networks become increasingly dense, energy-saving under quality of service (QoS) constraints is essential for energy-efficient radio access network (RAN) operation [1]. Prior work [2] analyzed the efficacy of leveraging advanced sleep modes (ASMs) to reduce base station (BS) energy consumption. This paper proposes an ASM-based energy-saving method driven by context-aware cellular traffic prediction to enable adaptive operation that accounts for spatiotemporal dependencies. Simulation results compare the proposed method with the *Always On* baseline to quantify the energy-saving gain.

II. System Model

Spatiotemporal traffic is represented on an $H \times W$ grid, where the traffic snapshot at time t is denoted by $\mathbf{D}_t \in \mathbb{R}^{H \times W}$. The traffic is predicted from three inputs, including a sequence of historical snapshots $\{\mathbf{D}_{t-1}, \ldots, \mathbf{D}_{t-p}\}$ capturing spatiotemporal patterns, static context features modeling external factors affecting traffic demand, and a vector of temporal characteristics encoding periodic patterns. The historical snapshots are processed by a convolutional long short-term memory (ConvLSTM) module to jointly learn temporal dynamics and spatial correlations, while the context features are encoded by a lightweight convolutional neural network (CNN) and the temporal characteristics are embedded by a feedforward network. The resulting feature maps are concatenated and passed through a densely connected convolutional block to yield the predicted traffic $\hat{\mathbf{D}}_t$ over the entire grid. The overall architecture of the context-aware traffic prediction is illustrated in Fig. 1.

The BS power consumption is characterized using the 3GPP energy model together with [3] for a more realistic parameterization. The 3GPP energy model provides BS power consumption for ASMs comprising micro sleep, light sleep, and deep sleep, as well as additional transition energy and

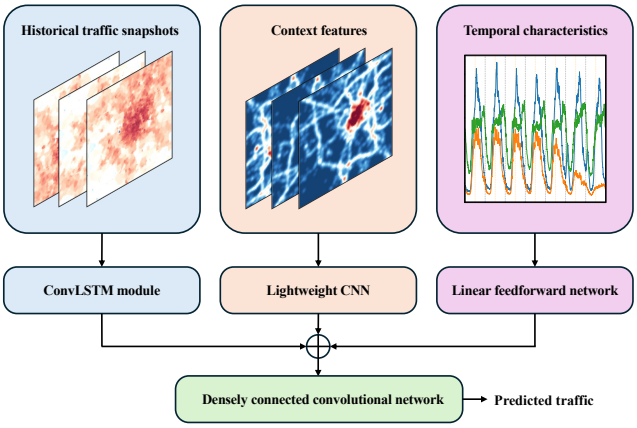

Fig. 1. A framework for the context-aware cellular traffic prediction.

TABLE I
BS Power Model

Power state	Relative power	Additional transition energy	Transition time
Deep sleep	33	1000	50 ms
Light sleep	80	90	6 ms
Micro sleep	98	0	0
Active downlink	$145 + 135 * s$	N/A	N/A

transition time associated with activating each sleep mode. In particular, the measurement-driven and analytically tractable energy model in [3] to reflect the power consumption of practical 5G active antenna units. The reference configuration for the BS power model is considered as *Set 1 FR1*, and the corresponding power model used in this paper is summarized in Table I. The relative power values are dimensionless and normalized to the deep sleep power, the additional transition energy is reported for 1 ms reference period, and $s \in [0, 1]$ denotes the physical resource block (PRB) utilization ratio.

III. Proposed Energy Saving Method

Given a target BS load level quantified by the PRB utilization ratio, the proposed method determines the ASM

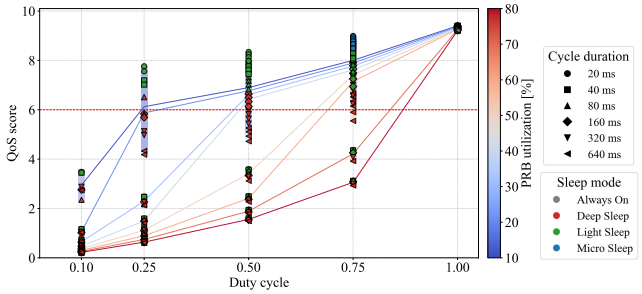

Fig. 2. QoS score of ASM configuration tuples under different load levels.

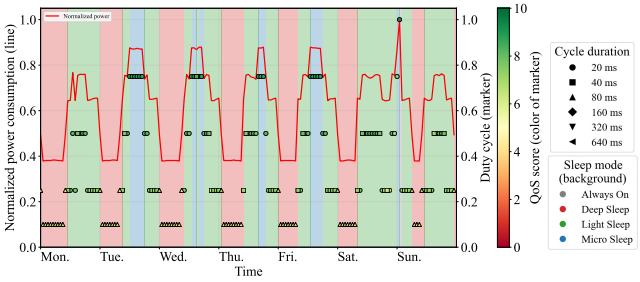

Fig. 3. ASM decisions for the grid $(50, 60)$, Milan's Duomo, and the resulting normalized BS power consumption.

configuration via an offline exhaustive search over a finite candidate set of cell discontinuous transmission (cell DTX) parameters. The search space is defined by the cycle duration $T \in \{20, 40, 80, 160, 320, 640\}$ ms, the duty cycle $\delta \in \{0.1, 0.25, 0.5, 0.75, 1\}$, and the sleep mode $m \in \{\text{micro}, \text{light}, \text{deep}\}$. For each tuple (T, δ, m), the power consumption is computed using the power model in Table I while satisfying mode transition feasibility constraints, which require the OFF duration to be long enough to enter the selected sleep mode. To achieve energy saving subject to a QoS constraint, a latency-based QoS score is employed to enforce the average latency threshold of 50 ms. In addition, the QoS score accounts for transmission reliability via the successful transmission ratio. As shown in Fig. 2, only tuples satisfying the QoS constraint are retained as feasible candidates in the exhaustive search. Among all feasible tuples, the optimal ASM parameters are selected to minimize the average BS power, yielding the resulting load-to-ASM lookup table, which is later applied online using predicted traffic.

The predicted traffic $\hat{\mathbf{D}}_t \in \mathbb{R}^{H \times W}$ is given as next-step traffic volume over all grids, whereas the lookup table is indexed by PRB utilization ratio. Thus, the predicted traffic is converted to a normalized load indicator by applying per-grid min–max normalization

$$\hat{\mathbf{S}}_t(h, w) = \frac{\hat{\mathbf{D}}_t(h, w) - D_{\min}(h, w)}{D_{\max}(h, w) - D_{\min}(h, w)}, \quad \forall (h, w) \tag{1}$$

where $D_{\min}(h, w)$ and $D_{\max}(h, w)$ denote the minimum and maximum traffic volumes observed at grid (h, w). The normalized load $\hat{\mathbf{S}}_t \in [0, 1]^{H \times W}$ is then mapped to a PRB utilization estimate $\hat{s}_t(h, w) \in [0, 1]$ via the identity mapping, after which the ASM parameters are obtained by indexing the lookup table with $\hat{s}_t(h, w)$.

IV. Simulation Results

For the traffic prediction, the Milan dataset [4], where the city is partitioned into $H \times W = 100 \times 100$ grids and traffic is collected over 62 days. The first seven weeks are used for training and the last week for testing, with a sliding-window input of length $p = 3$. The model is trained using Adam with a batch size of 32 for 300 epochs, starting from a learning rate of 0.01 and applying step-down scheduling at 50% and 75% of the total epochs. To evaluate the energy-saving gain under protocol dynamics, slot-level sleeping is implemented on a 5G system-level simulator. Traffic is generated by a user datagram protocol client by sweeping the packet inter-arrival time to match the desired PRB utilization.

Fig. 3 shows the ASM decisions for a representative grid during the test period, where the selected parameters vary over time according to the predicted PRB utilization. Low-load periods enable lower duty cycles and deeper sleep states, whereas high-load periods lead to selecting *Always On* to satisfy the QoS constraint. For the system-level evaluation, the proposed method is applied to the central 20×20 grids, and the energy-saving gain is computed compared with the *Always On* baseline, yielding an average energy-saving gain of 31.4%.

V. Conclusion

This paper proposed an energy-saving method that leverages ASMs with context-aware cellular traffic prediction. The proposed method predicts next-step traffic over spatial grids, maps the predicted traffic volume to a PRB utilization estimate, and indexes a load-to-ASM lookup table obtained via exhaustive search to select ASM parameters subject to a latency-based QoS constraint. Simulation results demonstrate that the proposed method achieves a notable energy-saving gain while satisfying the QoS constraint, highlighting its potential to support more sustainable RAN operation. Future work includes developing an open RAN-compliant approach using the RAN intelligent controller for near-real-time traffic prediction and ASM control.

Acknowledgment

This work was supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2024-00396828, Development of AI based low power 5G-A O-DU/O-CU).

References

[1] D. López-Pérez *et al.*, "A survey on 5G radio access network energy efficiency: Massive MIMO, lean carrier design, sleep modes, and machine learning," *IEEE Commun. Surv. Tut.*, vol. 24, no. 1, pp. 653–697, 2022.

[2] J. G. Borja, P. Bruhn, and M. Petrova, "Towards 6G: Leveraging advanced sleep modes to improve energy performance," in *2024 3rd Int. Conf. 6G Netw. (6GNet)*. IEEE, 2024, pp. 112–116.

[3] N. Piovesan *et al.*, "Machine learning and analytical power consumption models for 5G base stations," *IEEE Commun. Mag.*, vol. 60, no. 10, pp. 56–62, 2022.

[4] G. Barlacchi *et al.*, "A multi-source dataset of urban life in the city of Milan and the Province of Trentino," *Sci. Data*, vol. 2, no. 1, pp. 1–15, 2015.