

# Online Energy-efficient Resource Allocation in Integrated Terrestrial and Satellite 6G Networks

A. Mesodiakaki<sup>\*,†</sup>, M. Gatzianas<sup>\*,†</sup>, C. Bratsoudis<sup>\*,†</sup>, G. Kalfas<sup>\*,†</sup>, C. Vagionas<sup>\*,†</sup>, R. Maximidis<sup>\*,†</sup>  
A. Antonopoulos<sup>□</sup>, N. Pleros<sup>\*,†</sup>, and A. Miliou<sup>\*,†</sup>

<sup>\*</sup>Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

<sup>†</sup>Center for Interdisciplinary Research and Innovation, Thessaloniki, Greece

<sup>□</sup> Nearby Computing, Barcelona, Spain

Emails: {amesodia, mgkatzia, cbrat, gkalfas, chvagion, maximidis, npleros, amiliou}@csd.auth.gr,  
aantonopoulos@nearbycomputing.com

**Abstract**—In this paper, we jointly study the real-time user association, traffic routing and x-Network Function (xNF) placement problem in an integrated Terrestrial and Satellite 6G Network (TN-SN), with the aim of maximizing the network energy efficiency and user acceptance ratio. We formulate the aforementioned problem as a Mixed Integer Linear Program (MILP), considering various capacity, power and flow conservation constraints for the integrated network, while also meeting the specific service requirements of each user. To tackle the increased complexity of the optimal solution, we also develop an efficient heuristic (named as TERA). Through extensive simulations, TERA demonstrates a notable superiority in energy efficiency compared to the current State-of-the-Art (SoA), achieving up to 85% of the optimal with up to 87% lower execution time even under challenging traffic load conditions.

**Index Terms**—6G, Cloud, Edge, Fog, Geostationary Earth Orbit (GEO), Green Deal, Heuristic, Low Earth Orbit (LEO), Medium Earth Orbit (MEO), Resource Allocation, Service Function Chain (SFC), Space, x-Network Function (xNF).

## I. INTRODUCTION

The ever-increasing traffic growth of mobile data is about to push the capabilities of current networks to their limits in the upcoming years. Considering that, 6<sup>th</sup> generation (6G) networks are about to unify the terrestrial and satellite domain so as to offer a wider selection of access points. Thereby, the transformation of current Radio Access Network (RAN) is anticipated, leading to a 3D architecture comprising “a network of networks” [1], with the expanded network offering even more front/mid/back-haul (X-haul) link capabilities, through, e.g., Inter-Satellite Links (ISLs).

In this context, resource allocation becomes even more challenging, due to the additional satellite base stations (BSs) and X-haul links, requiring new channel modeling to support them. In parallel, the User Equipment (UE) Service Function Chains (SFCs) and their associated x-Network Functions (xNFs), which can be network functions of any type, e.g., physical (PNF), virtual (VNF) or cloud-native (CNF), have to be deployed ensuring that the capacity and computational constraints of each node are met and that the xNFs are executed in the same order as in the SFC without violating any link constraints [2]. xNFs can be placed either to BSs (in both terrestrial and satellite domain) collocated with Multi-Access Edge Computing (MEC) capabilities for lower latency,

or to farther cloud computing nodes with greater capabilities but with higher latency or to fog computing nodes in between offering a trade-off between computing capacity and latency. For wireless links, the use of the mmWave band is favored, while for the ISL links higher frequencies, i.e., optical connections from 20 to 375 THz [3], will be preferred exploiting the lack of atmosphere in space.

Due to the network densification dictated by the unprecedented data traffic growth, 6G networks are forced to maximize their energy efficiency, mainly to: a) reduce the associated Operational Expenditure (OPEX) of the network and b) decrease the energy consumption, thus leading to sustainable networks. As a result, due to these imminent additions and modifications to future mobile networks, strategies for online resource allocation should be designed that: i) consider different resource types, such as computational, communication and storage, as well as 6G technologies, e.g., THz bands, and their constraints, ii) enable real-time decision-making by optimizing the algorithmic computational complexity, iii) guarantee end-to-end (E2E) optimality by taking into account the E2E latency and data rate needs and iv) maximize the network energy efficiency. In a nutshell, integrated Terrestrial and Satellite 6G Networks (TN-SNs) call for the development of energy-friendly solutions for online user association, traffic routing and xNF placement, while guaranteeing the QoS of the UE and the SFC chaining.

### A. Related work and Contribution

One of the most anticipated key technologies of 6G will be the integration of TN-SNs, which has recently received a great research interest. In [4]–[11], the authors are focusing on a subset of problems, proposing, e.g., a power and frequency resource allocation scheme in TN-SNs [4], a disjoint user association and spectrum allocation solution for the TN and SN domains targeting rate maximization [5] or aim to reduce the latency of task processing for latency-dependent tasks [6]. In [7], user association, bandwidth assignment, and power allocation in the uplink is studied, while in [8], VNF placement in SNs is addressed. In [9], the authors compare different algorithms for VNF placement in TN-SNs, while [10] studies communication and computation resource allocation in

a SN with MEC-enabled Low Earth Orbit satellites (LEOs), targeting at energy consumption minimization. Power and delay-aware VNF placement and traffic routing strategies are proposed in [11]. In parallel, these papers do not employ all satellite orbits in their study, as they consider either only LEO [4], [5], [7], [10], only Medium Earth Orbit (MEO) [8] or only LEO and MEO [6], [9] in their system model or do not consider the satellite domain at all [11], [12], thus limiting the full potential of integrated TN-SNs, while simplifying the considered problem. As a result, their high performance in solving jointly the aforementioned problems (and not only a subset) in such complex 3D networks, comprising LEOs, MEOs, and Geostationary Earth Orbit satellites (GEOs) with different capabilities and constraints, cannot be guaranteed.

To that end, unlike the State-of-the-Art (SoA), in this paper, we substantially extend our previous work in [12] focusing solely on TNs, by considering the problem of online user association, traffic routing and xNF placement in a complex 3D heterogeneous TN-SN, comprising both terrestrial (gNB and SCs) and Satellite (LEO, MEO and GEO) BSs. We adopt a distributed 3D computing infrastructure across the compute continuum (edge, fog, cloud), each with different capabilities, which together with different BS types, and Access Network (AN) and X-haul links (being both wireless and fiber), are incorporated in the problem formulation, taking into account their power and capacity limitations. In addition, a detailed channel and power model have been developed for both the TN and SN, accounting for the fluctuations of the wireless channel. A highly complex optimal solution is derived as well as a low complexity algorithm (named as TERA), which is shown to achieve a good trade-off between energy efficiency and complexity, while significantly outperforming the SoA.

The structure of the paper is as follows: Section II presents the system model, the statement of the problem as well as the Optimal solution, while Section III describes the proposed heuristic algorithm (TERA). In Section IV, the proposed solutions (both Optimal and TERA) are compared to the SoA, while in Section V we conclude the paper. We denote sets as  $\mathcal{V}$ , indicator functions as  $\mathbb{I}[\cdot]$  and equality by definition as  $\triangleq$ .

## II. SYSTEM MODEL AND PROBLEM STATEMENT

We consider an integrated TN-SN under a 3D network deployment as depicted in Fig. 1, while accounting for the spatio-temporal dynamics of the Service Requests (SRs) and previously allocated resources. We model the network as a directed graph  $\mathcal{G}(\mathcal{V}, \mathcal{E})$ , with  $\mathcal{V}$  the set of non-UE nodes in the RAN and Core segments (both TN and SN) and  $\mathcal{E}$  the set of edges between them, as depicted in Fig. 1. Nodes in  $\mathcal{V}$  can be i) terrestrial gNBs and/or Small Cells (SCs), ii) satellite-based gNBs (we hereafter refer to all TN/SN gNBs and SCs as BSs), and iii) intermediate nodes offering forwarding or specialized middlebox services (e.g., firewall, gateway etc.). For each link  $e = (u, v) \in \mathcal{E}$ , we denote its capacity with  $c_e$  and its introduced delay (i.e., the sum of transmission, propagation and processing delay) with  $d_e$ . We partition  $\mathcal{E}$  into the sets of wired and wireless links,  $\mathcal{E}_{wi}$ ,  $\mathcal{E}_{wl}$ , respectively, and

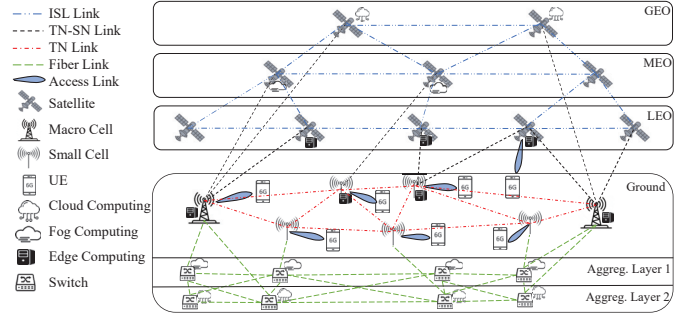


Fig. 1. System model example of an integrated terrestrial and satellite network leveraging a distributed compute continuum across the different layers.

denote with  $\mathcal{V}_{wi}$  the set of nodes with at least one incident (incoming or outgoing) wired link (hereafter referred to as “switches”, being the typical enablers of such links). We also write  $h(e) = u$ ,  $t(e) = v$  for the head and tail, respectively, of directed link  $(u, v)$ . Let  $\mathcal{S}$  be the set of subscribing UEs, where UE  $s \in \mathcal{S}$  can connect to a BS (either TN or SN) in set  $\mathcal{B}^{(s)} \subseteq \mathcal{V}$ . We focus on downlink and define the set of access links  $\mathcal{E}_{AN} \triangleq \{(b, s) : b \in \mathcal{B}^{(s)}, s \in \mathcal{S}\}$  and  $\tilde{\mathcal{E}} \triangleq \mathcal{E} \cup \mathcal{E}_{AN}$ . Let  $\bar{N}_b^{(RB)}$  be the number of Resource Blocks (RBs) that BS  $b \in \mathcal{B} \triangleq \bigcup_{s \in \mathcal{S}} \mathcal{B}^{(s)}$  can offer, in aggregate, to its served UEs.

We denote with  $\mathcal{V}_h \subseteq \mathcal{V}$  the set of nodes capable of hosting xNFs, where node  $x \in \mathcal{V}_h$  has  $c_x$  computational resources (measured in GFLOPS). Let  $\mathcal{F}$  be the set of available xNFs, where multiple instances of a given xNF in the same or different nodes are allowed. We define xNF  $f \in \mathcal{F}$  as  $f \triangleq (I_f, P_f, R_f, D_f)$ , where  $I_f$  identifies xNF functionality,  $P_f > 0$  is the xNF’s traffic processing capability (in Mbps),  $R_f > 0$  is the amount of CPU resources (in GFLOPS) consumed by the xNF, and  $D_f > 0$  is the delay incurred on an individual data packet when processed by  $f$ .

To capture the processing requirements for the packets of each requested flow, we introduce the concept of an SFC  $\pi \triangleq \langle f_1^{(\pi)}, \dots, f_{N_\pi}^{(\pi)} \rangle$  as an ordered tuple of xNFs. We denote with  $\mathcal{C}$  the set of all SFCs and impose the rule that a flow is successfully served only if its traffic passes through the xNFs in its corresponding SFC  $\pi$  in exactly the specified order. We write  $f \rightsquigarrow \pi$  to denote that  $\pi$  contains xNF  $f$  and define  $\mathcal{R}_\pi \triangleq \{f \in \mathcal{F} : f \rightsquigarrow \pi\}$ .  $\rho$  is equivalently described by a directed graph  $\mathcal{G}^{(\rho)}(\mathcal{V}^{(\rho)}, \mathcal{E}^{(\rho)})$ , where  $\mathcal{V}^{(\rho)}$  contains the xNFs in  $\mathcal{R}_\rho$  as virtual nodes and  $\mathcal{E}^{(\rho)}$  describes the respective order of the xNFs. For any virtual edge  $e' \in \mathcal{E}^{(\rho)}$ , the xNFs at the head, tail of  $e'$  are denoted as  $h(e')$ ,  $t(e')$ .

Each UE  $s$  issues, upon arriving at the network at a random epoch  $\tau_s$ , a SR  $q_s \triangleq (\beta_{q_s}, r_{q_s}, \delta_{q_s}, \pi_{q_s}, \eta_{q_s})$ , with  $\beta_{q_s} \in \mathcal{V}$  the source node of  $q_s$  (destination being  $s$  itself),  $r_{q_s} > 0$  the E2E required throughput,  $\delta_{q_s} > 0$  the required E2E latency threshold,  $\pi_{q_s} \in \mathcal{C}$  the requested SFC and  $\eta_{q_s} > 0$  the duration of the service that must be offered to the UE (i.e.,  $q_s$  must be continuously served in the time interval  $[\tau_s, \tau_s + \eta_{q_s}]$ ). We allow sharing of xNF instances among multiple SFCs under arbitrary continuous distributions of the SR inter-arrival intervals and denote with  $N_{b,s}^{(RB)}$  the number of RBs allocated

by BS  $b$  to  $s$  to achieve rate  $r_{q_s}$  on link  $(b, s)$ . Similar to [12], we assume that constant power  $p_b^{(RB)}$  is assigned to each RB of BS  $b$ . Defining  $\mathcal{A}(t) \triangleq \{s \in \mathcal{S} : t - \eta_{q_s} < \tau_s < t\}$  as the set of SRs that have arrived before  $t$  and are still served at  $t$ , we denote with  $\tilde{\tau}(k)$ , for each integer  $k > 0$ , the  $k$ -th earliest SR epoch and with  $\tilde{\gamma}(k) \triangleq \{s : \tau_s = \tilde{\tau}(k)\}$  the respective UE issuing the SR. Hence, all SRs in  $\mathcal{N}(\tilde{\tau}(k))$  are still served at  $\tilde{\tau}(k)$ . We denote  $\mathcal{N}_k \triangleq \mathcal{N}(\tilde{\tau}(k))$  and assume that all network and computational resources allocated for the SR of UE  $s$  arriving at  $\tau_s$  remain “reserved” for this UE until time  $\tau_s + \eta_{q_s}$  (i.e., when the service is no longer needed). In other words, no “migration” of allocated xNFs and resources is allowed, therefore resources are released only when all services using them have expired. This is motivated by the assumption that active services can tolerate no xNF redeployment during SR reconfiguration, as this would lead to service disruption.

We next formulate an optimization problem to “*minimize total network power consumption while defining the placement of the requested xNFs and E2E routing paths to satisfy an arbitrary number and type of SRs dynamically issued by UEs, given that no migration is permitted*”.

#### A. Power consumption model and problem formulation

We consistently use the following indices:  $s \in \mathcal{S}$ ,  $b \in \mathcal{B}$ ,  $x \in \mathcal{V}_h$ ,  $\hat{x} \in \mathcal{V}_h \cup \mathcal{S}$ ,  $f \in \mathcal{F}$ ,  $u, v, w \in \mathcal{V}$ ,  $m, n \in \mathcal{V}_{wi}$ . For each SR instance  $k$ , a new optimization problem is derived with decision variables  $a_{s,b} \triangleq \mathbb{I}[\text{UE } s \text{ attaches to BS } b]$ ,  $\varphi_{\hat{x},f,q_j} \triangleq \mathbb{I}[\text{xNF } f \text{ in SFC } \pi_{q_j} \text{ is deployed on node } \hat{x}]$  and  $\vartheta_e^{e',q_j} = \vartheta_{u,v}^{e',q_j} = \mathbb{I}[e \text{ belongs to the physical path in } \mathcal{G} \text{ onto which virtual link } e' \in \mathcal{E}^{(\rho_{q_j})} \text{ is mapped for SFC } \pi_{q_j}]$  for physical link  $e = (u, v) \in \mathcal{E}$ . Also,  $N_{f,\hat{x}}$  denotes the number of instances of xNF  $f$  deployed on  $\hat{x}$ , while  $\hat{a}_{s,b}$ ,  $\hat{\varphi}_{\hat{x},f,q_j}$ ,  $\hat{\vartheta}_e^{e',q_j}$  are the decision variables for the previous SR instance  $k-1$  (for  $k=1$ , it holds  $\hat{a}_{s,b} = \hat{\varphi}_{\hat{y},f,q_j} = \hat{\vartheta}_e^{e',q_j} = 0$ ). Note that, when solving for SR  $k$ , variables  $\hat{a}_{s,b}$ ,  $\hat{\varphi}_{\hat{x},f,q_j}$ ,  $\hat{\vartheta}_e^{e',q_j}$  become parameters and not decision variables of SR  $k$ . We also omit the explicit dependence of the above quantities on  $k$  to keep the notation manageable.

Let  $\zeta_x$  the Boolean variable indicating whether xNFs are actually deployed on node  $x$  (necessitating computational resources). The CPU of node  $x$  consumes power  $P_x^{(CPU)} = P_x^{(CPU,0)}\zeta_x + (P_x^{(CPU,m)} - P_x^{(CPU,0)})U_x$ , with  $P_x^{(CPU,m)}$ ,  $P_x^{(CPU,0)}$  the CPU power at maximum and idle condition and  $U_x \triangleq \sum_{f \in \mathcal{F}} \frac{N_{f,\hat{x}} R_f}{c_x}$  the CPU utilization [11]. For a node  $n \in \mathcal{V}_{wi}$  with wired links, we introduce the decision variable  $\varpi_n \triangleq \mathbb{I}[\text{node } n \text{ has active incident wired links carrying traffic}]$ . Similarly, for wired links  $e = (n, m) \in \mathcal{E}_{wi}$ , we define decision variables  $z_{n,m} \triangleq \mathbb{I}[e \text{ actually carries traffic, in either link direction}]$  and  $w_{n,m} \triangleq \mathbb{I}[e \text{ carries traffic from } n \text{ to } m]$ , so that  $z_{m,n} \geq w_{m,n}$ ,  $w_{n,m}$ , which impose the consistency conditions in (1), for a sufficiently large constant  $C_1 > 0$ .

$$\sum_{e' \in \mathcal{E}^{(\pi_{q_j}(k))}} \vartheta_e^{e',q_j(k)} + \sum_{s \in \mathcal{N}_k} \sum_{e' \in \mathcal{E}^{(\pi_{q_s})}} \hat{\vartheta}_e^{e',q_s} \leq C_1 w_e, \quad \forall e \in \mathcal{E}_{wi},$$

$$\sum_{e \in \mathcal{E}_{wi}: h(e)=n} w_e + \sum_{e \in \mathcal{E}_{wi}: t(e)=n} w_e \leq C_1 \varpi_n, \quad \forall n \in \mathcal{V}_{wi}. \quad (1)$$

Switch  $n$  consumes power  $P_n^{(sw)} = P^{(sw,0)}\varpi_n + P_{port}$ .  $\sum_{m \in \mathcal{V}_{wi}: (n,m) \in \mathcal{E}_{wi}} z_{n,m}$ , where  $P^{(sw,0)}$  is the switch’s idle power and the sum accounts for the switch’s active wired links, with  $P_{port}$  the consumed power per active port [11].

We employ the power model of [12] for X-haul (i.e., non-access) wireless links, where link  $e \in \mathcal{E}_{wl}$  consumes power  $P_e^{(Xh)} = N_e^{(RF)}(\chi_e P_e^{(Xh,0)} + \Delta_e^{(Xh)} F(\ell_e))$ , where  $N_e^{(RF)}$  is the number of the link’s transmitter (TX) RF chains,  $P_e^{(Xh,0)}$  is TX idle power,  $\chi_e \triangleq \mathbb{I}[e \text{ actually carries traffic}]$ ,  $\ell_e$  is the load-dependent link utilization,  $\Delta_e^{(Xh)}$  is a parameter describing TX electronics, and  $F(\cdot)$  is a piecewise-linear function representing the nonlinear throughput/power relation. In the BS-to-UE links, BS  $b$  consumes power  $P_b^{(BS)} = N_b^{(RF)}(\mu_b P_b^{(BS,0)} + \Delta_b^{(BS)} p_b^{(RB)}(a_{\tilde{\gamma}(k),b} N_{b,\tilde{\gamma}(k)}^{(RB)} + \sum_{s \in \mathcal{N}_k} \hat{a}_{s,b} N_{b,s}^{(RB)}))$ , with  $\mu_b \triangleq \mathbb{I}[b \text{ serves at least one UE}]$  and  $N_b^{(RF)}$ ,  $P_b^{(BS,0)}$ ,  $\Delta_b^{(BS)}$  having the same semantics as the X-haul links. Following [13], we employ auxiliary Boolean variables  $\sigma_e$ , for  $e \in \mathcal{E}_{wl}$ , and  $\mu_b, \nu_b$ , for  $b \in \mathcal{B}$ , to convert the activation constraints into the linear form of (2), where  $C_2 > 0$  is a sufficiently large constant.

$$\begin{aligned} \chi_e + C_2 \sigma_e &\geq 1, \quad \forall e \in \mathcal{E}_{wl}, \quad \mu_b + C_1 \nu_b \geq 1, \quad \forall b, \\ 1 - C_2 \chi_e &\leq \sum_{e' \in \mathcal{E}^{(\pi_{q_j}(k))}} \vartheta_e^{e',q_j(k)} + \sum_{s \in \mathcal{N}_k} \sum_{e' \in \mathcal{E}^{(\pi_{q_s})}} \hat{\vartheta}_e^{e',q_j} \\ &\leq C_2(1 - \sigma_e), \quad \forall e \in \mathcal{E}_{wl}, \\ 1 - C_1 \nu_b &\leq a_{\tilde{\gamma}(k),b} + \sum_{s \in \mathcal{N}_k} \hat{a}_{s,b} \leq C_2(1 - \nu_b), \quad \forall b. \end{aligned} \quad (2)$$

We impose the following constraints (along with basic consistency conditions, which are omitted due to space limitations):

$$\begin{aligned} \sum_{x \in \mathcal{V}_h} \varphi_{x,f,q_{\tilde{\gamma}(k)}} &= 1, \quad \forall f \in \mathcal{R}_{q_{\tilde{\gamma}(k)}}, \\ \sum_{b \in \mathcal{B}^{(\tilde{\gamma}(k))}} a_{\tilde{\gamma}(k),b} &= 1, \quad a_{\tilde{\gamma}(k),b'} = 0, \quad \forall b' \notin \mathcal{B}^{(\tilde{\gamma}(k))}, \end{aligned} \quad (3)$$

$$\begin{aligned} \varphi_{x,f,q_{\tilde{\gamma}(k)}} r_{q_{\tilde{\gamma}(k)}} + \sum_{s \in \mathcal{N}_k: f \rightsquigarrow \pi_{q_s}} \hat{\varphi}_{x,f,q_s} r_{q_s} &\leq N_{f,x} P_f, \quad \forall f, \forall x, \\ \sum_{f \in \mathcal{F}} N_{f,x} R_f &\leq c_x, \quad \sum_{f \in \mathcal{F}} N_{f,x} \leq C_1 \zeta_x, \quad \forall x, \end{aligned} \quad (4)$$

$$\begin{aligned} \sum_{e' \in \mathcal{E}^{(\pi_{q_j}(k))}} \vartheta_e^{e',q_j(k)} r_{q_{\tilde{\gamma}(k)}} + \sum_{s \in \mathcal{N}_k, e' \in \mathcal{E}^{(\pi_{q_s})}} \hat{\vartheta}_e^{e',q_s} r_{q_s} &\leq c_e, \quad \forall e, \\ a_{\tilde{\gamma}(k),b} N_{b,\tilde{\gamma}(k)}^{(RB)} + \sum_{s \in \mathcal{N}_k} \hat{a}_{s,b} N_{b,s}^{(RB)} &\leq \bar{N}_b^{(RB)}, \quad \forall b, \end{aligned} \quad (5)$$

$$\begin{aligned} \sum_{v:(u,v) \in \tilde{\mathcal{E}}} \vartheta_{u,v}^{e',q_s} - \sum_{w:(w,u) \in \tilde{\mathcal{E}}} \vartheta_{w,u}^{e',q_s} &= \varphi_{u,h(e'),q_s} - \varphi_{u,t(e'),q_s}, \\ s &= \tilde{\gamma}(k), \quad \forall e' \in \mathcal{E}^{(\pi_{q_s})}, \quad \forall u \in \tilde{\mathcal{V}}, \end{aligned} \quad (6)$$

$$\begin{aligned}
& \sum_{x \in \mathcal{V}_h} \sum_{f \in \mathcal{R}_{q_{\tilde{\gamma}(k)}}} \varphi_{x,f,q_{\tilde{\gamma}(k)}} D_f + \sum_{e \in \mathcal{E}} \sum_{e' \in \mathcal{E}^{(\pi_{q_{\tilde{\gamma}(k)}})}} \vartheta_{e'}^{e',q_{\tilde{\gamma}(k)}} D_e \\
& + \sum_{b \in \mathcal{B}(\tilde{\gamma}(k))} a_{\tilde{\gamma}(k),b} D(b,\tilde{\gamma}(k)) \leq \delta_{q_{\tilde{\gamma}(k)}}, \quad (7)
\end{aligned}$$

where (3) requires that the xNFs of each requested SFC are actually deployed and the arriving UE is served by exactly one BS. Eq. (4) make sure the deployed xNF instances on a node do not exceed the node's processing capabilities for the current traffic (including previous SRs still running), while computational nodes are only deployed as needed. Link capacity and RB constraints are captured in (5), while (6) is a flow routing condition which forces the packets of the new SR's SFC to pass through the corresponding xNFs in the correct order. Eq. (7) captures the new SR's E2E delay target.

Hence, for each SR instance  $k = 1, 2, \dots$ , the total power consumed by the network is  $P_{tot} \triangleq \sum_{n \in \mathcal{V}_{wi}} P_n^{(sw)} + \sum_{x \in \mathcal{V}_h} P_x^{(CPU)} + \sum_{e \in \mathcal{E}_{wl}} P_e^{(Xh)} + \sum_{b \in \mathcal{B}} P_b^{(BS)}$  and we formulate our problem as a sequence of MILPs (for each instance  $k$  a separate MILP needs to be solved).

$$\begin{aligned}
\text{For each instance } k : \quad & \text{minimize } P_{tot}, \\
& \text{s.t. } (1)-(7) \quad (8)
\end{aligned}$$

Solving (8) is computationally expensive due to the NP-hard nature of the MILP problem. Therefore, we also propose a low-complexity heuristic algorithm and evaluate its performance compared to the optimal solution of (8).

### III. TERA: ONLINE INTEGRATED TERRESTRIAL AND SATELLITE ENERGY-EFFICIENT RESOURCE ALLOCATION HEURISTIC ALGORITHM

In this section, a heuristic algorithm named as TERA, which studies the joint user association, traffic routing and xNF placement problem, is proposed, aiming to maximize user acceptance ratio as well as energy efficiency. As per Fig. 2, TERA is split into two main steps: a) first the algorithm decides upon the user association and traffic routing path and then, b) it places the xNFs of the SFC required by each UE, in the exact order specified by the SFC.

In the first step, every time a new UE SR arrives, TERA constructs a weighted graph and examines all available paths from the source to the destination based on their power consumption. In each path, all feasible wireless and fiber X-Haul transport links are included, as well as the AN link between the serving BS and the UE. The shortest-weighted path, i.e., with the minimum power consumption, is then selected to satisfy the UE demands, as long as the capacity and delay constraints are not violated. In case of a violation, TERA selects the next available shortest path, with no constraint violations and proceeds or, otherwise, if there is no other path to select, TERA blocks the UE and checks for new SR arrivals.

Once a path has been selected, TERA moves to the next step, i.e., the xNF placement. In order to place each xNF of the requested SFC in the available nodes specified by the selected path, the nodes are being sorted by a parameter denoted by  $\Omega$ ,

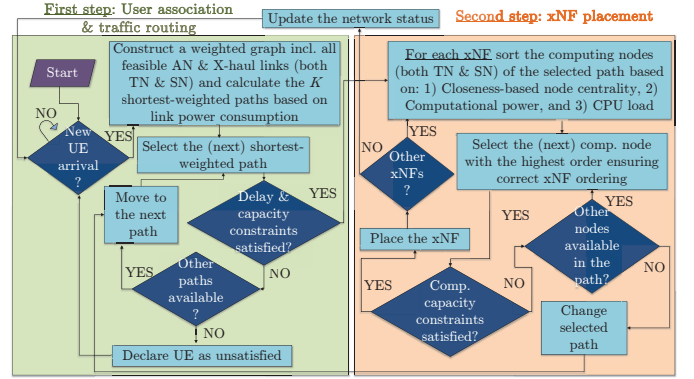


Fig. 2. Operation flowchart of TERA: Proposed Online Heuristic for Energy-efficient User association, Traffic Routing and xNF placement.

which consists of the node's closeness centrality, the maximum computational capacity and the load of CPU. As for the CPU load, three values are allowed: a) 1 (high priority) when the node has enough computational capacity and can host the xNF without the need for a new xNF instance initiation, b) 0.1 (low priority) when the node has enough computational capacity but needs a new instance initiation to host the xNF and c) 0 (no priority) when the node cannot host the xNF. When sorting is finished, TERA selects the highest ranked node and places the xNF, provided that the computational constraints are satisfied. If such a node cannot be found, TERA returns to the first step and selects the next shortest path, while repeating the process in the second step until all xNFs are placed or there are no other available paths, in which case it blocks the UE. If all xNFs from the SFC are placed, TERA updates the network state and awaits for new SR arrivals, repeating the same steps.

### IV. PERFORMANCE EVALUATION

We have executed extensive simulations (5 BS distribution scenarios with 10 UE hotspot traffic distribution snapshots each) in MATLAB R2022b for both the proposed heuristic and the reference algorithm, while the optimal solution is developed in IBM CPLEX. For the TN, without loss of generality, we focus on a gNB sector located at Thessaloniki, Greece (coordinates: [40.63, 22.94]) with 500 m radius, which is overlaid with two SC clusters, each one consisting of 4 SCs uniformly distributed in a 100 m radius from the cluster center [12]. We also assume mmWave backhaul links among TN BSs that are less than 200 m apart. The gNB and a randomly chosen SC per cluster have fiber connection to the transport network, which is divided into two levels, one corresponding to the fog and another to the cloud, with 4 randomly chosen nodes per layer being fiber-connected with each other. As shown in Fig. 3, where the different BSs' coordinates are depicted, for the sake of simplicity, we consider for the SN only the LEO and GEO satellites that have LoS during all the UE service duration, so that handovers are avoided for the time period studied. To that end, 2 Iridium, 28 OneWeb, 104 Starlink and 2 GEO satellites (HellasSat 3, 4) are considered, whose orbital elements were retrieved from [14]. For the TN AN, orthogonal

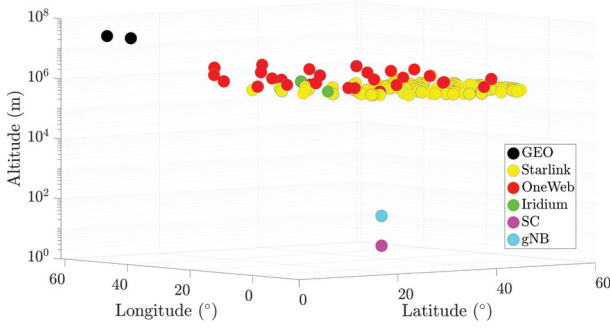


Fig. 3. Simulation setup example depicting both the terrestrial and satellite base station positions.

TABLE I  
SFC DETAILS [11]

SFC specifications				
Type	Rate (Mbps)	Latency (ms)	Duration (sec)	Share (%)
Web	[0.6-1]	500	20	20
VoIP	[0.384-0.64]	100	100	20
Streaming	[5-24]	100	360	39
Gaming	[0.24-0.5]	60	300	6
Ultra RT AI/ML	[15-25]	1	40	15

channels are employed between the gNB and the SCs (100 PRBs for each TN BS). However, SCs belonging to different clusters reuse the same frequencies and thus may interfere to each other. For the TN-SN and ISL links, the interference is considered negligible due to the high frequencies used, which attenuate greatly in other directions than the direct path.

For illustration purposes, 5 service types (SFCs) are considered, each one being an ordered sequence of VNFs (although any xNF could be employed) with specific data processing capacities and requirements, as explained in [12]. The details of each SFC are summarized in Table I, including the requested rate range, the E2E latency budget, the service duration and the percentage of UEs of this SFC type. The rest of the simulation parameters are summarized in Table II, while the CPU cores and power consumption values for the computational nodes of [12] are employed for the edge, fog and cloud both in the TN and SN domain. For the fiber links, the power consumption of a switch in idle state is  $P_{idle}^{(sw)}=315$  W, while the consumed power per port is  $P_{port}=7$  W and the fiber link rate 10 Gbps. Finally, a packet length of 1.5 KB is assumed.

We compare the proposed optimal and heuristic solutions with the SoA [11], which has been properly adapted to the studied scenarios for a fair comparison. As for the SoA, it first performs the VNF placement (criterion: highest betweenness centrality) and then the traffic routing (criterion: lowest delay) so that the E2E latency constraint is not violated. As the SoA does not take into account the user association problem, we apply the default criterion i.e., the UEs are connected to the BS from which they receive the highest signal power.

#### A. Simulation results

In Fig. 4 and 5, the energy efficiency (bits/Joule) and the computational time (s) are demonstrated, respectively, for the

TABLE II  
SIMULATION PARAMETERS [11], [12], [15]

Part of the network	Access Network			Backhaul Network		
	TN		SN	TN	TN-SN	ISL
	gNB	SC				
$f$ (GHz)	2		28	60	28	200000
$BW$ (GHz)	0.02		0.4	0.2	0.6	20
$N_{TRX_i}$	8	4	32	64	32	128
$P_{0_i}$ (W)	130	6.8	3.9	3.9	3.9	3.9
$\Delta_{p_i}$	4.7	4	4.5	5	4.5	10

proposed work compared to the SoA for various traffic load conditions, i.e., number of UEs in the system and arrival rates ( $1/\lambda$ ). As demonstrated, for high UE arrival rates ( $1/\lambda=1$ ), i.e., on average 1 UE arrival per second, the energy efficiency of both the proposed algorithm and the SoA increases, as the number of concurrent UEs in the system increases, and so does the total rate (please note that the total throughput increases at a higher rate than power consumption). On the contrary, for low arrival rates ( $1/\lambda=10$ ), i.e., on average 1 UE arrival per 10 sec, due to the long inter-arrival times, some UEs have exited the system, once their service time is ended, thus resulting in lower rate and consequently lower energy efficiency. TERA achieves up to 85% of the Optimal solution with up to 87% lower computational time.

In all cases, the user acceptance ratio is equal to 1 for all algorithms, i.e., no UE is being blocked. Focusing on the execution time, as depicted in Fig. 5, it increases with higher arrival rate for all algorithms, as expected. This stems from the fact that, although SRs always arrive in batches of 1 (i.e., 1 UE per SR), in case of higher arrival rate, the number of concurrent UEs in the system increases, resulting in a decreased number of feasible solutions, and consequently to higher computational time for all algorithms. TERA achieves an excellent trade-off between energy efficiency and complexity. Please also note that these results only illustrate the **minimum** achievable computational gains of TERA, due to the assumed 1 UE per SR and, in case an SR involves more UEs, higher computational gains compared to the optimal solution are expected.

Compared to the SoA, the proposed heuristic demonstrated significant energy efficiency gains in all cases, reaching up to 62% higher performance than the SoA, with low execution time. This is because, unlike the SoA, TERA jointly considers the user association problem, resulting in higher flexibility, at the expense of slightly higher complexity than the SoA. In particular, in the SoA, the serving BSs are already determined based on the best SINR, and then the optimal VNF placement and traffic routing are performed. This is supported by Fig. 6, which provides the power break-down of all algorithms under low ( $N = 10$ ) and high ( $N = 40$ ) traffic conditions for the most demanding case, i.e.,  $1/\lambda = 1$ . As can be observed, Optimal and TERA favor the selection of SCs and LEOs as serving BSs and not the gNB (contrary to the SoA), thus achieving much lower power consumption. We also note that the Optimal and TERA power consumption scales better than the SoA with increasing load (TERA still achieves 75% of the Optimal energy efficiency even in the most highly loaded

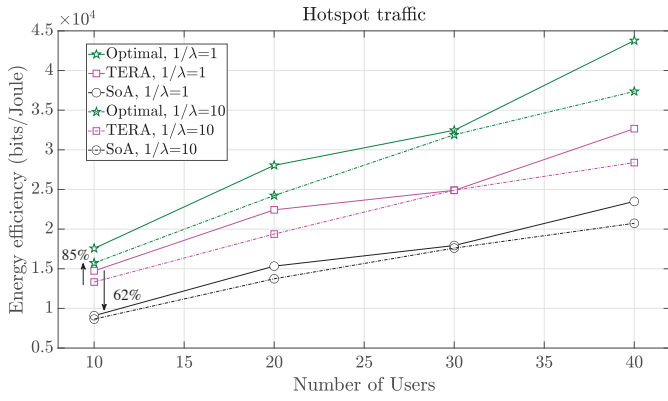


Fig. 4. Energy efficiency (bits/Joule) of all algorithms for different traffic load conditions and UE arrival rates ( $1/\lambda$ ).

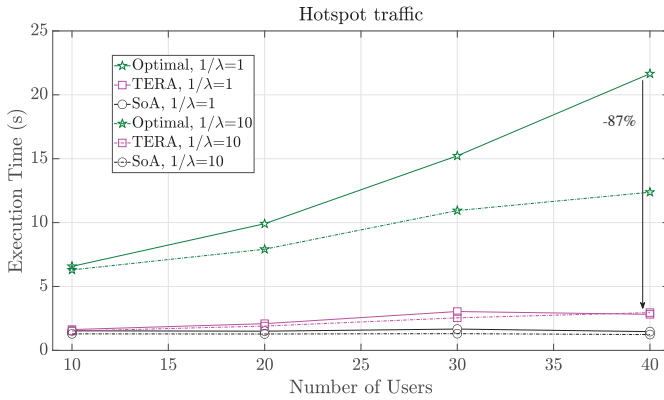


Fig. 5. Execution time in sec of all algorithms for different user traffic load conditions and UE arrival rates ( $1/\lambda$ ).

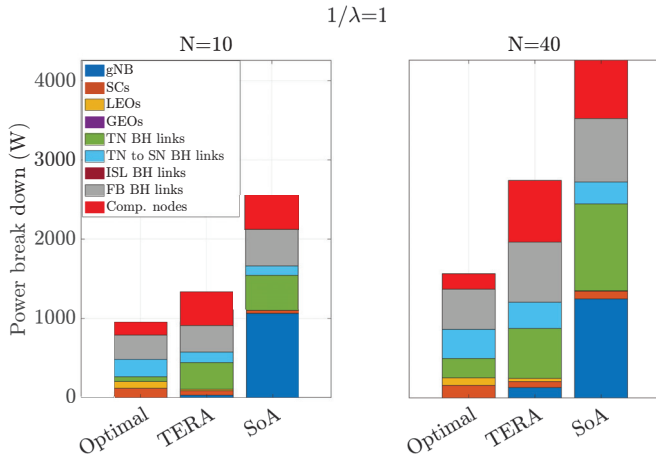


Fig. 6. Power break-down in W of all algorithms for low and high traffic ( $N=10$  and  $N=40$ , respectively) as well as high arrival rate ( $1/\lambda=1$ ).

scenarios ( $N = 40$  and  $1/\lambda = 1$ ), which further supports our approach of jointly studying user association, traffic routing and VNF placement to guarantee truly optimal E2E real-time network performance in integrated 6G networks.

## V. CONCLUSION

We jointly studied the online user association, traffic routing

and xNF placement in 6G integrated TN-SNs, targeting at maximizing the network energy efficiency, while minimizing the UE blocking probability. A computationally expensive MILP analytical solution was derived, capturing all characteristics of the employed technologies, service and resource types as well as their constraints, while minimal assumptions were made. To decrease the complexity of the optimal solution, we proposed TERA, an energy-efficient real-time heuristic for the integrated TN-SN. TERA was shown to achieve up to 85% of the Optimal with up to 87% lower complexity, while significantly outperforming the SoA, thus proving its suitability for integrated 6G TN-SNs.

## ACKNOWLEDGMENT

This work is supported by Horizon Europe SNS JU ETHER project (101096526), the HORIZON-MSCA-2022-DN-01 ELIXIRION project (101120135) and the AROMA3D project (TSI-063000-2021-70/71) under the UNICO5G-RPTR programme.

## REFERENCES

- [1] O. Bulakci *et al.*, “Towards Sustainable and Trustworthy 6G: Challenges, Enablers, and Architectural Design,” in *Boston-Delft: now publishers*, 2023.
- [2] L. Liu *et al.*, “Joint Dynamical VNF Placement and SFC Routing in NFV-Enabled SDNs,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, pp. 4263–4276, Dec. 2021.
- [3] ITU-R S.1590, “Technical and operational characteristics of satellites operating in the range 20-375 THz,” ITU-R, Tech. Rep., 2002.
- [4] S. Chan *et al.*, “Intelligent Low Complexity Resource Allocation Method for Integrated Satellite-Terrestrial Systems,” *IEEE Wireless Communications Letters*, vol. 11, no. 5, pp. 1087–1091, 2022.
- [5] W. Qiu, A. Liu, and C. Han, “Joint Downlink Spectrum Allocation and User Association for Satellite-Terrestrial Integrated Networks,” in *2022 IEEE 14th International Conference on Advanced Infocomm Technology (ICAIT)*, 2022, pp. 19–24.
- [6] B. Ma *et al.*, “Computation-Dependent Routing Based Low-Latency Decentralized Collaborative Computing Strategy for Satellite-Terrestrial Integrated Network,” in *2022 14th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2022, pp. 1–5.
- [7] Y. Zhang *et al.*, “Resource Allocation in Terrestrial-Satellite-Based Next Generation Multiple Access Networks With Interference Cooperation,” *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 4, pp. 1210–1221, 2022.
- [8] Y. Yue *et al.*, “Delay-aware and resource-efficient vnf placement in 6g non-terrestrial networks,” in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 2023, pp. 1–6.
- [9] X. Qin *et al.*, “Service-aware resource orchestration in ultra-dense leo satellite-terrestrial integrated 6g: A service function chain approach,” *IEEE Transactions on Wireless Communications*, vol. 22, no. 9, pp. 6003–6017, 2023.
- [10] X. Cao *et al.*, “Edge-assisted multi-layer offloading optimization of leo satellite-terrestrial integrated networks,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 2, pp. 381–398, 2023.
- [11] A. Varasteh *et al.*, “Holu: Power-aware and delay-constrained vnf placement and chaining,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1524–1539, 2021.
- [12] A. Mesodiakaki *et al.*, “ONE: Online Energy-efficient User Association, VNF Placement and Traffic Routing in 6G HetNets,” in *2022 IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 304–309.
- [13] A. Mesodiakaki *et al.*, “Optimal User Association, Backhaul Routing and Switching off in 5G HetNets with Mesh Millimeter Wave Backhaul Links,” *Ad Hoc Networks*, vol. 78, pp. 99–114, Sep. 2018.
- [14] T.S. Kelso, “Celestrak,” <https://celestrak.org/>.
- [15] W. V. Heddeghem *et al.*, “Power Consumption Modeling in Optical Multilayer Networks,” *Photon. Netw. Commun.*, vol. 24, pp. 86–102, Jan. 2012.