A Multi-Armed Bandit Framework for Online Optimisation in Green Integrated Terrestrial and Non-Terrestrial Networks

Henri Alam[†‡], Antonio De Domenico[†], Tareq Si Salem[†], Florian Kaltenberger[‡]
[†]Huawei Technologies, Paris Research Center, 20 quai du Point du Jour, Boulogne Billancourt, France.
[‡]EURECOM, 2229 route des Crêtes, 06904 Sophia Antipolis Cedex, France.

***Abstract*—Integrated terrestrial and non-terrestrial network (TN-NTN) architectures offer a promising solution for expanding coverage and improving capacity for the network. While non-terrestrial networks (NTNs) are primarily exploited for these specific reasons, their role in alleviating terrestrial network (TN) load and enabling energy-efficient operation has received comparatively less attention. In light of growing concerns associated with the densification of terrestrial deployments, this work aims to explore the potential of NTNs in supporting a more sustainable network. In this paper, we propose a novel online optimisation framework for integrated TN-NTN architectures, built on a multi-armed bandit (MAB) formulation and leveraging the Bandit-feedback Constrained Online Mirror Descent (BCOMD) algorithm. Our approach adaptively optimises key system parameters—including bandwidth allocation, user equipment (UE) association, and macro base station (MBS) shutdown—to balance network capacity and energy efficiency in real time. Extensive system-level simulations over a 24-hour period show that our framework significantly reduces the proportion of unsatisfied UEs during peak hours and achieves up to 19 % throughput gains and 5 % energy savings in low-traffic periods, outperforming standard network settings following 3GPP recommendations.**

I. Introduction

Recent advancements in cellular communications have sharply increased the demand for high-speed data, driving the need for broader network coverage and higher capacity. To address these challenges, mobile operators have intensified their deployment of terrestrial MBSs. This constant expansion has resulted in increased energy consumption, raising environmental concerns from a societal perspective as well as economic challenges for network operators. Thus, minimising energy consumption while maintaining Quality of Service (QoS) standards has become a key objective in mobile network management [1].

NTNs have emerged as a practical solution to complement TNs and extend coverage to underserved areas in the past few years. NTNs use airborne platforms such as drones or satellites as relay nodes or MBSs to provide connectivity to user equipment UEs. Their key advantage lies in offering wide-area coverage, particularly in remote regions where deploying terrestrial MBSs is costly or impractical. In particular, low-earth orbit (LEO) satellites are poised to play a central role in delivering high-capacity space-based connectivity [2], as they benefit from reduced latency, stronger signals, and lower energy requirements for both launch and communication, effectively building an integrated TN-NTN capable of delivering seamless, high-capacity communication services while ensuring efficient and reliable connectivity for the UEs [3].

Typically, these UEs are associated with the MBS offering the highest reference signal received power (RSRP). However, this approach ignores traffic demand variations, often leading to poor load distribution and degraded network performance. An effective policy for load balancing usually involves a pricing-based association strategy [4] which considers both signal quality and cell load. Alam et al. [5] propose a similar approach in the context of an integrated TN-NTN, leveraging satellite resources to distribute the load to improve overall network capacity and coverage.

Moreover, keeping all MBSs active during low traffic can lead to inefficient energy and resource use, as many may be underutilised. In an integrated TN-NTN, selectively turning off some MBSs and offloading UEs to satellites can help reduce energy consumption. To that end, our previous work [6] proposed a framework designed to balance network fairness and energy consumption by adjusting to varying traffic conditions in an integrated TN-NTN.

In this paper, we propose a novel framework that dynamically balances network capacity and terrestrial energy consumption in an integrated TN-NTN architecture. The framework is formulated as a MAB problem and leverages a constrained online learning algorithm named BCOMD, introduced in [7], to adaptively select the optimal configuration of system parameters in response to time-varying traffic demands. By jointly optimising UE association, bandwidth allocation, and MBS shutdown decisions, the framework effectively improves load distribution, reduces energy usage during low-traffic periods, and enhances user satisfaction in high-traffic scenarios.

II. System Model

We consider a downlink (DL) cellular network, operating over T time slots, which consists of M terrestrial MBSs and N LEO satellites mounted with MBSs, for a total of L MBSs. K UEs are deployed in the study area. The network operates in the S-band at approximately 2 GHz, where the total available system bandwidth W is allocated by the mobile network operator between terrestrial and non-terrestrial tiers, each using orthogonal frequency bands. Throughout this paper, we denote the set of terrestrial MBSs by $\mathcal{T}$ and the set of satellite MBSs

by $\mathcal{S}$. The set of UEs is represented as $\mathcal{U} = \{1, \ldots, K\}$, and the complete set of MBSs is given by $\mathcal{B} = \mathcal{T} \cup \mathcal{S} = \{1, \ldots, L\}$. For the channel model, the large-scale channel gain between a terrestrial MBS j and a UE i is computed as follows:

$$\beta_{ij} = \mathcal{M}\Big(G_{\mathrm{T_X}} + G_{\mathrm{UE}} + \mathrm{PL}_{ij}^{b} + \mathrm{SF}_{ij} + \mathrm{PL}_{ij}^{\mathrm{tw}} + \mathrm{PL}_{ij}^{\mathrm{in}} + \mathcal{N}\left(0, \sigma_p^2\right)\Big). \tag{1}$$

where all terms are in dB and are mapped to linear space using the operator $\mathcal{M}(\cdot)$. Here, G_{UE} and $G_{\mathrm{T_X}}$ represent the UE and MBS antenna gain respectively, PL_{ij}^{b} is the basic outdoor path loss [8, Table 7.4.1-1], and SF_{ij} is the shadow fading, normally distributed with mean 0 and variance σ_{SF}^2. The remaining terms ($\mathrm{PL}_{ij}^{\mathrm{tw}}$, $\mathrm{PL}_{ij}^{\mathrm{in}}$, and $\mathcal{N}(0, \sigma_p^2)$) account for building penetration losses, detailed in [8]. The line-of-sight (LoS) condition for each UE is computed as in [8, Table 7.4.2-1]. In contrast, for a satellite MBS j serving UE i, the large-scale channel gain can be expressed as follows [9]:

$$\beta_{ij} = \mathcal{M}\left(G_{\mathrm{T_X}} + G_{\mathrm{UE}} + \mathrm{PL}_{ij}^{b} + \mathrm{SF}_{ij} + \mathrm{CL} + \mathrm{PL}_{ij}^{s} + \mathrm{PL}_{ij}^{e}\right) \tag{2}$$

In (2), CL denotes clutter loss, which is the attenuation caused by buildings and vegetation near the UE, and PL_s represents scintillation loss, capturing rapid fluctuations in signal amplitude and phase due to ionospheric conditions. Lastly, PL_{ij}^{e} denotes the building entry loss, representing the attenuation experienced by all UEs positioned indoors. Since interference between both terrestrial and non-terrestrial tiers is negligible due to orthogonal bandwidth allocations, the large-scale signal-to-interference-plus-noise ratio (SINR) for each UE i can be calculated as follows:

$$\gamma_{ij} = \frac{\beta_{ij} p_j}{\sum\limits_{j' \in \mathcal{I}_j} \beta_{ij'} p_{j'} + \sigma^2}, \tag{3}$$

where p_j represents the transmit power per resource element (RE) allocated by MBS j, $\mathcal{I}_j$ denotes the set of MBSs interfering with serving MBS j, and σ^2 accounts for the noise power per RE. In our study, each UE i has a specific data-rate demand ρ_i, modelled as a random variable following an exponential distribution of parameter λ_{U}. Then, the number of physical resource blocks (PRBs) assigned to the UE by the associated MBS j is computed as:

$$B_{ij} = \left\lceil \frac{\rho_i}{\Delta \log_2\left(1 + \gamma_{ij}\right)} \right\rceil. \tag{4}$$

In (4), the denominator is the product between the spectral efficiency and Δ, which represents the total bandwidth of a single PRB in 5G new radio (NR). Finally, $\lceil \cdot \rceil$ denotes the ceiling function, which rounds up the input number to the nearest integer. The load ν_j for MBS j is then defined as the fraction of PRBs being utilised. Using this, we can calculate the mean throughput for UE i served by MBS j as:

$$R_{ij} = \Delta B_{ij} \log_2(1 + \gamma_{ij}). \tag{5}$$

Finally, the terrestrial MBS energy consumption model, dependent on various parameters as described in [10], can be succinctly represented as the sum of three components. The baseline component refers to the energy consumed by elements that remain active even when the MBS is shut down. The static component is the fixed energy consumption required to maintain essential systems operational, independent of traffic load. Lastly, the dynamic component varies with the traffic load, increasing when the MBS transmits at higher power levels or utilises additional PRBs. For a MBS j, this model is written as:

$$Q_j = P_0 + p_j + \psi_j \mathbb{1}_{\{p_j > 0\}}, \tag{6}$$

where P_0 denotes the baseline energy consumption, ψ_j indicates the static component, and p_j, the transmission power of MBS j, corresponds to the dynamic component. Additionally, $\mathbb{1}_{\{\cdot\}}$ is the indicator function that equals 1 if the inputted condition is True, and 0 otherwise. The satellite is presumed to harvest its energy from solar panels. We also define the parameters that we intend to optimise in our framework: ε is the proportion of the bandwidth which is allocated to the LEO satellites. τ_ν is the threshold considered for the load of a MBS, which determines whether we attempt to shut it down or not, while τ_{RSRP} is also a threshold for the perceived RSRP. Finally, α is a weight that controls the influence of MBS load on the UE association decision. The role of each parameter will be explained in more detail in Section IV.

III. Problem Formulation

Similar to the model proposed by Alam et al. [6], we aim to design a framework that jointly optimises network capacity and TN energy consumption by dynamically adjusting resource allocation based on network load, while satisfying the data-rate requirements of each UE.

More specifically, our objective is to identify the optimal policy, defined as an action sequence selected from a set of n distinct configurations of $\theta = [\varepsilon, \tau_\nu, \tau_{\mathrm{RSRP}}, \alpha]$, each referred to as an *arm*. Indeed, each arm represents a specific setting for those parameters, chosen from the n different combinations possible. These arms directly impact network behaviour through a heuristic, which is detailed in Section IV. To evaluate the performance of the network at time t, we define a cost function that captures the key trade-offs introduced by selecting a given arm a_t, which is drawn according to a probability distribution x_t over the action space Δ_n (the n-dimensional probability simplex):

$$f_t\left(a_t, x_t\right) = \zeta \sum_{j \in \mathcal{B}} Q_j(\theta) - \sum_{i \in \mathcal{U}} \log(R_i(\theta)), \tag{7}$$

where $R_i(\theta)$ denotes the throughput perceived by UE i and $Q_j(\theta)$ represents the energy consumption of MBS j, both for a given configuration θ. ζ is a regularisation factor that allows balancing the trade-off between UE performance (i.e., sum log-throughput (SLT)) and network energy consumption. Without loss of generality, the cost function values are normalised between 0 and 1. In parallel, we define the constraint violation incurred by selecting arm a_t as:

$$g_t\left(a_t, x_t\right) = \frac{1}{K} \sum_{i \in \mathcal{U}} \mathbb{1}_{\{R_i < \rho_i\}}. \tag{8}$$

Note that if a UE perceives a RSRP lower than a set threshold $\mathrm{RSRP}_{\min}$, it is considered unsatisfied.

Naturally, the cost distribution associated with each arm evolves over time as network demand fluctuates with the traffic load. Indeed, an arm that enhances capacity under high traffic may be suboptimal in low-traffic scenarios—emphasising the need for context-aware arm selection. This non-stationarity in cost aligns well with the adversarial bandit-feedback setting proposed in [7]. Accordingly, we leverage the algorithm introduced in [7] to handle dynamic costs while ensuring long-term constraint satisfaction.

We denote by x^* the *oracle* policy, i.e., the action sequence that achieves the minimum cumulative loss over the time horizon:

$$\{x_t^\star\}_{t=1}^T \in \arg \min_{\{x_t\}_{t=1}^T \in \bigcap_{t=1}^T \Delta_{n,t}} \left\{ \sum_{t=1}^T f_t(a_t, x_t) \right\}, \quad (9)$$

where $\Delta_{n,t}$ is the set of feasible points within the simplex at time t:

$$\Delta_{n,t} \triangleq \{x \in \Delta_n : g_t(a_t, x) = 0\}. \quad (10)$$

Our goal is to find a policy π that minimises cumulative loss relative to the oracle, while also satisfying time-varying constraints. Adopting the notation from [7], we define the regret and constraint violation which we want to minimise as:

$$\mathfrak{R}_T(\pi) \triangleq \mathbb{E}_\pi \left[\sum_{t=1}^T f_t(a_t, \pi_t) \right] - \sum_{t=1}^T f_t(a_t, x_t^*), \quad (11)$$

$$\mathfrak{V}_T(\pi) \triangleq \mathbb{E}_\pi \left[\sum_{t=1}^T g_t(a_t, \pi_t) \right]. \quad (12)$$

IV. Finding Optimal Policy Using BCOMD

In this section, we present the designed framework, which takes as input the set of parameters θ and associates a cost that quantifies both the network performance and the quality of the parameter configuration. Then, we describe the method used to determine the optimal setting of θ, which minimises Eq. (11) and (12).

A. Network Optimisation Framework

The framework proposed to measure network performance given θ can be broken down into the following steps:

1) **Initialisation:** Given the input $\theta = [\varepsilon, \tau_\nu, \tau_{\mathrm{RSRP}}, \alpha]$, we first associate each UE using the max-RSRP criterion, compute the resulting load on each MBS, and redistribute the resources based on ε.
2) **UE association:** For each UE i, we propose a pricing function which takes into account the load of MBS j:

$$P_i(j) = \mathrm{RSRP}_{ij} - \alpha \nu_j. \quad (13)$$

A positive value of α discourages highly loaded MBSs from serving additional UEs, thereby promoting load balancing across the terrestrial network. Conversely, a negative value encourages UEs to associate with loaded MBSs, leading to a higher number of inactive MBSs. We associate the UE to the MBS which maximises the pricing function.
3) **MBS Shutdown:** For each terrestrial MBS j, we check if the sum of its load and the load of the satellite is smaller than τ_ν. If true, and all UEs served by this MBS perceive a RSRP greater than τ_{RSRP} from the satellite, we handover the UEs to the satellite and shutdown the MBS.
4) **Cost and Constraint:** We compute the incurred cost and constraint violations based on (7) and (8).

B. Bandit-feedback Constrained Online Mirror Descent

We now give an overview of the BCOMD algorithm originally presented in [7], which works closely with the framework presented in Section IV-A to derive the optimal policy.

Firstly, by denoting $f_t \in [0,1]^n$ and $g_t \in [0,1]^n$ as the cost and constraint violation vectors, we can compute the expected loss and constraint violation at time t respectively as:

$$f_t(x) \triangleq f_t \cdot x, \quad g_t(x) \triangleq g_t \cdot x, \quad (14)$$

where $x \in \Delta_n$.

The algorithm leverages a Lagrangian function defined as:

$$\Psi(x, \lambda) \triangleq f_t(x) + \lambda g_t(x). \quad (15)$$

The first term of the Lagrangian function corresponds to the cost function, while the second term imposes a penalty for soft constraint violations using the Lagrange multiplier, which acts as a weighting factor. In the bandit-feedback setting, we derive unbiased estimators for the gradients of $f_t(x)$ and $g_t(x)$ as follows:

$$\tilde{f}_t = \frac{f_t(a_t, x_t)}{x_{t,a_t}} e_{a_t}, \quad \tilde{g}_t = \frac{g_t(a_t, x_t)}{x_{t,a_t}} e_{a_t}, \quad (16)$$

where a_t is the arm selected at time step t, and e_{a_t} denotes the unit basis vector corresponding to that arm. However, the unbounded variance of these estimators poses significant challenges in establishing reliable performance guarantees under the bandit regime. To address that, [7] proposes to use OMD, as it has proven to be effective in controlling the variance while claiming enhanced convergence speed compared to classic online gradient methods.

In OMD, the updates are first performed in the dual space and then projected back to the primal space via a mirror map (e.g. the negative entropy). Algorithm 1 outlines the iterative refining of the action distribution through an OMD framework. Indeed, at each iteration, the update direction is determined by combining a gradient estimate of the cost function with a weighted estimate of the constraint gradients (Lines 21-22). These weights are not fixed and are adaptively modified based on the accumulated constraint violations (Line 23). This dynamic adjustment allows the policy to effectively balance cost minimisation with constraint satisfaction over time. Note that the probability of selecting any action is maintained above a predefined threshold γ. The BCOMD algorithm, adapted to our framework, is outlined in Algorithm 1.

Algorithm 1: BCOMD - Network Optimisation

Data: Initial $x_1 = (1/n)_{a \in \mathcal{A}}$, $\lambda_1 = 0$, Mirror map $\Phi : \mathbb{R}^n \to \mathbb{R}$, learning rate $\eta > 0$, $\gamma \in [0, 1/n]$, $\Omega > 0$

for $t = 1, \ldots, T$ **do**

  Sample action $a_t \sim x_t$ // Run Framework

  Data: K UEs, L MBs, $\theta = [\varepsilon, \tau_\nu, \tau_{\text{RSRP}}, \alpha]$.

  Initialisation: Association done through max-RSRP;

  Compute the load for MBS;

  Redistribute the resources according to ε;

  UE association: for *all UEs* u **do**

    Associate UE u to MBS j^* such that:

    $j^* = \arg\max_j P_u(j)$ (13)

  Recompute the load for MBS;

  MBS Shutdown:

  for *all MBSs* j **do**

    if $\nu_j + \nu_{\text{sat}} \leq \tau_\nu$ **then**

      if *RSRP from satellite of all served UEs by MBS* j $\geq \tau_{\text{RSRP}}$ **then**

        Associate each UE to the satellite;

        Shutdown MBS j;

  Incur $f_t(a_t, x_t)$ and $g_t(a_t, x_t)$; // Bandit-feedback

  $\tilde{f}_t \leftarrow (f_t(a_t, x_t)/x_{t,a_t})\, e_{a_t}$; // Loss gradient estimate

  $\tilde{g}_t \leftarrow (g_t(a_t, x_t)/x_{t,a_t})\, e_{a_t}$; // Constraint gradient estimate

  $\tilde{\omega}_t \leftarrow (\Omega/x_{t,a_t}) e_{a_t}$; // Bias term

  $\tilde{b}_t \leftarrow \tilde{\omega}_t + \tilde{f}_t + \lambda_t \tilde{g}_t$; // Gradient for $\Psi(\cdot, \lambda_t)$

  $y_{t+1} \leftarrow (\nabla\Phi)^{-1} \left(\nabla\Phi(x_t) - \eta \tilde{b}_t \right)$; // Update primal action distribution

  $x_{t+1} \leftarrow \Pi_{\Delta_{n,\gamma}}(y_{t+1})$; // Project to feasible simplex

  $\lambda_{t+1} \leftarrow (\lambda_t + \mu g_t(a_t))_+$; // Update dual variable

V. Simulation results and analysis

In this section, we assess the performance of our framework over 24 hours, with the number of UEs varying at each hour, similar to [6]. Using a custom-built system-level simulator, we collected $7 \cdot 10^3$ snapshots of the network for each hour of the day, yielding a total of $168 \cdot 10^3$ samples. Then, we used the learned policy to sample an action for each hour of the day to evaluate the resulting performance. ε and τ_ν take values in $[0.25,\ 0.50,\ 0.75,\ 0.85,\ 0.90]$, while τ_{RSRP} and α take values in $[-80,\ -90,\ -100,\ -110,\ -120]$ and $[-3,\ -2,\ -1,\ 0,\ 1,\ 2,\ 3]$ respectively. Our study focuses on an area of approximately $2500\ \text{km}^2$, corresponding to the coverage area of an LEO satellite beam [11], and encompasses both urban and rural regions. Additionally, we assume that the LEO constellation employs Earth-fixed beams [9]. The UEs are deployed uniformly across the study area, with a higher density in the urban region compared to the rural area. Similarly, the terrestrial MBSs are arranged in a hexagonal grid layout in both urban and rural areas, with a higher density of MBSs in the urban area. Two benchmark configurations are provided to compare performances: the `3GPP-TN` scenario, where no satellite tier is present and the terrestrial tier is allocated a total bandwidth of 10 MHz, and the `3GPP-NTN` scenario, where the total bandwidth is allocated as per 3GPP specifications [11], with 30 MHz for the satellite tier and 10 MHz for the terrestrial tier. In both scenarios, each UE associates with the MBS according to the max-RSRP rule, and only inactive MBSs are shut down. The parameter ζ is set to be inversely proportional to the number of UEs in the network. Detailed simulation parameters are provided in Table I and are set based on [8], [9], [11]–[14].

Parameter	Value
Total Bandwidth W	40 MHz
Urban/Rural Inter-Site Distance	$500/1732\ m$
Number of Macro BSs	1776
Satellite Altitude [11]	600 km
Number of arms n	875
Terrestrial Max Tx Power per RE $p_{\max}$ [13]	17.7 dBm
Satellite Max Tx Power per RE $p_{\max}$ [11]	15.8 dBm
Antenna gain (Terrestrial) G_{T_X} [14]	14 dBi
Antenna gain (Satellite) G_{T_X} [11]	30 dBi
Shadowing Loss (Terrestrial) SF [8]	4 – 8 dB
Shadowing Loss (Satellite) SF [9]	0 – 12 dB
Line-of-Sight Probability (Terrestrial / Satellite)	Refer to [8] / [9]
White Noise Power Density	-174 dBm/Hz
Coverage threshold $\text{RSRP}_{\min}$	-120 dBm
Urban/Rural UEs distribution proportion	40%/60%
UE Antenna gain G_{UE} [9]	0 dBi

Table I: Simulation parameters.

A. UE Satisfaction Analysis

Firstly, we study the UE satisfaction constraint violation. To that end, Fig. 1 depicts the proportion of UEs who are not satisfied throughout the day for our framework as well as the two benchmarks mentioned previously. We notice straight away that the proportion of UEs unsatisfied is steady around 3 % throughout the day for `3GPP-TN`. This is explained by the fact that this setting does not include a satellite, leading to several cell-edge UEs being out of coverage and thereby unsatisfied. Conversely, `3GPP-NTN` and `COMD` bring this proportion down to nearly 0 % in low-traffic hours (0 AM - 9 AM) solely by the addition of the satellite. However, as the traffic demand increases, we notice that the proportion of unsatisfied UEs jumps to roughly 6 %. Indeed, the max-RSRP association does not consider the load of each cell, leading to the overload of the satellite and a deteriorated data-rate for served UEs. Nevertheless, our framework is able to improve on both benchmarks during high-traffic, as the optimal policy learns the best setting of the parameters θ (α in particular), leading to a more efficient load distribution and, consequently, a decrease in the number of unsatisfied UEs.

B. Network Performance Analysis

In this section, we analyse the network performance in terms of total achieved capacity, as well as the total TN energy consumption. To that end, Fig. 2 shows the evolution of the sum throughput (ST) throughout the day, while Fig. 3 shows the TN energy consumption. Since each UE has a specific demand, the ST is inherently bounded, and we cannot exceed this threshold, which limits the potential gains that can be observed. The gains are directly explained by the number of unsatisfied UEs in the network, as we see an average ST improvement in high-traffic hours of 1 % and 4 % compared

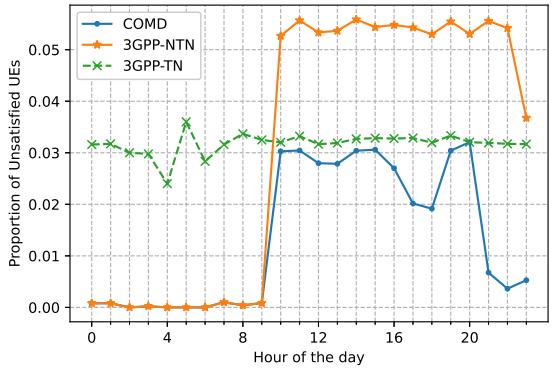

Figure 1: Daily satisfied UE proportion profile for various settings.

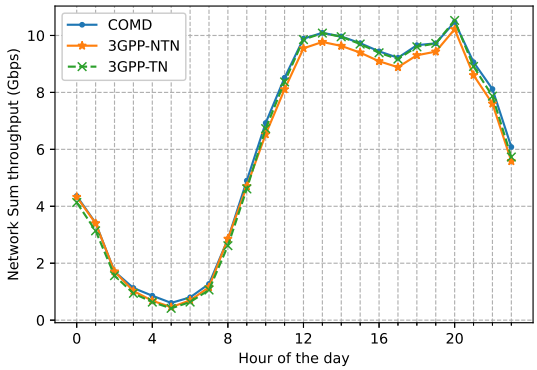

Figure 2: Daily network ST profile for various settings.

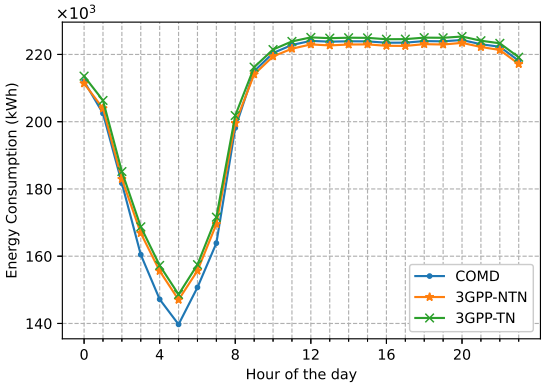

Figure 3: Daily TN energy consumption profile for various settings.

to 3GPP−TN and 3GPP−NTN, respectively. In low-traffic, the average ST gain soars up to 19 % and 10 %, respectively. In terms of energy consumption, we see an average decrease of roughly 5 % compared to both benchmarks in low traffic hours. Again, this is due to our policy learning the best configuration of θ (especially ε and τ_ν) that could facilitate the shutdown of terrestrial MBSs. In high-traffic scenarios, the energy consumption for COMD is slightly worse, as the emphasis is on load balancing. Indeed, the optimal policy selects a θ configuration which enables handovers to inactive MBSs, resulting in fewer MBSs shutdowns compared to the two benchmarks while maintaining a higher satisfaction rate, as seen in Section V-A.

VI. Conclusion

In this work, we proposed a novel framework for online optimisation in integrated TN-NTN, aimed at jointly improving UE satisfaction through load balancing and reducing TN energy consumption. By formulating the problem as a MAB and leveraging the BCOMD algorithm, our framework adaptively optimises a set of control parameters to select the most suitable system configuration in response to time-varying network conditions, striking a balance between enhanced capacity and energy efficiency. Through extensive simulations over a 24-hour period, we demonstrated that our approach significantly improves performance compared to standard 3GPP−TN and 3GPP−NTN benchmarks. Notably, our method reduces the proportion of unsatisfied users during peak hours, enables up to 19 % higher ST and 5 % lower energy consumption in low-traffic scenarios. Our future works will include a theoretical study on the dynamic regret bound that our algorithm can achieve if we change our mirror map to the Tsallis entropy, as well as exploring alternate, more robust estimators for the gradients.

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 Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 653–697, 2022.

[2] M. Giordani *et al.*, "Non-terrestrial networks in the 6g era: Challenges and opportunities," *IEEE Network*, vol. 35, no. 2, 2021.

[3] M. Benzaghta *et al.*, "Uav communications in integrated terrestrial and non-terrestrial networks," in *2022 IEEE GLOBECOM*, December 2022, pp. 1–6.

[4] K. Shen *et al.*, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1100–1113, 2014.

[5] H. Alam *et al.*, "Throughput and coverage trade-off in integrated terrestrial and non-terrestrial networks: an optimization framework," in *IEEE ICC Workshops*, 2023.

[6] ——, "Optimizing integrated terrestrial and non-terrestrial networks performance with traffic-aware resource management," 2025. [Online]. Available: https://arxiv.org/abs/2410.06700

[7] "Adversarial multi-armed bandits with constraints in dynamic environments," *Under Submission*, 2025.

[8] 3GPP TSG RAN, "TR 38.901, Study on channel model for frequencies from 0.5 to 100 GHz ," *V17.0.0*, March 2022.

[9] ——, "TR 38.811, Study on New Radio (NR) to support non-terrestrial networks," *V15.4.0*, September 2020.

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

[11] 3GPP TSG RAN, "TR 38.821, Solutions for NR to support non-terrestrial networks (NTN)," *V16.1.0*, May 2021.

[12] ——, "TR 36.763, Study on NB-IoT / eMTC support for NTN," *V17.0.0*, June 2021.

[13] ——, "TR 36.814, E-UTRA; Further advancements for E-UTRA physical layer aspects," *V9.2.0*, March 2017.

[14] ——, "TR 36.931, E-UTRA; Radio Frequency (RF) requirements for LTE Pico Node B," *V17.0.0*, March 2022.