

Throughput-Aware Hierarchical Energy Saving in Cellular Networks via Joint MIMO Reconfiguration and Capacity-Cell Sleep

Ahmed AlAlwani ^{*||}, Mohammad Galal Khafagy [§], Karim Banawan ^{*‡},
Mariam Aboelwafa [†], Ayman Gaber [§], Karim Seddik ^{*}

^{*} School of Sciences and Engineering, The American University in Cairo (AUC), New Cairo, Egypt.

[§] Mobile Network Development, Technology Department, Vodafone Egypt, Smart Village, Giza, Egypt.

^{||} Network Strategy & Engineering, Vodafone Intelligent Solutions (NSE VOIS), Egypt.

[†] Computer, Communications and Autonomous Systems Engineering, NewGiza University, Egypt.

[‡] Electrical Engineering Department, Faculty of Engineering, Alexandria University, Egypt.

E-mail: {aalalwani, karim.banawan, kseddik}@aucegypt.edu,

{mohamed.khafagy, ayman.gaber}@vodafone.com, mariam.nabil@ngu.edu.eg

Abstract—Energy consumption is a major operational challenge in cellular networks, especially with the densification of capacity layers to support growing traffic demand. This paper proposes a traffic-aware hierarchical energy-saving framework for a sectorized two-layer cellular network, where capacity cells progressively select among full sleep mode, reduced 1×1 or 2×2 multiple-input multiple-output (MIMO) operation, and full 4×4 operation according to load and throughput conditions, while coverage cells remain active to preserve service continuity. To evaluate the proposed framework under realistic conditions, a system-level simulation model is developed using an anonymized base stations (BSs) deployment from a real operator network in Egypt, incorporating realistic sector orientations, user activity periods, traffic demand, interference-coupled SINR calculation, MIMO rank adaptation, modulation and coding scheme (MCS) selection, and physical resource block (PRB) allocation. Cell-level energy consumption is estimated using XGBoost regression models trained on real-network measurements. The threshold-based control policy is tuned via Bayesian optimization and benchmarked against full-configuration, MIMO-only, sleep-only, and static-threshold operation. Simulation results show that the optimized joint MIMO-and-sleep policy achieves approximately 15.3% energy saving, compared to about 9.9% using static thresholds under the same quality of service (QoS) requirement.

Index Terms—cellular networks, energy saving, cell sleep, MIMO, green mobile networks

I. INTRODUCTION

The rapid growth of mobile data traffic has made energy efficiency a critical design objective for modern cellular networks. As mobile networks evolve toward dense, multi-band, and high-capacity deployments, operators are required to support increasing traffic demand while controlling operational costs and reducing environmental impact. Mobile networks contribute a non-negligible share of global energy consumption and carbon emissions, with the radio access network (RAN) accounting for a dominant portion of the total mobile-network energy usage [1], [2]. Therefore, improving RAN energy efficiency has become an important research direction for both academia and industry [3].

Several energy-saving techniques have been studied, including base-station sleep modes, user association, load bal-

ancing, transmit-power control, and learning-based optimization [4]–[8]. Sleep-mode mechanisms reduce energy consumption by deactivating lightly loaded cells, while user association and load-balancing approaches distribute users across available cells or frequency bands to improve resource utilization and avoid congestion [9]–[11]. More recently, machine learning and reinforcement learning have been introduced for energy-efficient network optimization, as they can capture complex relationships between traffic load, radio conditions, user experience, and energy consumption [12]. These learning-based methods have been applied to traffic-adaptive sleep-mode management, cell activation, power allocation, and multiple-input multiple-output (MIMO)-specific energy-saving decisions [7], [13]–[15].

Despite these advances, existing solutions often consider energy-saving mechanisms separately. Sleep-mode approaches typically focus on whether a capacity cell should be active or inactive, without considering intermediate MIMO configurations. Conversely, MIMO-based methods usually adapt spatial transmission features without jointly considering complete capacity-cell shutdown. Moreover, the interaction between MIMO rank adaptation, interference variation, physical resource block (PRB) allocation, and cell sleep remains insufficiently modeled. Therefore, the joint impact of MIMO reconfiguration and capacity-cell shutdown on energy consumption and user experience requires further investigation.

In this work, the energy-saving problem is formulated as the selection of the capacity-layer operating state over time, where each capacity cell can remain active with a selected MIMO configuration or enter sleep mode, while the coverage layer remains active for service continuity. The objective is to minimize the total network energy consumption while maintaining acceptable user throughput. The main contributions of this paper are summarized as follows:

- We develop a realistic dynamic simulator for a sectorized two-layer cellular network. The simulator is designed to closely reflect practical network operation by using an

anonymized real operator base stations (BSs) deployment from Egypt, realistic sector orientations, user association rules, traffic demand, and user activity periods.

- We model joint capacity-cell shutdown and MIMO re-configuration, where capacity cells can switch among $(4 \times 4, 2 \times 2, \text{ and } 1 \times 1)$ operation or enter full sleep mode, while coverage cells remain active for service continuity.
- We design a hierarchical threshold-based control policy and tune its decision thresholds using Bayesian optimization to progressively select the most energy-efficient feasible capacity-cell state based on coverage-layer load, capacity-layer load, and throughput constraints.
- We integrate XGBoost-based energy-consumption models trained on real-network data to estimate cell-level energy consumption from traffic volume, resource utilization, rank statistics, activity time, and sleep duration.
- We evaluate the proposed joint MIMO-and-sleep policy against baseline full-configuration, MIMO-only, sleep-only, and static-threshold operation. The proposed policy achieves approximately 15.3% energy saving, compared to about 9.9% using static thresholds under the same quality of service (QoS) requirement.

The remainder of this paper is organized as follows. Section II presents the system model. Section III formulates the energy-saving optimization problem and describes the proposed dynamic MIMO and shutdown control framework. Section IV presents the simulation setup and performance evaluation. Section V concludes the paper.

II. SYSTEM MODEL

A. Network Architecture

We consider the downlink (DL) of a sectorized cellular network consisting of a set of BSs $\mathcal{B} = \{1, \dots, B\}$ and a set of user equipments (UEs) $\mathcal{U} = \{1, \dots, U\}$. Each BS $b \in \mathcal{B}$ is composed of a set of sectors $\mathcal{S}_b = \{1, \dots, S_b\}$. Within each sector, two frequency layers are deployed, denoted by $\mathcal{F} = \{f_{cap}, f_{cov}\}$, where f_{cap} represents the capacity layer frequency and f_{cov} represents the coverage layer frequency. The capacity layer is deployed on a higher-frequency carrier and is mainly used to provide additional capacity, whereas the coverage layer operates on a lower-frequency carrier and remains active to ensure basic service continuity. Accordingly, each cell in the network is identified by the triplet (b, s, f) , where b is the BS index, s is the sector index, and f is the frequency-layer index. Each cell is equipped with N_t transmit antennas, while each user has N_r receive antennas. Time is discretized into equal time steps $t \in \mathcal{T} = \{1, 2, \dots, T\}$, where each step has a duration of Δt_s .

For each cell (b, s, f) , let $a_{b,s,f}[t] \in \{0, 1\}$ denote the activity state of the cell at time step t , where $a_{b,s,f}[t] = 1$ indicates an active cell and $a_{b,s,f}[t] = 0$ indicates sleep mode. We assume that coverage cells remain active throughout the evaluation horizon with the full antenna configuration:

$$a_{b,s,f_{cov}}[t] = 1 \quad \text{and} \quad n_{b,s,f_{cov}}^{tx}[t] = 4, \quad \forall b, s, t \quad (1)$$

On the other hand, the capacity cell of each sector is dynamically controlled by the energy management policy and can operate in $(4 \times 4, 2 \times 2, \text{ or } 1 \times 1)$ MIMO, or enter sleep mode.

Each UE is connected to at most one serving cell at any time following a capacity-first rule. Each UE first searches for the capacity cell with the highest reference signal received power (RSRP). If this RSRP is above the capacity-layer admission threshold τ_{cap} , the UE connects to that cell. Otherwise, the UE searches for the coverage cell with the highest RSRP. If this RSRP is above the minimum coverage threshold τ_{cov} , the UE connects to that cell. If neither condition is satisfied, the UE is considered out of service.

B. Channel Model

The wireless propagation environment between the BSs and the UEs is modeled using the 3GPP-compliant geometry-based stochastic channel model 3GPP_38.901_UMa_NLOS [16], as implemented in the QuaDRiGa channel generator [17]. Let $\mathbf{H}_{u,b,s,f} \in \mathbb{C}^{N_r \times N_t}$ denote the full MIMO channel matrix between UE u and cell (b, s, f) . The channel model captures the effects of path loss, shadowing, small-scale fading, antenna patterns, and sector orientation. When a capacity cell operates with a reduced MIMO configuration, only a subset of its transmit antennas is active. Therefore, the effective channel used for transmission is obtained by selecting the active transmit antennas from the full channel matrix. For cell (b, s, f) , the effective channel at time step t is $\mathbf{H}_{u,b,s,f}^{eff}[t] \in \mathbb{C}^{N_r \times n_{b,s,f}^{tx}[t]}$.

To evaluate the spatial multiplexing capability of the effective channel, singular value decomposition (SVD) is applied; $\mathbf{H}_{u,b,s,f}^{eff}[t] = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices, and $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{R_{max}} \geq 0$. The maximum achievable MIMO rank is inherently bounded by the number of active transmit antennas and the user's receive antennas, given by:

$$R_{max}[t] = \min(n_{b,s,f}^{tx}[t], N_r) \quad (2)$$

Let (b_u^*, s_u^*, f_u^*) denote the serving cell of UE u . The total useful received signal power from this serving cell is denoted by $P_u^{rx}[t]$. For a given transmission rank $R \leq R_{max}[t]$, the signal power allocated to the ℓ -th spatial layer is distributed proportionally to the channel's singular values [18]:

$$P_{u,\ell}(R, t) = P_u^{rx}[t] \cdot \frac{\sigma_\ell^2}{\sum_{j=1}^R \sigma_j^2}, \quad \forall \ell \in \{1, \dots, R\} \quad (3)$$

The interference experienced by UE u depends dynamically on the activity states and the traffic loads of the neighboring cells operating on the same serving frequency layer f . Let $\rho_{b,s,f}[t] \in [0, 1]$ denote the resource utilization ratio (i.e., load) of cell (b, s, f) at time step t . The average effective interference power $I_{eff,u}[t]$ is modeled as the sum of the raw interference powers from all non-serving cells, weighted by their respective loads and activity states:

$$I_{eff,u}[t] = \sum_{(b,s) \neq (b_u^*, s_u^*)} a_{b,s,f_u^*}[t] \cdot \rho_{b,s,f_u^*}[t] \cdot I_{u,b,s,f_u^*}^{raw}[t] \quad (4)$$

where $I_{u,b,s,f_u}^{\text{raw}}[t]$ is the raw received interference power from neighboring cell (b, s, f_u^*) under full-load conditions. By including $a_{b,s,f_u^*}[t]$, cells placed in sleep mode contribute zero interference. The SINR for the ℓ -th spatial layer of UE u under candidate rank R is therefore formulated as:

$$\gamma_{u,\ell}(R, t) = \frac{P_{u,\ell}(R, t)}{I_{\text{eff},u}[t] + N_0} \quad (5)$$

where N_0 is the thermal noise power. To maximize spatial multiplexing efficiency, the serving cell dynamically selects the optimal MIMO rank $R_u^*[t]$ that maximizes the theoretical sum capacity across all active layers:

$$R_u^*[t] = \arg \max_{R \in \{1, \dots, R_{\text{max}}[t]\}} \sum_{\ell=1}^R \log_2(1 + \gamma_{u,\ell}(R, t)) \quad (6)$$

To keep track of the MIMO-rank behavior of each cell, we define the number of rank- r reports of cell (b, s, f) as:

$$\Gamma_{b,s,f}^{(r)} = \sum_{t \in \mathcal{T}} \sum_{u \in \mathcal{U}_{b,s,f}[t]} \mathbb{I}\{R_u^*[t] = r\}, \quad r \in \{1, 2, 3, 4\} \quad (7)$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. Thus, each active UE contributes one rank report to its serving cell at each active time step, according to the selected rank $R_u^*[t]$.

C. PRB Allocation, User Activity, and Throughput

Once the optimal rank $R_u^*[t]$ is established, the SINR of each active layer is mapped to a discrete spectral efficiency (SE) value using a predefined modulation and coding scheme (MCS) mapping function, $\mathcal{F}_{\text{MCS}}(\cdot)$. The total SE for UE u is the sum of the efficiencies across all its spatial layers:

$$SE_u[t] = \sum_{\ell=1}^{R_u^*[t]} \mathcal{F}_{\text{MCS}}(\gamma_{u,\ell}(R_u^*[t], t)) \quad (8)$$

Based on the total SE and the bandwidth of a single PRB $W_{\text{PRB}} = 180\text{kHz}$, the data rate provided by one PRB to UE u is $c_u[t] = SE_u[t] \cdot W_{\text{PRB}}$. Consequently, the number of PRBs required to satisfy the user's rate demand D_u is:

$$Q_{\text{req},u}[t] = \left\lceil \frac{D_u}{c_u[t]} \right\rceil \quad (9)$$

If the aggregate PRB demand of the connected UEs does not exceed the available PRBs, each UE is allocated its requested number of PRBs, $Q_{\text{alloc},u}[t] = Q_{\text{req},u}[t]$. Otherwise, the available Q_{total} PRBs are fully distributed among the connected UEs using a proportional-fair allocation rule.

$$Q_{\text{alloc},u}[t] = \left\lfloor \frac{Q_{\text{req},u}[t]}{\sum_{k \in \mathcal{U}_{b,s,f}[t]} Q_{\text{req},k}[t]} Q_{\text{total}} \right\rfloor \quad (10)$$

for all $u \in \mathcal{U}_{b,s,f}[t]$, with any remaining PRBs assigned to users with the largest residual demand. The instantaneous throughput achieved by UE u at step t is then defined as:

$$\zeta_u[t] = Q_{\text{alloc},u}[t] \cdot c_u[t] \quad (11)$$

Each UE u is further assigned a random activity window, reflecting the fact that users are not always active. The

activity duration Δ_u^{act} is drawn from a log-normal distribution, $\Delta_u^{\text{act}} \sim \text{Lognormal}(\mu, \sigma^2)$, parameterized to match a target mean and standard deviation and truncated to the range $[\Delta_{\text{min}}^{\text{act}}, \Delta_{\text{max}}^{\text{act}}]$ to exclude unrealistically short or long sessions [19]. Given Δ_u^{act} , the activity start time t_u^{start} is drawn uniformly over $[1, T - \Delta_u^{\text{act}}]$, so that UE u is active over the discrete time steps

$$\mathcal{T}_u = \{t \in \mathcal{T} : t_u^{\text{start}} \leq t \leq t_u^{\text{start}} + \Delta_u^{\text{act}}\}. \quad (12)$$

The average throughput of UE u is given by:

$$\bar{\zeta}_u = \frac{1}{|\mathcal{T}_u|} \sum_{t \in \mathcal{T}_u} \zeta_u[t] \quad (13)$$

where $|\mathcal{T}_u|$ is the total number of active time steps for UE u .

The total traffic volume carried by cell (b, s, f) is:

$$V_{b,s,f} = \sum_{t=1}^T \sum_{u \in \mathcal{U}_{b,s,f}[t]} \zeta_u[t] \Delta t_s \quad (14)$$

The aggregate network-wide carried traffic volume can be expressed equivalently as the sum over all cells:

$$V_{\text{tot}} = \sum_{b \in \mathcal{B}} \sum_{s \in \mathcal{S}_b} \sum_{f \in \mathcal{F}} V_{b,s,f} \quad (15)$$

The network-wide average user throughput is defined as:

$$\bar{\zeta} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \bar{\zeta}_u \quad (16)$$

Let $\hat{E}_{b,s,f}$ denote the estimated energy consumed by cell (b, s, f) . The total network estimated energy consumption is

$$\hat{E}_{\text{tot}} = \sum_{b \in \mathcal{B}} \sum_{s \in \mathcal{S}_b} \sum_{f \in \mathcal{F}} \hat{E}_{b,s,f} \quad (17)$$

The value of $\hat{E}_{b,s,f}$ depends on the resulting cell-level statistics, including traffic volume, PRB utilization, MIMO-rank distribution, activity time, and sleep duration. In Section III, we show the data-driven model used to estimate $\hat{E}_{b,s,f}$.

III. PROBLEM FORMULATION AND PROPOSED HIERARCHICAL ENERGY-SAVING FRAMEWORK

The energy management problem focuses on selecting the activity state and MIMO configuration of each capacity cell over the time horizon $\mathcal{T}$. The objective is to minimize $\hat{E}_{\text{tot}}$ while satisfying operator-defined service constraints related to throughput and coverage-layer utilization. Accordingly, the energy-minimization problem is formulated as:

$$\begin{aligned} & \min_{\substack{\{a_{b,s,f_{\text{cap}}}[t]\} \\ n_{b,s,f_{\text{cap}}}^{\text{tx}}[t]\}} \hat{E}_{\text{tot}} \\ & \text{s.t. } \bar{\zeta}_{b,s,f} \geq \bar{\zeta}^{\min} \end{aligned} \quad (18a)$$

$$\rho_{b,s,f_{\text{cov}}}[t] \leq \rho_{b,s,f_{\text{cov}}}^{\max}, \text{ if } a_{b,s,f_{\text{cap}}}[t] = 0, \forall b, s, t \quad (18b)$$

$$a_{b,s,f_{\text{cap}}}[t] \in \{0, 1\}, \quad \forall b, s, t \quad (18c)$$

$$n_{b,s,f_{\text{cap}}}^{\text{tx}}[t] \in \{1, 2, 4\}, \quad \forall b, s, t \quad (18d)$$

Here, $\bar{\zeta}^{\min}$ in Constraint (18a) represents the minimum allowable average user throughput per cell, and $\rho_{b,s,f_{cov}}^{\max}$ in Constraint (18b) denotes the maximum permissible PRB utilization threshold for coverage cells only when the corresponding capacity cell is switched off. These parameters serve as operator-configurable operational knobs: tightening them forces a conservative policy that prioritizes QoS, whereas relaxing them allows for aggressive energy savings.

Solving this problem directly is computationally prohibitive because each capacity cell can choose one of four states at every time step, leading to $4^{K|T|}$ possible state trajectories for K capacity cells. Moreover, each trajectory affects user association, interference, SINR, rank selection, PRB allocation, throughput, and energy consumption, making every evaluation costly. Therefore, we propose a practical hierarchical energy-saving framework that combines threshold-based capacity-layer control, data-driven energy estimation, and Bayesian optimization. For each candidate policy, the system model evaluates the resulting user association, SINR, rank selection, PRB allocation, and throughput; the trained energy models then estimate the total network energy consumption, and Bayesian optimization is used to tune the thresholds governing MIMO reconfiguration and capacity-cell shutdown.

A. Data-Driven Energy Consumption Model

To evaluate the objective function in Problem (18), the proposed framework employs data-driven regression models. The proposed approach estimates the energy consumed by each cell from the operational statistics produced by the system-level simulation that accurately mimics the real LTE network. Since coverage and capacity cells have different operating characteristics, two separate XGBoost regression models are trained using related but not identical input features. Both models follow the same general structure, where a cell-level feature vector is mapped to an estimated energy-consumption value. For a coverage cell (b, s, f_{cov}) , the feature vector is

$$\mathbf{z}_{b,s,f_{cov}} = [V_{b,s,f_{cov}}, \bar{\rho}_{b,s,f_{cov}}, \bar{U}_{b,s,f_{cov}}, \Gamma_{b,s,f_{cov}}^{(1)}, \Gamma_{b,s,f_{cov}}^{(2)}, \Gamma_{b,s,f_{cov}}^{(3)}, \Gamma_{b,s,f_{cov}}^{(4)}, A_{b,s,f_{cov}}]^T \quad (19)$$

where $V_{b,s,f_{cov}}$ is the carried traffic volume, $\bar{\rho}_{b,s,f_{cov}}$ is the average PRB utilization, $\bar{U}_{b,s,f_{cov}}$ is the average number of connected users, $\Gamma_{b,s,f_{cov}}^{(r)}$ is the number of rank- r reports, and $A_{b,s,f_{cov}}$ is the cell activity time.

For a capacity-layer cell (b, s, f_{cap}) , the feature vector additionally includes the accumulated sleep duration

$$\mathbf{z}_{b,s,f_{cap}} = [V_{b,s,f_{cap}}, \bar{\rho}_{b,s,f_{cap}}, \bar{U}_{b,s,f_{cap}}, \Gamma_{b,s,f_{cap}}^{(1)}, \Gamma_{b,s,f_{cap}}^{(2)}, \Gamma_{b,s,f_{cap}}^{(3)}, \Gamma_{b,s,f_{cap}}^{(4)}, A_{b,s,f_{cap}}, S_{b,s,f_{cap}}]^T \quad (20)$$

where $S_{b,s,f_{cap}}$ denotes the accumulated sleep duration of the capacity cell over the evaluation horizon. The predicted energy consumption of cell (b, s, f) is then obtained using the layer-specific regression model as

$$\hat{E}_{b,s,f} = \begin{cases} \hat{\mathcal{E}}_v(\mathbf{z}_{b,s,f_{cov}}), & f = f_{cov}, \\ \hat{\mathcal{E}}_c(\mathbf{z}_{b,s,f_{cap}}), & f = f_{cap}, \end{cases} \quad (21)$$

where $\hat{\mathcal{E}}_v(\cdot)$ and $\hat{\mathcal{E}}_c(\cdot)$ denote the trained XGBoost regression models for the coverage and capacity cells, respectively.

The total predicted network energy consumption used in the optimization objective is therefore

$$\hat{E}_{\text{tot}} = \sum_{b \in \mathcal{B}} \sum_{s \in \mathcal{S}_b} (\hat{E}_{b,s,f_{cap}} + \hat{E}_{b,s,f_{cov}}) \quad (22)$$

Since the input features of the energy models are affected by the selected activity and MIMO states, $\hat{E}_{\text{tot}}$ naturally captures the impact of the proposed control policy on both capacity-layer and coverage cells.

B. Hierarchical Threshold-Based Control Policy

To make the problem tractable, we restrict the solution to a threshold-based control policy that applies energy-saving actions in an ordered manner. At each time step, the controller attempts to use the deepest feasible energy-saving state: full capacity cell sleep is considered first, followed by reduced MIMO configurations using (1×1) or (2×2) MIMO, while full 4×4 operation is retained whenever the load or throughput indicate that additional capacity is required.

Let the threshold vector be denoted by

$$\boldsymbol{\theta} = [\rho_{1 \times 1}, \rho_{2 \times 2}, \bar{\zeta}_{cap}^{\min}, \rho_{\text{sleep}}, \rho_{\text{wake}}, \bar{\zeta}_{cov}^{\text{sleep}}, \bar{\zeta}_{cov}^{\text{wake}}]^T \quad (23)$$

where $\rho_{1 \times 1}$ and $\rho_{2 \times 2}$ are the PRB-utilization thresholds used for MIMO reconfiguration of active capacity cells, and $\bar{\zeta}_{cap}^{\min}$ is the capacity-layer throughput guard used to prevent MIMO downgrading when capacity-layer users already experience insufficient throughput. The parameter ρ_{sleep} is the maximum projected coverage-layer utilization allowed for capacity-cell shutdown, while ρ_{wake} is the coverage-layer utilization threshold used to reactivate a sleeping capacity-layer cell. Finally, $\bar{\zeta}_{cov}^{\text{sleep}}$ is the minimum coverage-layer throughput required to allow capacity-cell shutdown, and $\bar{\zeta}_{cov}^{\text{wake}}$ is the minimum acceptable coverage-layer throughput below which a sleeping capacity-layer cell is reactivated. The thresholds satisfy

$$\rho_{1 \times 1} < \rho_{2 \times 2}, \quad \rho_{\text{sleep}} < \rho_{\text{wake}}, \quad \bar{\zeta}_{cov}^{\text{wake}} \leq \bar{\zeta}_{cov}^{\text{sleep}} \quad (24)$$

The first condition ensures an ordered MIMO reconfiguration rule. The second introduces utilization-based hysteresis between sleep and wake-up decisions. The third introduces throughput-based hysteresis. These conditions reduce unnecessary state oscillations.

For each sector (b, s) and time step t , the controller first evaluates whether the capacity cell can be switched off. If the capacity-layer traffic were transferred to the coverage layer, the projected coverage-layer utilization is approximated as

$$\rho_{b,s,f_{cov}}^{\text{proj}}[t] = \rho_{b,s,f_{cov}}[t] + \rho_{b,s,f_{cap}}[t] \quad (25)$$

The capacity cell is allowed to enter sleep mode only if

$$\rho_{b,s,f_{cov}}^{\text{proj}}[t] \leq \rho_{\text{sleep}}, \quad \bar{\zeta}_{b,s,f_{cov}}[t] \geq \bar{\zeta}_{cov}^{\text{sleep}} \quad (26)$$

where $\bar{\zeta}_{b,s,f_{cov}}[t]$ denotes the average throughput of users served by the coverage cell (b, s, f_{cov}) at time step t . These conditions prevent shutdown decisions that would overload

the coverage layer or degrade the service quality of users already served by it.

If a capacity cell is already in sleep mode, it is reactivated when either the coverage-layer utilization exceeds the wake-up threshold or the coverage-layer user throughput falls below the wake-up throughput guard:

$$\rho_{b,s,f_{cov}}[t] \geq \rho_{wake} \quad || \quad \bar{\zeta}_{b,s,f_{cov}}[t] \leq \bar{\zeta}_{cov}^{wake}. \quad (27)$$

Upon reactivation, the capacity cell returns to 4×4 MIMO configuration. If the capacity cell remains active, the number of active transmit antennas is then selected as

$$n_{b,s,f_{cap}}^{tx}[t] = \begin{cases} 1, & \rho_{b,s,f_{cap}}[t] \leq \rho_{1 \times 1}, \\ & \bar{\zeta}_{b,s,f_{cap}}[t] \geq \bar{\zeta}_{cap}^{min}, \\ 2, & \rho_{1 \times 1} < \rho_{b,s,f_{cap}}[t] \leq \rho_{2 \times 2}, \\ & \bar{\zeta}_{b,s,f_{cap}}[t] \geq \bar{\zeta}_{cap}^{min}, \\ 4, & \text{otherwise.} \end{cases} \quad (28)$$

Accordingly, low utilization allows capacity cells to operate with fewer transmit antennas, whereas high utilization or insufficient throughput forces the full 4×4 configuration.

C. Bayesian Optimization of Control Parameters

The performance of the hierarchical policy depends on the selected threshold vector θ . Thus, Bayesian optimization is used to efficiently tune θ . Let Θ denote the feasible threshold search space. For each candidate $\theta \in \Theta$, the threshold-based policy is applied over the time horizon $\mathcal{T}$. The simulation then produces the resulting traffic volume, PRB utilization, MIMO-rank reports, activity time, sleep time, user throughput, and coverage-layer load. These statistics are passed to the energy models in Section III-A to compute $\hat{E}_{tot}(\theta)$.

Bayesian optimization minimizes the objective function:

$$J(\theta) = \hat{E}_{tot}(\theta) \quad (29)$$

Bayesian optimization constructs a stochastic surrogate model of $J(\theta)$ using the set of previously evaluated candidates as

$$\mathcal{D}_x = \left\{ (\theta_i, J(\theta_i)) \right\}_{i=1}^x \quad (30)$$

In iteration $x + 1$, an acquisition function selects the next threshold vector to evaluate:

$$\theta_{x+1} = \arg \max_{\theta \in \Theta} \alpha(\theta | \mathcal{D}_x) \quad (31)$$

where $\alpha(\cdot)$ balances exploration of uncertain regions of the search space and exploitation of promising threshold values.

After X Bayesian-optimization iterations, the optimized threshold vector is selected as

$$\theta^* = \arg \min_{\theta_i \in \mathcal{D}_X} J(\theta_i). \quad (32)$$

The corresponding policy is then used as the optimized energy-saving policy in the performance evaluation.

For X evaluated threshold vectors, fitting an exact Gaussian-process surrogate at iteration x has complexity $\mathcal{O}(x^3)$. If the surrogate is refitted after each evaluation, the cumulative Bayesian optimization overhead is $\mathcal{O}(X^4)$. [20]

Algorithm 1 Bayesian Optimization of the Hierarchical Energy-Saving Policy

- 1: Define the feasible threshold search space Θ
- 2: Initialize the evaluated set $\mathcal{D}_0 = \emptyset$
- 3: **for** $x = 1, \dots, X$ **do**
- 4: Select θ_x using the acquisition function
- 5: Apply the threshold-based policy over $\mathcal{T}$
- 6: Extract traffic, PRB utilization, rank reports, activity time, sleep time, and throughput statistics
- 7: Estimate $\hat{E}_{tot}(\theta_x)$ using the trained energy models
- 8: Compute $J(\theta_x)$
- 9: Update $\mathcal{D}_x = \mathcal{D}_{x-1} \cup \{(\theta_x, J(\theta_x))\}$
- 10: **end for**
- 11: Select $\theta^* = \arg \min_{\theta_x \in \mathcal{D}_X} J(\theta_x)$
- 12: **return** θ^*

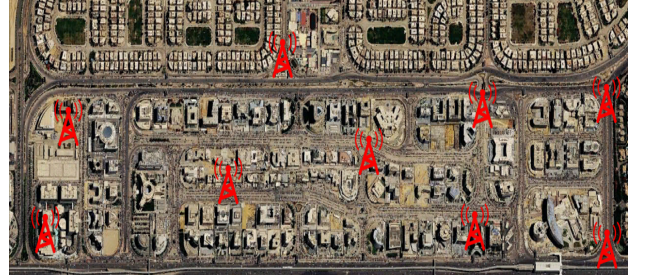


Fig. 1: Network Topology.

IV. PERFORMANCE EVALUATION

This section evaluates the proposed hierarchical energy-saving framework through system-level simulations. The objective is to quantify the energy reduction achieved by the proposed policy while verifying that the resulting user throughput and coverage-layer utilization remain within acceptable limits.

The BSs locations used in the simulation are selected from a real network deployment in Egypt as shown in Figure 1. This allows the simulation to capture a realistic, non-uniform site layout. Real sector coordinates and azimuths are used to construct the simulated topology, and the locations are presented in a local Cartesian coordinate system to preserve network confidentiality. UEs are randomly distributed within the bounding region of the selected area with an additional margin around the sites. The simulated network follows the two-layer sectorized architecture described in Section II. Each sector contains one capacity cell and one coverage cell. The main simulation parameters are summarized in Table I.

Table II shows the prediction accuracy of the two XGBoost energy-consumption models, confirming that the trained models provide a sufficiently accurate basis for estimating energy consumption within the proposed framework.

To isolate the contribution of each control mechanism, the proposed joint MIMO-and-sleep policy, referred to as the hierarchical policy [4], is compared against four benchmark cases. The first is the baseline configuration, where all capacity cells remain active with full 4×4 MIMO operation. The second is a MIMO-only policy, in which Bayesian optimization is

TABLE I: Simulation Parameters

Parameter	Value
Number of BSs	9
Number of Sectors	34
Number of UEs	100
Number of Simulations	25
Simulation duration	3600 s
Δt_s	10 s
Capacity-layer carrier frequency	2.6 GHz
Coverage-layer carrier frequency	2.1 GHz
System bandwidth per layer	20 MHz
Number of PRBs per cell	100
N_t, N_r	4
BS height	30 m
UE height	1.5 m
Antenna downtilt	6°
τ_{cap}	-105 dBm
τ_{cov}	-124 dBm
User Demand	8 Mbps
μ	1800 s
σ^2	900 s
Δ_{act}^{act}	60 s
Δ_{act}^{min}	60 s
Δ_{act}^{max}	3600 s
Rank-reporting interval	10 ms
Bayesian-optimization evaluations	100
Acquisition function	Expected improvement plus
ζ_{min}	6 Mbps
$\zeta_{cap}^{min}, \zeta_{cov}^{sleep}, \zeta_{cov}^{wake}$	[6, 12] Mbps
$\rho_{1 \times 1}$	[10, 35]%
$\rho_{2 \times 2}$	[30, 70]%
ρ_{sleep}	[20, 70]%
ρ_{wake}	[30, 80]%

TABLE II: Energy consumption models performance metrics

Model	MAE	MAPE
Coverage Layer	9.9 Wh	3.66%
Capacity Layer	16 Wh	10.2%

used to tune the MIMO reconfiguration thresholds while capacity-cell shutdown is disabled. The third is a sleep-only policy, in which Bayesian optimization is used to tune the shutdown and wake-up thresholds while active capacity cells remain fixed at 4×4 MIMO. The fourth is a static-threshold hierarchical policy, which applies the same joint MIMO-and-sleep decision logic but without Bayesian optimization [4]. For the static hierarchical policy, the fixed thresholds are set to $\rho_{1 \times 1} = 30\%$, $\rho_{2 \times 2} = 50\%$, $\zeta_{cap}^{min} = 8$ Mbps, $\rho_{sleep} = 40\%$, $\rho_{wake} = 60\%$, $\zeta_{cov}^{sleep} = 10$ Mbps, and $\zeta_{cov}^{wake} = 7$ Mbps.

For the Bayesian-optimized policies, the throughput and load constraints are embedded in the threshold search space reported in Table I. Thus, the optimizer tunes the energy-saving policy within predefined ranges, rather than minimizing energy without service constraints.

Fig. 2 compares the total energy consumption of the evaluated policies. The baseline, where all capacity cells remain active with full 4×4 MIMO, records the highest consumption of 14.648 kWh. MIMO-only control provides a limited reduction to 14.365 kWh, corresponding to about 1.93% saving. The sleep-only policy consumes 12.826 kWh, while the proposed joint MIMO-and-sleep policy achieves the lowest consumption of 12.4 kWh, corresponding to approximately 15.3% saving. Compared to the static-threshold

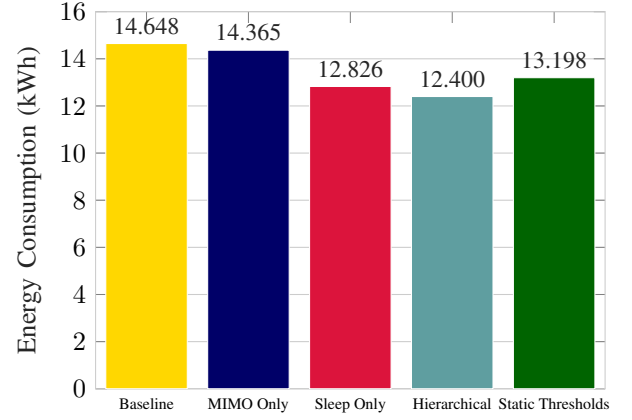


Fig. 2: Energy consumption (kWh).

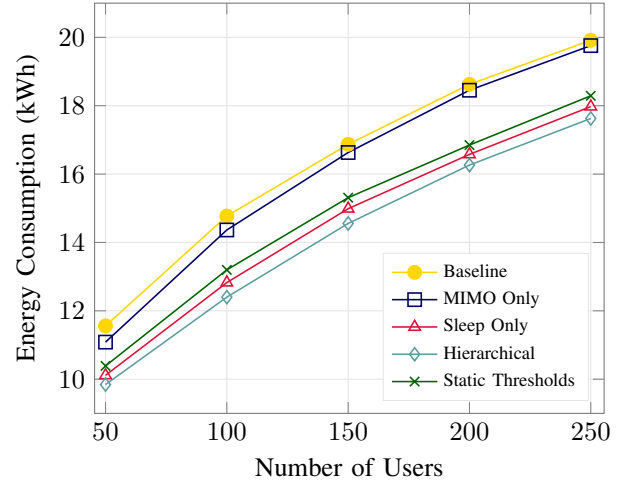


Fig. 3: Energy consumption (kWh).

policy, which consumes 13.198 kWh and achieves about 9.89% saving, the proposed optimized policy provides higher energy reduction by jointly adapting shutdown and MIMO reconfiguration decisions to the network operating conditions.

Fig. 3 evaluates the scalability of the evaluated policies as the number of users increases from 50 to 250. Energy consumption increases for all policies due to the higher offered traffic load and resource utilization. The baseline consistently has the highest consumption, rising from 11.559 kWh to 19.915 kWh, confirming that keeping all capacity cells active with full 4×4 MIMO is the least energy-efficient strategy. The proposed hierarchical policy achieves the lowest consumption across the full user-density range, increasing from 9.843 kWh at 50 users to 17.627 kWh at 250 users. This corresponds to savings of approximately 14.9% and 11.5%, respectively, relative to the baseline. Although the relative saving decreases at higher user densities because more capacity-layer resources must remain active, the proposed policy remains more efficient than the MIMO-only, sleep-only, and static-threshold policies by jointly combining capacity-cell shutdown with adaptive MIMO reconfiguration.

Fig. 4 evaluates the impact of increasing the per-user throughput demand from 4 Mbps to 12 Mbps. As expected,

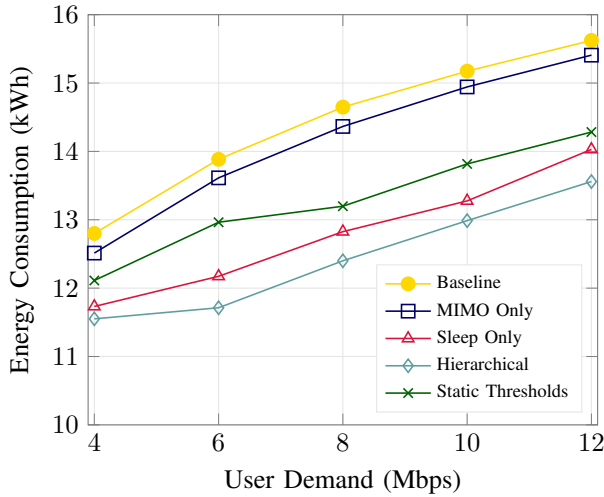


Fig. 4: Energy consumption (kWh).

energy consumption increases for all policies because higher demand requires more PRBs, increases cell utilization, and reduces opportunities for sleep mode or MIMO downgrading. The baseline consistently has the highest consumption, increasing from 12.800 kWh to 15.623 kWh, while the proposed hierarchical policy achieves the lowest consumption, increasing from 11.551 kWh to 13.559 kWh. This corresponds to energy savings of approximately 9.8% and 13.2%, respectively, relative to the baseline. These results show that the optimized joint control remains effective across different demand levels, outperforming the MIMO-only, sleep-only, and static-threshold policies by combining capacity-cell shutdown with adaptive MIMO reconfiguration and by tuning the thresholds to the network operating conditions.

V. CONCLUSION

This paper investigated the joint use of capacity-cell sleep and adaptive MIMO reconfiguration to improve energy efficiency in a two-layer cellular network. The results show that the two mechanisms complement each other; sleep mode provides a strong opportunity to reduce energy consumption during low-load conditions, while MIMO reconfiguration offers a finer level of adaptation when complete shutdown is infeasible. This highlights the importance of jointly optimizing energy-saving actions rather than applying each mechanism independently or relying on fixed control thresholds.

Future work will extend the framework towards joint energy saving and load balancing, where users and traffic can be dynamically steered across layers or neighboring cells before applying sleep decisions. This can help avoid excessive coverage-layer congestion and improve the robustness of energy-saving operation. Further extensions may include coordinated multi-cell control, online learning-based policies, and mobility-aware evaluation.

REFERENCES

[1] M. Pickavet, W. Vereecken, S. Demeyer, P. Audenaert, B. Vermeulen, C. Develder, D. Colle, B. Dhoedt, and P. Demeester, "Worldwide

energy needs for ICT: The rise of power-aware networking," in *Proc. 2008 2nd IEEE International Symposium on Advanced Networks and Telecommunication Systems*, Mumbai, India, May 2008.

[2] A. Fehske, G. Fettweis, J. Malmolin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 55–62, 2011.

[3] P. Frenger, Y. Jading, and A. Nader, "Energy performance of 6G radio access networks: A once in a decade opportunity," Ericsson, Tech. Rep., 2024. [Online]. Available: <https://www.ericsson.com/4ab133/assets/local/reports-papers/white-papers/2024/6g-energy-performance.pdf>

[4] M. Zaki, A. Gaber, M. G. Khafagy, M. Beshara, and N. Abdelbaki, "Autonomous traffic-aware and QoS-constrained capacity cell shutdown for green mobile networks," in *2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. IEEE, 2023, pp. 142–145.

[5] A. AlAlwani, A. Itman, A. Gaber, M. Zaki, M. G. Khafagy, K. Banawan, and K. Seddik, "ML-aided traffic-aware base station sleep threshold design with user throughput guarantees," in *2025 IEEE 11th International Conference on Network Softwarization (NetSoft)*. IEEE, 2025, pp. 194–198.

[6] A. Arbi and T. O'Farrell, "Energy efficiency in 5G access networks: Small cell densification and high order sectorisation," in *Proc. 2015 IEEE ICCW*, London, UK, Jun. 2015.

[7] M. Masoudi, M. G. Khafagy, E. Soroush, D. Giacomelli, S. Morosi, and C. Cavdar, "Reinforcement learning for traffic-adaptive sleep mode management in 5G networks," in *Proc. IEEE PIMRC'20*, London, UK, Oct. 2020.

[8] M. Ozturk, A. I. Abubakar, J. P. B. Nadas, R. N. B. Rais, S. Hussain, and M. A. Imran, "Energy optimization in ultra-dense radio access networks via traffic-aware cell switching," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 2, pp. 832–845, 2021.

[9] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. Leung, and H. V. Poor, "Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1936–1947, 2017.

[10] T. Zhou, Z. Liu, J. Zhao, C. Li, and L. Yang, "Joint user association and power control for load balancing in downlink heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2582–2593, 2017.

[11] X. Huang, W. Xu, H. Shen, H. Zhang, and X. You, "Utility-energy efficiency oriented user association with power control in heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 526–529, 2018.

[12] F. Hussain, S. A. Hassan, R. Hussain, and E. Hossain, "Machine learning for resource management in cellular and iot networks: Potentials, current solutions, and open challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1251–1275, 2020.

[13] Y. L. Lee, W. L. Tan, S. B. Y. Lau, T. C. Chuah, A. A. El-Saleh, and D. Qin, "Joint cell activation and user association for backhaul load balancing in green HetNets," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1486–1490, 2020.

[14] M. Aboelwafa, M. Zaki, A. Gaber, K. Seddik, Y. Gadallah, and A. Elezabi, "Machine learning-based MIMO enabling techniques for energy optimization in cellular networks," in *Proc. IEEE CCNC'20*, Las Vegas, NV, USA, Jan. 2020.

[15] A. El Soukkary and K. G. Seddik, "Energy efficient load balancing in multiband cellular networks via reinforcement learning," in *Proc. IEEE PIMRC'25*, Istanbul, Turkiye, Sep. 2025.

[16] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Technical Report TR 38.901, May 2024, Release 18.

[17] S. Jaeckel, L. Raschkowski, K. Börner, L. Thiele, F. Burkhardt, and E. Eberlein, "QuaDRiGa-quasi deterministic radio channel generator, user manual and documentation," *Fraunhofer Heinrich Hertz Institute, Tech. Rep.* v2.8.1, 2023.

[18] E. Telatar, "Capacity of multi-antenna gaussian channels," *European transactions on telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

[19] M. Hamid, N. Björnell, and S. B. Slimane, "Empirical statistical model for lte downlink channel occupancy," *Wireless personal communications*, vol. 96, no. 1, pp. 855–866, 2017.

[20] A. E. Siemenn and T. Buonassisi, "Decreasing the computing time of bayesian optimization using generalizable memory pruning," in *2023 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2023, pp. 1–7.