IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 September 2022, accepted 24 September 2022, date of publication 27 September 2022,
date of current version 7 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3210327

RESEARCH ARTICLE

Service-Aware User Association and Resource Allocation in Integrated Terrestrial and Non-Terrestrial Networks: A Genetic Algorithm Approach

DENISE JOANITAH BIRABWA[1,2], (Student Member, IEEE), DANIEL RAMOTSOELA[1], (Member, IEEE), AND NECO VENTURA[1], (Life Member, IEEE)
[1]Department of Electrical Engineering, University of Cape Town, Rondebosch 7700, South Africa
[2]Department of Electrical and Electronic Engineering, Kyambogo University, Kampala, Uganda

Corresponding author: Denise Joanitah Birabwa (brbden001@myuct.ac.za)

This work was supported in part by Kyambogo University Staff Development Fund, in part by the University of Cape Town, and in part by the Telkom Centre of Excellence (COE) in Broadband Networks, South Africa.

ABSTRACT In 6G networks and beyond, multiple radio access networks (RANs), including; the satellite, high altitude platforms, low altitude platforms, and the terrestrial network, will co-exist. These networks are characterized by different capabilities and limitations in meeting the envisioned 6G contrasting user requirements. Therefore, associating users with the appropriate radio access network (RAN) in such an integrated network is rigorous and complex. In this work, the user association and resource allocation problem is formulated as a multi-objective optimization problem (MOOP), aiming to maximize data rate while minimizing mobility-induced handoff in the integrated network. Moreover, the problem is formulated in such a way as to prioritize the service provisioning of mission-critical users. The weighted sum method is adopted to simplify and transform the MOOP into a single-objective optimization problem (SOOP). In order to solve the formulated NP-hard SOOP, a genetic algorithm (GA) whose fitness value is based on the user's service group is proposed. The performance of the proposed algorithm is evaluated by comparing it to the optimal solution, the greedy signal-to-interference-plus-noise ratio (SINR) based association, and the random user association algorithms. Simulation results show that as the number of access nodes in the network increases, the GA's spectrum efficiency (SE) remains within 0.4% of the optimal solution. Moreover, the GA outperforms all three schemes in user acceptance ratio (AR) and handoff probability.

INDEX TERMS RAN user association, resource allocation, terrestrial networks, non-terrestrial networks, genetic algorithm.

I. INTRODUCTION

A. BACKGROUND

In recent times, the information communication technology sector has witnessed an explosive growth in demand for high-speed wireless access, which has increasingly strained the terrestrial network (TN) [1]. While technologies such as ultra-dense networks (UDNs) and device-to-device (D2D) communications have shown great potential in increasing the capacity of the terrestrial networks (TNs), they are faced with different challenges [2], [3]. UDNs are limited by frequent handoff, interference, and backhaul challenges, while D2D communication is faced with frequency planning and resource allocation implementation complexity.

To solve such challenges and increase the TN's capabilities in providing ubiquitous broadband connectivity, one of the key enabling features of the sixth generation of wireless networks (6G) is the integration of TNs with non-terrestrial networks (NTNs) [4], [5], [6], [7], [8]. The considered NTNs include satellite communications (SatComs) and unmanned

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaodong Xu.

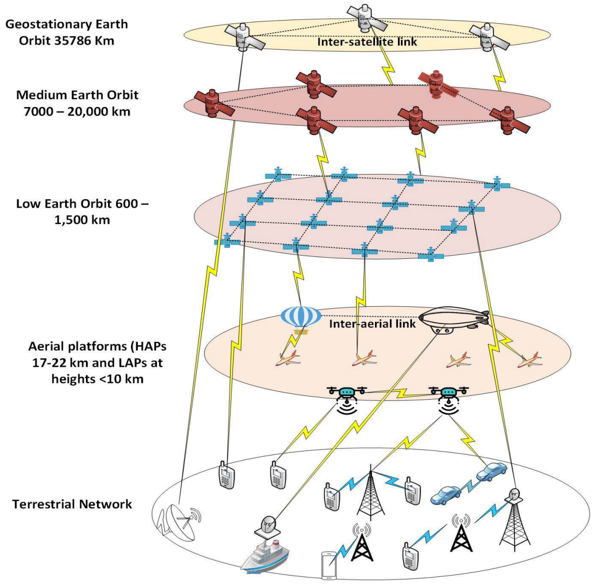

Geostationary Earth Orbit 35786 Km
Inter-satellite link
Medium Earth Orbit 7000 – 20,000 km
Low Earth Orbit 600 – 1,500 km
Aerial platforms (HAPs 17-22 km and LAPs at heights <10 km
Inter-aerial link
Terrestrial Network

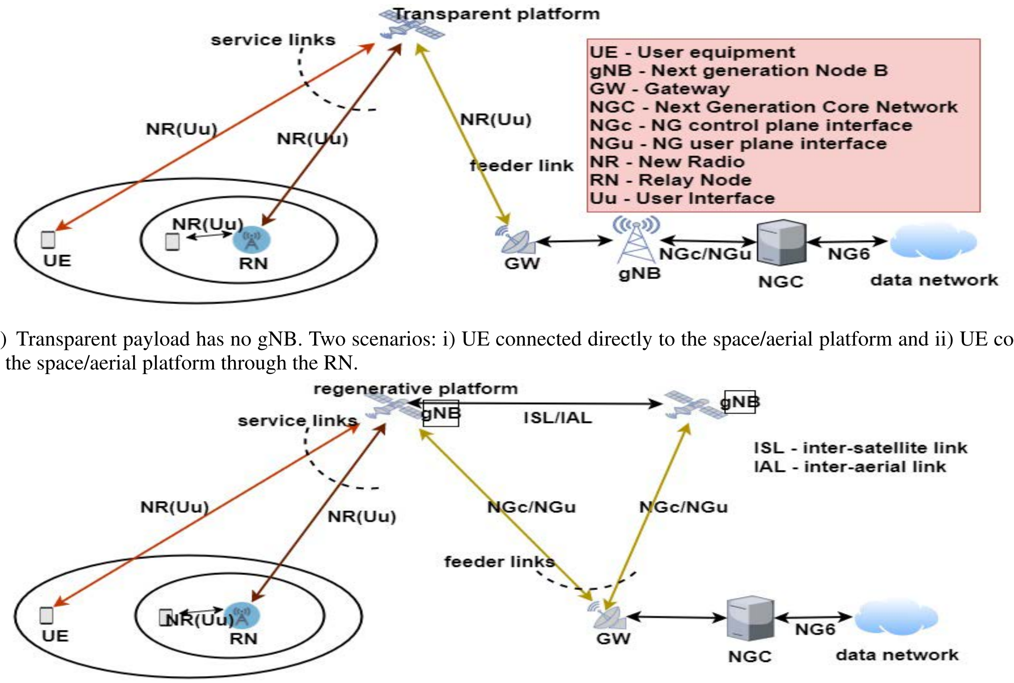
FIGURE 1. Layered architecture of the terrestrial and non-terrestrial integrated network.

aerial vehicles (UAVs), which can be classified into high altitude platforms (HAPs) and low altitude platforms (LAPs) [2]. Fig. 1 depicts the integrated terrestrial and non-terrestrial network (ITNTN) architecture, which is a layered and 3-dimensional integration of SatComs, aerial platforms, i.e., HAPs and LAPs, and the TN. SatComs consists of the Geostationary Earth Orbit (GEO), Medium Earth Orbit (MEO), and Low Earth Orbit (LEO) located at altitudes of 35786 km, 7000 – 20,000 km, and 600 – 1500 km, respectively [9]. On the other hand, HAPs are repeaters flying at an altitude of 17-22 km in the stratosphere [10]. They can be classified as aerostatic and aerodynamic, with aerostatic HAPs taking the form of either balloons or airships. Balloons are designed to stay still in space, while airships are quasi-stationary with onboard electric motors and propellers for station keeping [10]. On the contrary, the aerodynamic HAP is an aircraft that has to stay in a forward motion to keep afloat. Compared to the other networks of the ITNTN, LAPs have the lowest cost and are characterized by fast and easy deployment [10], [11]. This attribute makes them suitable for providing communication services for emergency response and acting as aerial BSs for direct user equipment (UE) connectivity and traffic offloading during limited-duration events such as festivals and sports events. They are relatively small and light and operate at low altitudes, not exceeding 10 km above the earth's surface. The TN layer includes fixed and mobile users using radio access technologies such as 5G, 4G, and WiFi to access various heterogeneous networks consisting of macro, micro, pico, and femtocells.

The integration of TN with NTNs is motivated by the technological advancement in manufacturing and launching processes that has resulted in the massive deployment of LEO and MEO satellites by companies such as OneWeb, and SpaceX [12]. Furthermore, companies such as Airbus and Google have heavily invested in HAPs [13], [14]. Moreover, the third generation partnership project (3GPP) has successfully conducted feasibility studies on radio access through the NTNs [9], [15]. According to 3GPP, the NTNs will support the TN in providing radio access via two radio access network (RAN) architectures depicted by Fig. 2. First is the transparent/bent pipe payload architecture portrayed by Fig. 2(a), in which the NTN access node (AN) acts only as a relay, with its function being frequency filtering, frequency conversion from uplink to downlink, and signal amplification. The second architecture is the regenerative payload configuration illustrated in Fig. 2(b) in which the AN not only performs the functions of a transparent payload but also implements demodulation/decoding, switch and/or routing, coding/modulation. Such an AN is considered a next-generation node B (gNB) as it performs the functions of a base station. In both architectures, the UE can connect to the NTN AN either directly or through a relay node (RN) using the new radio (NR) interface, as depicted by Fig. 2. On the other hand, the connection from the transponder to the gateway (GW) is through the NR air interface for the transparent configuration, while the regenerative architecture uses the NG (Next Generation) interface. An interested reader is referred to [9] for a detailed description of the NTN RAN architecture. Indeed, the future wireless networks will have the TN co-existing with the NTNs to provide radio access to multi-mode UE.

Motivated by these trends, this paper seeks to address the problem of user association and resource allocation in the ITNTN. In particular, and similar to [16] and [17], the work considers the usage scenario in which there exists a large number of users whose traffic cannot be entirely supported by the TN RAN, for example, in an urban area during a carnival event. Such a usage scenario necessitates the deployment of NTNs to de-congest and support the TN RAN in providing radio access to the different users. The UE in this integrated network is considered a multi-radio terminal with the ability to access either the terrestrial access network or the space/airborne communication networks, including the SatComs, HAPs, and LAPs. Without loss of generality, the NTN access nodes (ANs) are assumed to be regenerative, and users can connect to them directly. The different radio access networks (RANs) in this ITNTN have diverse constraints and capabilities towards meeting the heterogeneous requirements of the beyond 5G (B5G) users.

One of the salient features of the fifth generation of wireless networks (5G) is differentiated service provisioning. Unlike the previous generations of mobile networks that were based on the traditional one-size-fits-all architecture, 5G leverages technologies such as network function virtualization (NFV) [18], software defined networking (SDN) [3], and network slicing [19] to provide an agile and flexible network. Such a network is designed for service provisioning based on the heterogeneous quality of service (QoS) requirements of diverse users. To this end, the various user services are classified into different service groups (use-cases), with



(a) Transparent payload has no gNB. Two scenarios: i) UE connected directly to the space/aerial platform and ii) UE connected to the space/aerial platform through the RN.

(b) Regenerative payload consists of gNB. Two scenarios: i) UE connected directly to the space/aerial platform and ii) UE connected to the space/aerial platform through the RN.

FIGURE 2. NTN radio access network architecture.

each service group comprising services of related attributes and priorities [20], [21]. The envisaged service groups in 6G according to [5] include further enhanced mobile broadband (feMBB), enhanced ultra-reliable and low-latency communications (euRLLC), ultra-massive machine-type communications (umMTC), long-distance and high-mobility communication (LDHMC), and extremely low power communications (ELPC).

The problem that arises then is how to map such heterogeneous requirements to the appropriate RAN in the ITNTN. For example, although the TN is endowed with a rich pool of resources that result in high throughput and low latency, it still cannot support an increased number of users in situations of traffic overload; hence it is essential to decide which users are served by the TN and which by the NTN RANs. Moreover, the TN is limited in coverage per base station (BS), making it an unsuitable RAN for LDHMC users. In comparison, the NTNs are characterized by a wide coverage that reduces the number of handoffs experienced by the LDHMC users, ultimately decreasing the associated delays and signaling overheads. Along the same lines, SatComs is limited by a high end-to-end delay and is not suitable for direct connectivity of the mission-critical euRLLC service group. Similarly, the HAPs and LAPs are constrained by the number of available channels; hence, it is crucial to prioritize their use for traffic whose QoS requirements may not be met satisfactorily by the other RANs. Clearly, associating the different users to appropriate RANs in the ITNTN while maximizing resource utilization and providing the required user QoS is a rigorous and complex task.

B. RELATED LITERATURE AND CONTRIBUTIONS

Several works have addressed resource allocation in the ITNTN in recent times. The authors in [22] propose a user association and resource allocation problem that maximizes the total data rate of an integrated UAVs and terrestrial RAN together with satellite and macro BSs backhaul links. This work is supplemented in [16] by the same authors by taking into account the access node's energy cost and the limited life endurance of UAVs. However, the work does not consider the satellite as an AN but only as a backhaul link and does not consider HAPs in the integration. Therefore, the authors did not consider the idea of having space (satellite) and aerial (LAPs and HAPs) nodes complement the TN in providing radio access to users. Moreover, the work did not account for how the different user requirements were met.

In [23], the authors propose a load balancing algorithm for an integrated satellite and TN. They define the radio resource utilization ratio as a metric used to measure each cell's load status and group traffic into delay-sensitive and delay-tolerant. Traffic is offloaded from an overloaded terrestrial cell to neighboring terrestrial cells first and then to the satellite cell. Delay-intolerant traffic is not offloaded to the satellite network. However, for 6G networks to have an effective and efficient resource allocation scheme, traffic can not only be classified into two but in different use-cases that effectively capture all future network demands. Moreover, this work only associates users to the NTN when the TN is overloaded. Besides, the authors did not consider other NTNs such as LAPs and HAPs.

In [24], the authors jointly optimize user association and resource allocation of both access and backhaul links, together with HAPs' locations, to maximize users' throughput in an integrated satellite, airborne, and TN. These authors neither considered the heterogeneous user QoS requirements nor the different limitations of the networks in the integration. The authors in [25] propose a joint algorithm that optimizes user association and resource allocation to a terrestrial macro base station (MBS) and multiple UAVs mounted BSs using in-band wireless backhaul. However, the authors only consider UAVs and TNs in their analysis, and they do not account for the provisioning of the different service group QoS requirements.

The authors in [26] propose an algorithm that maximizes the energy efficiency (EE) of an integrated satellite/terrestrial cache-enabled RAN. Both terrestrial ANs and LEO satellites provide content distribution and retrieval services, thereby offloading such traffic from the MBS. The work does not consider the implementation of an integrated terrestrial-aerial-space RAN. It neither addresses QoS provisioning to different use-cases, as it considers only one service: content distribution. In [17], the authors propose an optimization problem that maximizes the network data rate while ensuring QoS by minimizing interference in an integrated satellite and TN. The work fails to distinguish users according to their different QoS requirements, such as data rates, and does not guarantee that these QoS demands are satisfied. Besides, it assumes that all the available networks can support all different users.

The authors in [27] maximized the minimum ergodic achievable rate of a user-UAV link. This work optimizes only the UAV RAN and not the entire integrated terrestrial-satellite-UAV RAN. Besides, the work did not incorporate QoS provisioning of the different use-cases. The authors in [28] first optimize multi-beam dynamic radio resource allocation for LEO-ground downlinks. Next, they optimize the dynamic resource allocation for HAP-ground downlinks when HAPs and LEO satellites share the same spectrum. This work does not consider joint resource management of the different RANs, as it optimizes the resource allocation of each RAN separately. The authors in [29] propose a traffic offloading scheme that considers the diversity of user demands. The optimization algorithm maximizes the eMBB users' data rate, subjected to stringent outage probability of the uRLLC use-case. This work does not consider using SatComs directly for radio access, and neither does it incorporate the UAVs (i.e., LAPs and HAPs) in the radio resource management.

Different from all the above work, we seek to find an optimal dynamic radio access user association and resource allocation scheme for an ITNTN comprising the TN, LAP, HAP, and SatComs ANs. We consider the heterogeneity in users' demands and the different RANs' uniqueness in meeting these diverse demands. The work considers three service groups: feMBB, euRLLC, and LDHMC.

The user association and resource allocation problem is formulated as a multi-objective optimization problem (MOOP), maximizing the total network data rate while simultaneously prioritizing large coverage NTNs over the TNs for service provisioning of the mobile LDHMC service group, with the objective of minimizing mobility-induced handoff probability. Moreover, given that denial of service to the euRLLC service group may result in catastrophic events, the optimization problem prioritizes the service provisioning of this service group over others and limits its access to the SatComs AN. The MOOP is simplified and transformed into a weighted sum single-objective optimization problem (SOOP). The formulated combinatorial and non-convex SOOP is NP-hard, with no efficient polynomial-time solution. Owing to its simplicity and efficiency in solving non-convex and combinatorial problems [30], we solve the SOOP using the genetic algorithm (GA).

The GA is a meta-heuristic search algorithm that uses the theory of evolution and natural selection to solve optimization problems. It is well suited for multi-objective and non-mathematical optimization problems, efficiently and easily enforcing different constraints, and also searching over multiple sets of solutions in a large search space to return a near-optimal solution [31]. Given its ease of implementation and optimization of discrete and continuous radio parameters, the GA is an excellent optimization tool for radio resource management in the ITNTN. The performance of the proposed GA is compared to three other algorithms; the optimal solution simulated using the gurobi solver, the greedy signal-to-interference-plus-noise ratio (SINR) based solution [32], and the random user association (RUA) solution. Simulation results show that as the number of ANs increase in the ITNTN, the GA's spectrum efficiency (SE) performance remains within 0.4% of the optimal solution and outperforms the greedy and RUA by 1.23% and 0.97%, respectively. Moreover, the GA outperforms the optimal, greedy, and RUA algorithms in handoff probability by 8.4%, 14.9%, and 51.8% on average, respectively. Furthermore, the GA shows an acceptance ratio (AR) performance that is better than the optimal, the greedy, and the RUA solutions by 1.41%, 10.8%, and 7.6%, respectively. The key contributions of this work can therefore be summarised as follows:

- Formulation of a user association and resource allocation optimization problem that maximizes the data rate of the ITNTN while simultaneously minimizing the probability of mobility-induced handoff. Handoff is minimized by prioritizing the use of large coverage NTNs by the mobile LDHMC service group. Moreover, service provisioning of the mission-critical euRLLC use-case is prioritized over other use-cases.
- The formulated multi-objective problem is transformed into a single-objective problem which is solved using the GA by encoding the problem into a sequence of chromosomes, with the genes representing user association solutions. Service group-dependent fitness functions are formulated to determine the near-optimal user association and resource allocation solution.
- Numerical results are presented, comparing the proposed GA to three algorithms; the optimal solution

simulated using the gurobi solver, the greedy algorithm whose association is based on maximum SINR [32], and the random user association solution.

C. ORGANISATION

The remainder of this paper is organized as follows: Section II gives a detailed description of the system model and assumptions, while section III defines the problem formulation. In Section IV, the GA process is reviewed, and the solution to the formulated problem based on the GA described. The results are analyzed in section V, and finally, the conclusion presented in section VI. The notations used in this paper are presented in Table 1.

TABLE 1. Notations.

Symbol	Description
b, u, h, s	MBS, LAP, HAP, SatComs AN
$\mathcal{B}, \mathcal{U}, \mathcal{H}, \mathcal{S}$	Set of ANs in the MBS, LAP, HAP, SatComs RAN
j, o_j	A RAN, AN in the j^{th} RAN
$\mathcal{I}$	The set of users in the ITNTN
$\upsilon, \mathcal{E}, \mathcal{R}, \mathcal{D}$	A use-case, feMBB, euRLLC, LDHMC use-case
$\mathcal{I}_{\mathcal{E}}, \mathcal{I}_{\mathcal{R}} \mathcal{I}_{\mathcal{D}}$	set of users demanding $\mathcal{E}, \mathcal{R}, \mathcal{D}$
$i, i_{\mathcal{E}}, i_{\mathcal{R}}, i_{\mathcal{D}}$	User, An feMBB, euRLLC, LDHMC user
$\mathcal{C}_j$	set of BBUs owned by RAN j
c_j	A BBU owned by RAN j
$\mathcal{T}_{c_j}, \Phi_{o_j}$	Bandwidth of: a BBU c_j, AN o_j
$\mathcal{O}$	Set of all ANs in the ITNTN
$z_{o_j}, R_{o_j}, (x_{o_j}, y_{o_j})$	Altitude, cell radius, ground coordinates of o_j
(x_i, y_i)	user i coordinates
$d_{o_j,i}, \theta_{o_j,i}$	Distance, elevation angle from user i to AN o_j
h_{o_j}, τ_{o_j}	MBS height, loss due to shadow fading
$PL_{o_j,i,c_j}, \Gamma_{o_j,i,c_j}$ $\mathcal{L}_{o_j,i,c_j}$	Path loss, channel gain , data rate, of a user i using BBU c_j of AN o_j
$\gamma_{o_j,i_q,c_j}, P_{o_j,i_q,c_j}$	SINR , transmit power of a user i_q using BBU c_j of AN o_j
$f_{o_j,c_j}, \beta_{o_j,i_q,c_j}$	Carrier frequency, small-scale fast fading
$Prob_{o_j,i}^{LoS}, Prob_{o_j,i}^{NLoS}$	line of sight (LoS), non line of sight (NLoS) Probability
x, y	Environmental constants
$PL_{o_j,i,c_j}^{LoS}, PL_{o_j,i,c_j}^{NLoS}$	LoS, NLoS path loss
η_{LoS}, η_{NLoS}	Additional loss in a UAV LoS, NLoS propagation
CL, PL^k, PL^e, PL^y	loss due to clutter, atmospheric gases, scintillation, building entry.
$\mu_{o_j,i}, \omega_{o_j,i,c_j}, \pi_{o_j,i}$	User association variable, resource allocation variable, coverage index
$\delta_1, \delta_2, \delta, \zeta$	Normalization factors
L_{thres}^{υ}	Service group υ threshold data rate
$\rho_{i_\upsilon}, \alpha$	user priority factor, objective function trade-off factor
$\xi_{i_\upsilon}, \xi_{i_{\mathcal{D}}}$	user penalty factors
M, G, P_c, P_m, E_r	Population size, Number of iterations, cross over probability, mutation probability, Elitism ratio
$\mu_k, \mu_{o_j,i}^k, x_{i,k}$	A k^{th}chromosome, a gene of the k^{th} chromosome indicating user i is associated with AN o_j, validity status of user $i's$ gene in the k^{th} chromosome

II. SYSTEM MODEL

This section describes the deployment scenario, mobility model, channel model, signal quality model and assumptions considered in this work.

A. DEPLOYMENT SCENARIO

A downlink transmission of an integrated communication network consisting of four RANs namely, the MBS, LAP, HAP and the LEO SatComs as depicted in Fig. 3, is considered. An AN in the MBS RAN is indexed as $b \in \mathcal{B}$ while that in the LAP RAN denoted as $u \in \mathcal{U}$. Similarly, a HAP AN is represented as $h \in \mathcal{H}$ while the SatComs AN indexed as $s \in \mathcal{S}$. A RAN in the ITNTN is denoted by $j \in \{\mathcal{B}, \mathcal{U}, \mathcal{H}, \mathcal{S}\}$, while an AN in the j^{th} RAN represented by o_j.

Similar to [33], [34], [35], and [36], this work is premised on the assumption that each UAV AN $o_j \in \{u, h\}$ is stationary or quasi-stationary with negligible mobility. Immobility of UAVs is assumed to avoid disconnections due to the UAV AN moving out of coverage of already connected users, some of whom could be mission critical. Consequently, we assume to use the rotary-wing LAPs and balloon or airship HAPs that have the ability to remain quasi-stationary [11], [37]. Moreover, we assume that the placement of the UAV ANs has already been optimized to cater to the usage scenario in which there is a large number of users, say in an urban area, during a carnival event.

The system provides downlink communications to a set of users $\mathcal{I} = \{1, 2, 3, \ldots, |\mathcal{I}|\}$ and supports three use-cases namely; feMBB, euRLLC, and LDHMC, denoted by $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{D}$ respectively, such that $\upsilon \in \{\mathcal{E}, \mathcal{R}, \mathcal{D}\}$. Users demanding use-cases $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{D}$ are grouped in sets denoted by $\mathcal{I}_{\mathcal{E}}$, $\mathcal{I}_{\mathcal{R}}$, and $\mathcal{I}_{\mathcal{D}}$ respectively, such that, $i_{\mathcal{E}} \in \mathcal{I}_{\mathcal{E}}$, $i_{\mathcal{R}} \in \mathcal{I}_{\mathcal{R}}$, $i_{\mathcal{D}} \in \mathcal{I}_{\mathcal{D}}$, $\mathcal{I} = \mathcal{I}_{\mathcal{E}} \cup \mathcal{I}_{\mathcal{R}} \cup \mathcal{I}_{\mathcal{D}}$ and $\mathcal{I}_{\mathcal{E}} \cap \mathcal{I}_{\mathcal{R}} \cap \mathcal{I}_{\mathcal{D}} = \emptyset$. Without loss of generality, the $\mathcal{I}_{\mathcal{E}}$ and $\mathcal{I}_{\mathcal{R}}$ users are assumed to be static while the $\mathcal{I}_{\mathcal{D}}$ users are mobile. An example of a static use-case belonging to the $\mathcal{I}_{\mathcal{R}}$ service group is remote surgery within a hospital or private clinic in the considered urban area. Also, a user $i \in \mathcal{I}$ is assumed to be embedded with multiple RAN interfaces, and thus can access any of the available RAN $j \in \{\mathcal{B}, \mathcal{U}, \mathcal{H}, \mathcal{S}\}$ within its coverage.

Since the different RANs in the ITNTN may have different multiple access schemes such as OFDMA, TDMA, and FDMA, the basic bandwidth unit (BBU) is used to represent the unit of radio resources as was done in [38]. Therefore,

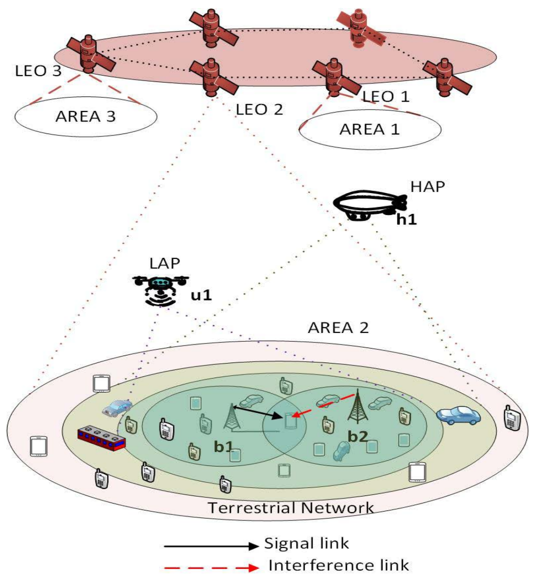


FIGURE 3. System model.

no matter what type of access technique is used, the system capacity is represented in terms of bandwidth. Similar to [39], this work assumes shared spectrum mode, such that all ANs that belong to the same RAN j own the same set $\mathcal{C}_j$ of BBUs. A BBU $c_j \in \mathcal{C}_j$ has a bandwidth denoted by $\mathcal{T}_{c_j}$, while the bandwidth of an AN $o_j \in \mathcal{O}$ where $\mathcal{O} = \{\mathcal{B} \cup \mathcal{U} \cup \mathcal{H} \cup \mathcal{S}\}$ is represented as Φ_{o_j}.

Furthermore, as was done by [24], we make the following assumptions: i) A user $i \in \mathcal{I}$ can associate with at most one AN. ii) For simplicity, the different RANs use frequencies sparsely separated from each other, thus have no cross-tier interference. iii) Intra-cell interference on the downlink between users associated with the same AN is negligible, as it can be effectively controlled through multiple access techniques. iv) Co-tier interference exists, where a user receives signals from different ANs in the same RAN. This interference is added to the thermal noise in the SINR expression defined in (9).

The system model used for the LEO SatComs RAN is adopted from [26]. A LEO satellite follows a specific pattern in which it periodically serves one area followed by another. Consequently, in this work, we assume that each non-overlapping area depicted in Fig. 3 is served by one LEO satellite for a given duration of time. Hence, a user in a given location can access only one satellite at any given time. As an illustration, let us consider each LEO satellite's view time of any given area to be t. Then if LEO 2 starts to view area 2 at time $t1$, this view continues until a time $t1 + t$, after which service provisioning of area 2 is handed over to LEO 1, which views from $t1 + t$ until $t1 + 2t$. Handover from one satellite to another is assumed to be managed by the Network Control Center [26]. In this work, this handover process is out of scope and thus shall not be considered. As the entire considered terrestrial area is under coverage by a single satellite during its service period, this implies that seamless coverage of the terrestrial RAN is guaranteed by the satellite, irrespective of its mobility. Moreover, in practice, thousands of LEO satellites are deployed, for example, in the Starlink project, implying that a particular area can be covered by multiple LEO satellites simultaneously [40], [41]. Therefore, the mobility model of the satellite shall not be considered in this work.

B. USER MOBILITY MODEL

Due to its simplicity and analytical tractability, the random walk mobility model [42] is used to imitate the LDHMC users' movement patterns. In each new interval, a user $i_{\mathcal{D}} \in \mathcal{I}_{\mathcal{D}}$ chooses a direction $\theta_{i_{\mathcal{D}}} \in [0 \;\; 2\pi]$ that is randomly and uniformly distributed. In the same manner, the user's speed $V_{i_{\mathcal{D}}} \in [V_{min} \;\; V_{max}]$ is randomly assigned following a uniform distribution, with V_{min} and V_{max} being the minimum and maximum velocity respectively, that a user can have. Considering the user's location in time interval t as $(x_{i_{\mathcal{D}}}(t), y_{i_{\mathcal{D}}}(t))$, then the location $(x_{i_{\mathcal{D}}}(t+1), y_{i_{\mathcal{D}}}(t+1))$ in interval $t+1$ is given by

$$\begin{aligned} x_{i_{\mathcal{D}}}(t+1) &= x_{i_{\mathcal{D}}}(t) + \frac{V_{i_{\mathcal{D}}}}{V_{max}} * D_{max} * \cos\theta_{i_{\mathcal{D}}}, \\ y_{i_{\mathcal{D}}}(t+1) &= y_{i_{\mathcal{D}}}(t) + \frac{V_{i_{\mathcal{D}}}}{V_{max}} * D_{max} * \sin\theta_{i_{\mathcal{D}}}, \end{aligned} \tag{1}$$

where D_{max} is the maximum distance that can be moved by a user in a given flight interval. We consider LDHMC users to move within the LEO satellite coverage area, assumed to be 5 km, such that they are reflected off the boundary of the circular region.

C. CHANNEL MODEL

The channel modelling is similar to our work in [32]. The ground location of any AN $o_j \in \{\mathcal{B} \cup \mathcal{U} \cup \mathcal{H} \cup \mathcal{S}\}$ is represented by $t_{o_j} = \{[x_{o_j}, y_{o_j}]^T \in \mathbb{R}^2\}$. In this case, for the NTN ANs, t_{o_j} is their projection on the ground. On the other hand, the coordinate of a user $i \in \mathcal{I}$ is given by $t_i = \{[x_i, y_i]^T \in \mathbb{R}^2 | i \in \mathcal{I}\}$. Consequently, the distance $d_{o_j,i}$ from an AN $o_j \in \{\mathcal{B} \cup \mathcal{U} \cup \mathcal{H} \cup \mathcal{S}\}$ to a user i can be calculated using

$$d_{o_j,i} = \sqrt{\|t_i - t_{o_j}\|_2^2 + z_{o_j}^2} \quad \forall j \in \{\mathcal{U}, \mathcal{H}, \mathcal{S}\}, \quad \forall i \in \mathcal{I}, \tag{2}$$

where z_{o_j} is the height of the AN and $\|.\|_2$ is the 2-norm operator. Path loss modeling is divided into three categories: (i) MBS TN, (ii) UAV, i.e., HAPs and LAPs, and (iii) SatComs.

1) MBS TERRESTRIAL NETWORK PATH LOSS MODEL

The Path loss of a user $i \in \mathcal{I}$ using a BBU $c_j \in \mathcal{C}_j | j \in \mathcal{B}$ to communicate to an urban MBS $o_j \in \mathcal{B}$ that is located a distance $d_{o_j,i}$ meters away is given by [43]

$$\begin{aligned} PL_{o_j,i,c_j} &= 40(1 - 4 \times 10^{-3} h_{o_j}) \log_{10}(\frac{d_{o_j,i}}{1000}) - 18 \log_{10} h_{o_j} \\ &\quad + 21 \log_{10} f_{o_j,c_j} + 80 + \tau_{o_j,i}, \quad \forall j \in \mathcal{B}, \quad \forall i \in \mathcal{I}. \end{aligned} \tag{3}$$

The term h_{o_j} represents the MBS height in meters, f_{o_j,c_j} is the carrier frequency in MHz, and $\tau_{o_j,i}$ is the path loss due to shadow fading, assumed to be a Gaussian random variable with zero mean and σ standard deviation in dB. $\tau_{o_j,i}$ can be expressed as $\tau_{o_j,i} = \log_{10}(F_{o_j,i})$ where $F_{o_j,i}$ is the log-normal shadow fading path loss between the user i and AN o_j [43].

2) UAV PATH LOSS MODEL

The path loss from a UAV AN $o_j \in \{\mathcal{U} \cup \mathcal{H}\}$ to a user $i \in \mathcal{I}$ is modelled according to [44]. In this model, the probability that a user i has a line of sight (LoS) link from a UAV AN $o_j \in \{\mathcal{U} \cup \mathcal{H}\}$ is given by

$$Prob^{LoS}_{o_j,i} = \frac{1}{1 + x\exp(-y(\theta_{o_j,i} - x))} \quad \forall o_j \in \{\mathcal{U} \cup \mathcal{H}\}\,. \quad (4)$$

The constants x and y are dependent on the environment while the elevation angle $\theta_{o_j,i}$ is given by $\frac{180}{\pi}\arctan(\frac{z_{o_j}}{\|t_i - t_{o_j}\|_2})$. The signal from the UAV AN propagates through free space until it reaches the environment on earth, where it undergoes additional loss due to shadowing, scattering, and reflections caused by buildings, foliage, etcetera. Consequently, the path loss from the UAV AN $o_j \in \{\mathcal{U} \cup \mathcal{H}\}$ and a user i located at distance $d_{o_j,i}$ is given by

$$PL_{o_j,i,c_j} = \begin{cases} PL^{LoS}_{o_j,i,c_j} = 20\log_{10} d_{o_j,i} + 20\log_{10} f_{o_j,c_j} \\ -27.55 + \eta_{LoS}\,, \qquad \text{LoS scenario} \\ PL^{NLoS}_{o_j,i,c_j} = 20\log_{10} d_{o_j,i} + 20\log_{10} f_{o_j,c_j} \\ -27.55 + \eta_{NLoS} \ \ \text{non LoS (NLoS) scenario}\,, \end{cases} \quad (5)$$

where $d_{o_j,i}$ is distance in meters computed using (2), f_{o_j,c_j} is carrier frequency in MHz, and η_{LoS} or η_{NLoS} is the additional loss experienced in LoS or NLOS propagation respectively. The addition loss η_{LoS} or η_{NLoS} has a Gaussian distribution [44]. The equivalent path loss is then given by [25]

$$PL^{equiv}_{o_j,i,c_j} = Prob^{LoS}_{o_j,i} * PL^{LoS}_{o_j,i,c_j} + Prob^{NLoS}_{o_j,i} * PL^{NLoS}_{o_j,i,c_j}\,, \quad (6)$$

where $Prob^{NLoS}_{o_j,i} = 1 - Prob^{LoS}_{o_j,i}$ is the probability that a user experiences a NLoS link.

3) SATELLITE COMMUNICATIONS PATH LOSS MODEL

Assuming clear sky conditions, the path loss between an AN $o_j \in \{\mathcal{S}\}$ and a user i at a distance $d_{o_j,i}$ from the node, is given by [9]

$$PL_{o_j,i,c_j} = 20\log_{10} d_{o_j,i} + 20\log_{10} f_{o_j,c_j} - 27.55 + \tau_{o_j} \\ +CL + PL^k + PL^e + PL^y, \ \ \forall\, o_j \in \mathcal{S}\,. \quad (7)$$

The frequency f_{o_j,c_j} is in MHz, while the range $d_{o_j,i}$ in meters and can be determined using (2). On the other hand, τ_{o_j} in (7) depicts the loss due to shadow fading, while CL gives the clutter loss resulting from reflections and scattering caused by surrounding buildings and objects on the ground. τ_{o_j} and CL are dependent on whether the propagation is LoS or NLoS. The terms PL^k, PL^e, and PL^y represent attenuation due to atmospheric gases, ionospheric or tropospheric scintillation, and building entry loss. The equivalent path loss is determined using (6) where by the LoS probability $P^{LoS}_{o_j,i}$ which depends on the elevation angle and UE environment is obtained from Table 6.6.1-1 in [9].

4) SIGNAL QUALITY MODEL

Using the results from the preceding sections II-C1 to II-C3, the linear channel gain due to large scale fading effects of path loss and shadowing from an AN $o_j \in \mathcal{O}$ and user i, communicating via a BBU c_j is given by [43]

$$\Gamma_{o_j,i,c_j} = 10^{-\frac{PL_{o_j,i,c_j}}{10}}\,. \quad (8)$$

In practice, for time division duplex systems, the ANs can exploit reciprocity between downlink and uplink channels to estimate the downlink channel [45], [46]. In comparison, for the frequency division duplex systems, there often exists weaker reciprocity in the uplink and downlink frequencies [45], [46]. Consequently, the channel state information (CSI) at the ANs for these systems can be obtained as feedback from the UE [45], [46]. Several works in literature are dedicated to reducing the amount of CSI feedback in the FDD systems [46], [47], but this is out of the scope of this work. The obtained CSI feedback from UEs suffers high latency for ANs at very high altitudes, like the SatComs; hence, this CSI is usually outdated [48]. Therefore, for such RANs, CSI can be obtained by utilizing the widely used training data-based CSI estimation techniques [49], [50], [51]. However, due to the complexities involved in simulating the CSI feedback and estimation techniques, this work assumes that the large-scale CSI is available at the ANs as was done in [24], [43], [52], and [27]. Therefore, the corresponding channel gain can be calculated using (8).

The SINR ratio experienced by the user i using BBU c_j is then determined using

$$\gamma_{o_j,i_q,c_j} = \begin{cases} \dfrac{P_{o_j,i_q,c_j}\ \Gamma_{o_j,i_q,c_j}}{\sigma^2 \ + \sum\limits_{\substack{l_j \in \mathcal{O} \\ l_j \neq o_j}} P_{l_j,i_q,c_j}\ \Gamma_{l_j,i_q,c_j}} \\ \forall i_q \in \{\mathcal{I}_{\mathcal{E}} \cup \mathcal{I}_{\mathcal{R}}\},\ \forall j \in \{\mathcal{B}, \mathcal{U}, \mathcal{H}, \mathcal{S}\} \text{ static user} \\ \dfrac{P_{o_j,i_q,c_j}\ \Gamma_{o_j,i_q,c_j}\, |\beta_{o_j,i_q,c_j}|^2}{\sigma^2 \ + \sum\limits_{\substack{l_j \in \mathcal{O} \\ l_j \neq o_j}} P_{l_j,i_q,c_j}\ \Gamma_{l_j,i_q,c_j}\ |\beta_{l_j,i_q,c_j}|^2} \\ \forall i_q \in \mathcal{I}_{\mathcal{D}},\ \forall j \in \{\mathcal{B}, \mathcal{U}, \mathcal{H}, \mathcal{S}\} \ \text{ mobile user}\,. \end{cases} \quad (9)$$

The terms β_{o_j,i_q,c_j} and β_{l_j,i_q,c_j} denote the small-scale fast fading component that accounts for mobility of a user and is assumed to be independent and identically distributed (i.i.d) as $\mathcal{CN}(0, 1)$ [53]. P_{o_j,i_q,c_j} is the transmit power from an AN o_j to a user i_q using BBU c_j. Similar to [54], [55], and [56], and in a bid to reduce complexity, this work shall not optimize power allocation. Therefore, the transmit power P_{o_j,i_q,c_j} is assumed to be fixed and uniformly allocated to all AN's BBUs. The power P_{o_j,i_q,c_j} is given as $P^{thres}_{o_j}/|\mathcal{C}_j|$, where $P^{thres}_{o_j}$ is the maximum available power at the AN o_j and $|.|$ represents the cardinality of the set. P_{l_j,i_q,c_j} depicts the co-tier interference from any other AN l_j in the j^{th} RAN reaching the same user i_q.

The maximum data rate that can be achieved by a user $i \in \mathcal{I}$ on the BBU c_j of AN o_j is then given by the Shannon capacity as

$$\mathcal{L}_{o_j,i,c_j} = \mathcal{T}_{c_j} \log_2(1 + \gamma_{o_j,i,c_j}) , \tag{10}$$

where $\mathcal{T}_{c_j}$ is the bandwidth for BBU $c_j \in \mathcal{C}_j$. Ultimately, the overall system data rate is then expressed by

$$\Omega = \sum_{o_j \in \mathcal{O}} \sum_{i \in \mathcal{I}} \sum_{c_j \in \mathcal{C}_j} \rho_{i_\upsilon} \; \mu_{o_j,i} \; \omega_{o_j,i,c_j} \; \mathcal{L}_{o_j,i,c_j}, \tag{11}$$

where $\rho_{i_\upsilon} \in [0, 1]$ is a weighting factor for a user $i \in \mathcal{I}$ demanding a given service belonging to service group $\upsilon \in \{\mathcal{E}, \mathcal{R}, \mathcal{D}\}$. ρ_{i_υ} prioritizes service provisioning of users based on the service they demand. $\mu_{o_j,i} \in \{0, 1\}$ is a binary user association variable that specifies whether the user $i \in \mathcal{I}$ is associated with the AN $o_j \in \mathcal{O}$ of the j^{th} RAN, and is defined as:

$$\mu_{o_j,i} = \begin{cases} 1, & \text{if user } i \text{ is associated with AN } o_j \\ 0, & \text{otherwise .} \end{cases} \tag{12}$$

On the other hand, $\omega_{o_j,i,c_j} \in \{0, 1\}$ is a binary resource allocation factor that is 1 if the BBU $c_j \in \mathcal{C}_j$ of AN o_j is allocated to the user i, and 0 otherwise, that is;

$$\omega_{o_j,i,c_j} \begin{cases} 1, & \text{if BBU } c_j \text{ of AN } o_j \text{ is allocated} \\ & \text{to the user } i. \\ 0, & \text{otherwise .} \end{cases} \tag{13}$$

We assume that an AN's BBU can be allocated to at most one user, as illustrated by (14).

$$\sum_{i \in \mathcal{I}} \omega_{o_j,i,c_j} \leqslant 1, \;\; \forall\, o_j \in \mathcal{O}, \;\; \forall c_j \in \mathcal{C}_j \tag{14}$$

Also, a user $i \in \mathcal{I}$ can only associate with ANs in whose coverage the user lies. Therefore, we define a binary index $\pi_{o_j,i}$ that indicates whether a user i is within AN $o_j's$ coverage or not. If the distance between a user $i's$ location (x_i, y_i) and the ground location of the AN o_j (x_{o_j}, y_{o_j}) is less than the AN's cell radius R_{o_j}, then $\pi_{o_j,i} = 1$, otherwise, $\pi_{o_j,i} = 0$, as illustrated by (15).

$$\pi_{o_j,i} = \begin{cases} 1, & if \;\sqrt{(x_i - x_{o_j})^2 + (y_i - y_{o_j})^2} \leqslant R_{o_j} \\ 0, & \text{otherwise .} \end{cases} \tag{15}$$

This work maximizes the overall system data rate while simultaneously minimizing the probability of mobility-induced handoff. We define the probability of handoff as the ratio of the number of users that experienced a handoff to the total number of mobile users during a given transmission time interval (TTI). To minimize the probability of handoff, a handoff reduction function given in (16) is maximized in each TTI.

$$\Lambda = \sum_{o_j \in \mathcal{O}} \sum_{i \in \mathcal{I}} \sum_{c_j \in \mathcal{C}_j} \rho_{i_\upsilon} \; \mu_{o_j,i} \; \omega_{o_j,i,c_j} \; \frac{R_{o_j}}{\zeta} \tag{16}$$

The term ζ in (16) represents the largest cell radius in the ITNTN. A mobile user traverses several small cells during an ongoing call. Hence, the smaller the ANs' cell radii, the more the number of handoffs due to user mobility. In order to limit the mobility-induced handoff, we introduce the ratio $(R_{o_j}/\zeta) \in (0, 1]$ in (16) to prioritize the association of mobile users to ANs with the largest cell radius. In this way, a maximum value of the ratio (R_{o_j}/ζ) implies a reduced number of handoffs since the mobile user will be associated with an available AN having the largest cell radius in each TTI, ultimately reducing the probability of handoff.

III. PROBLEM FORMULATION

In this section, the user association and resource allocation problem is formulated as a MOOP that maximizes the total system data rate while at the same time minimizing the probability of mobility-induced handoff. The probability of handoff is minimized by maximizing the association of mobile users to ANs with a large cell radius. Moreover, ρ_{i_υ} in (11) and (16) can be used to prioritize the service group $\mathcal{R}$ users over other users. The MOOP is formulated as

$$\max_{(\mu,\, \omega)\, \in\, \Delta} \quad \{\Omega, \Lambda\}, \tag{17}$$

where Δ is the set consisting of all feasible user association and resource allocation solutions satisfying the following constraints

$$C1 : \mu_{o_j,i} \leq \pi_{o_j,i}, \;\; \forall\, o_j \in \mathcal{O}, \;\; \forall\, i \in \mathcal{I} \tag{17a}$$

$$C2 : \sum_{o_j \in \{\mathcal{B} \,\cup\, \mathcal{U} \,\cup\, \mathcal{H}\}} \pi_{o_j,i} \; \mu_{o_j,i} \; \omega_{o_j,i,c_j} \;\; \leqslant \;\; 1, \quad \forall\, i_{\mathcal{R}} \in \mathcal{I}_{\mathcal{R}} \tag{17b}$$

$$C3 : \sum_{o_j \in \{\mathcal{S}\}} \pi_{o_j,i} \; \mu_{o_j,i} \; \omega_{o_j,i,c_j} \;\; = \;\; 0, \quad \forall\, i_{\mathcal{R}} \in \mathcal{I}_{\mathcal{R}} \tag{17c}$$

$$C4 : \sum_{o_j \in \mathcal{O}} \pi_{o_j,i} \; \mu_{o_j,i} \; \omega_{o_j,i,c_j} \;\; \leqslant \;\; 1, \;\; \forall\, i_{\mathcal{E}} \in \mathcal{I}_{\mathcal{E}}, \;\; \forall\, i_{\mathcal{D}} \in \mathcal{I}_{\mathcal{D}} \tag{17d}$$

$$C5 : \sum_{o_j \in \mathcal{O}} \sum_{c_j \in \mathcal{C}_j} \pi_{o_j,i} \; \mu_{o_j,i} \; \omega_{o_j,i,c_j} \mathcal{L}_{o_j,i,c_j} \geqslant \mathcal{L}^{\upsilon}_{thres}, \;\; \forall i \in \mathcal{I}_{served}, \;\; \forall\, \upsilon \in \{\mathcal{E}, \mathcal{R}, \mathcal{D}\} \tag{17e}$$

$$C6 : \sum_{i \in \mathcal{I}} \sum_{c_j \in \mathcal{C}_j} \pi_{o_j,i} \; \mu_{o_j,i} \; \omega_{o_j,i,c_j} \mathcal{T}_{c_j} \leq \Phi_{o_j}, \quad \forall\, o_j \in \mathcal{O} \tag{17f}$$

$$C7 : \sum_{i \in \mathcal{I}} \pi_{o_j,i} \; \mu_{o_j,i} \; \omega_{o_j,i,c_j} \leqslant 1, \;\; \forall\, o_j \in \mathcal{O}, \;\; \forall c_j \in \mathcal{C}_j \tag{17g}$$

$$C8 : \mu_{o_j,i} = \{0, 1\}, \;\; \omega_{o_j,i,c_j} = \{0, 1\}, \;\; \forall\, j \in \{\mathcal{B}, \mathcal{H}, \mathcal{U}, \mathcal{S}\}, \;\; \forall\, o_j \in \mathcal{O}, \forall\, c_j \in \mathcal{C}_j, \;\; \forall\, i \in \mathcal{I} \tag{17h}$$

Constraint C1 ensures that a user can only associate with an AN in whose coverage radius the user lies. C2 indicates that an $i_{\mathcal{R}}$ user can be served by only one of either the MBS, LAP, or HAP AN at a time. C3 ensures that an euRLLc user is not attached to the satellite AN. In constraint C4, a feMBB or LDHSC user can associate with only one of the available ANs at any given time. Constraint C5 guarantees served users a minimum QoS in terms of data rate. $\mathcal{I}_{served}$ is a set of users for which $\pi_{o_j,i}\ \mu_{o_j,i}\ \omega_{o_j,i,c_j} = 1$ while $\mathcal{L}^{\upsilon}_{thres}$ is the minimum required data rate of a user demanding service group υ. In C6, the total radio resources allocated to all users attached to an AN o_j must not exceed the AN's radio resource budget Φ_{o_j}. Constraint C7 ensures that an AN's BBU can only be allocated to at most one user, while C8 gives the decision variables that are binary in nature.

To solve the MOOP in (17), the concept of Pareto optimality as defined in [57] is utilized, and this states that:

Definition 1: A point $\Delta_0 \in \Delta$ is said to be Pareto optimal if and only if there is no other point $\Delta_1 \in \Delta$ such that $\Omega(\Delta_1) \geq \Omega(\Delta_0)$, $\Lambda(\Delta_1) \geq \Lambda(\Delta_0)$ and at least one Ω or Λ has been strictly improved.

In simple terms, a point is Pareto Optimal if there is no other point that can improve both Ω and Λ simultaneously. The set of all Pareto optimal points gives an optimal trade-off between Ω and Λ by providing the maximum value of Ω for any given value of Λ and vice verse. The weighted sum method is capable of providing a complete set of Pareto optimal solutions to the MOOP in (17). Therefore, similar to [58], [59], and [60] and owing to its simplicity and low computational complexity; the weighted sum method is adopted to transform the MOOP in (17) into a SOOP, which is a weighted sum of the two objective functions as shown below:

$$\begin{aligned} \max_{\mu_{o_j,i},\ \omega_{o_j,i,c_j}} \quad & \alpha\ \delta_1\ \Omega + (1-\alpha)\ \delta_2\ \Lambda \\ \text{s.t.} \quad & 17a,\ 17b,\ 17c,\ 17d,\ 17e,\ 17f,\ 17g,\ 17h. \end{aligned} \tag{18}$$

The terms δ_1 and δ_2 in (18) are normalization factors associated with Ω and Λ respectively, introduced to put the two functions on the same scale. α in (18) is used to provide a trade-off between data rate maximization and mobility-induced handoff minimization for the mobile users. This parameter is particularly useful since the ANs that maximize data rate may not necessarily minimize handoff, and hence is used to define the importance of the two objective functions. It is important to note that for static users, i.e., the euRLLC and feMBB service groups, α is 1 since these users do not experience mobility-induced handoff. On the other hand, when the values of α are varied in the range [0,1] for the mobile users, the set of all Pareto optimal points to the MOOP in (17) can be obtained [61]. However, in this work, we consider the priority of the mobile LDHMC service group to be handoff minimization, and as such, we set α to 0 for these users.

The optimization problem described in (18) is non-convex and combinatorial, making it NP-hard with no efficient polynomial-time solution. Exact algorithms such as branch and bound exist that return a global optimal solution for such problems. However, these algorithms have to search for all possible solutions (in the worst-case scenario) in the search space. Consequently, the exact algorithms have a high computational complexity that increases exponentially with the network's number of ANs and users. However, the GA has been proven to give near-optimal solutions with a polynomial-time complexity to non-convex and combinatorial problems [30]. Consequently, in this work, we adopt the GA to solve the problem formulated in (18).

IV. PROPOSED SOLUTION

The proposed GA solution is described in this section, but first, a brief review of GA is elucidated.

A. A REVIEW OF THE GENETIC ALGORITHM (GA)

The GA is a search meta-heuristic based on the principle of natural selection in which the fittest individuals of a population are selected to produce children of the next generation. The algorithm starts with generating an initial population consisting of a set of randomly generated solutions, also called chromosomes. Each chromosome is made up of genes, which are essentially the decision variables of the optimization problem.

A fitness function corresponding to the objective function is defined and used to measure the fitness of each chromosome in the population. Parents are selected from the population for reproduction based on their fitness values. In the crossover phase, the parents exchange genes, producing new chromosomes, otherwise called children. The crossover phase is governed by the probability of crossover P_c, which determines whether to consider a child or parent chromosome in the new population.

The mutation operator is used to randomly change one or more genes of the chromosomes to create diversity in the new population and prevent the GA from converging to a local optimum. The mutation is dependent on the probability of mutation P_m, with high values of P_m changing the algorithm to random search. After mutation, the elitism operator ensures that the best solutions/chromosomes in the old population are not lost through crossover and mutation. Therefore, depending on the elitism ratio E_r, a given fraction of the best chromosomes in the old population replace other chromosomes in the new population. The algorithm then terminates when the termination criteria are satisfied, and the chromosome with the best fitness value is the solution to the optimization problem. There are several termination criteria used in literature [62], including; (i) when a fixed number of generations is reached, (ii) when a certain fitness level is reached, and (iii) when there is no improvement in the best fitness value. In this work, the termination criterion is either when there is no improvement in the best fitness value for a given number of consecutive iterations or when a fixed number of generations is reached. The GA process is illustrated in Fig. 4.

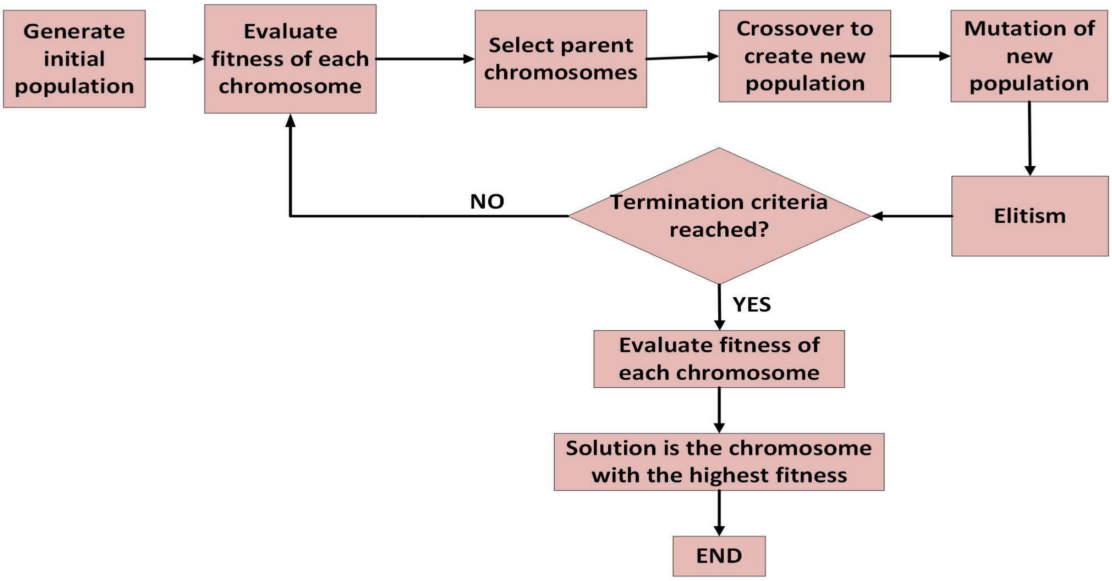

Generate initial population
Evaluate fitness of each chromosome
Select parent chromosomes
Crossover to create new population
Mutation of new population
Elitism
Termination criteria reached?
NO
YES
Evaluate fitness of each chromosome
Solution is the chromosome with the highest fitness
END

FIGURE 4. The GA process.

B. THE PROPOSED GA

A population set $\mathcal{M}$ consisting of M chromosomes is defined. A chromosome $\mu_k \in \mathcal{M}$ is defined as an $|\mathcal{I}|$ dimensional vector that represents a user association solution $\mu_k = (\mu^k_{o_j,1}, \mu^k_{o_j,2}, \ldots, \mu^k_{o_j,|\mathcal{I}|}) \in \mathbb{R}^{|\mathcal{I}|}\ \forall\ \mu_k \in \mathcal{M},\ k \in \{1, 2, \ldots, M\}$, such that $\mu^k_{o_j,i} \in \{1, 2, \ldots, |\mathcal{O}|\}\ \forall\ i \in \mathcal{I}$.

In any given iteration, the fitness value of a gene $\mu^k_{o_j,i}$ of a chromosome μ_k, that represents the optimization problem in (18), is defined depending on the service group of user i.

1) FITNESS VALUE FOR AN euRLLC OR feMBB GENE (DATA RATE MAXIMIZATION FITNESS VALUE)

The fitness value of a gene belonging to an $i_{\mathcal{R}}$ or $i_{\mathcal{E}}$ user is given by

$$\Theta_{\upsilon}(\mu^k_{o_j,i}) = x_{i,k}\left(\delta\,\mathcal{L}_{o_j,i,c_j}\right) - \xi_{i_\upsilon}(1-x_{i,k}),\quad \forall\, i_{\mathcal{R}} \in \mathcal{I}_{\mathcal{R}},\ i_{\mathcal{E}} \in \mathcal{I}_{\mathcal{E}},\ \upsilon \in \{\mathcal{R},\mathcal{E}\} \tag{19}$$

where $\delta\mathcal{L}_{o_j,i,c_j}$ is the normalized data rate a user i can achieve over one BBU c_j of AN o_j. $x_{i,k} \in \{0, 1\}$ denotes the validity status of the gene $\mu^k_{o_j,i}$. $x_{i,k} = 1$ if the user $i's$ association variable/gene $\mu^k_{o_j,i}$ is valid and 0 otherwise. A gene $\mu^k_{o_j,i}$ is valid if the AN o_j is within coverage of user i, is capable of serving the user, and has sufficient resources to meet the QoS requirements of the user. $\xi_{i_\upsilon} \in [0, 1]$ is the penalty cost for not admitting a user of service group $\upsilon \in \{\mathcal{E}, \mathcal{R}\}$.

2) FITNESS VALUE FOR AN LDHMC GENE (HANDOFF MINIMIZATION FITNESS VALUE)

The fitness value for users demanding service group $\mathcal{D}$ prioritizes association to ANs with large cell radius in order to minimize handoff probability. Consequently, the fitness value for these users is expressed as

$$\Theta_{\mathcal{D}}(\mu^k_{o_j,i}) = x_{i,k}\,\frac{R_{o_j}}{\zeta} - \xi_{i_{\mathcal{D}}}(1-x_{i,k}),\ \forall\, i_{\mathcal{D}} \in \mathcal{I}_{\mathcal{D}}, \tag{20}$$

where $\xi_{i_{\mathcal{D}}}$ is a penalty for not admitting an $i_{\mathcal{D}}$ user.

The penalties ξ_{i_υ} in (19) and $\xi_{i_{\mathcal{D}}}$ in (20) are varied depending on the priority of a given service group. We prioritize the euRLLC service group in this work since denial of its service may lead to catastrophic consequences. We also prioritize access to the NTN with a large coverage area for the mobile LDHMC use-case to minimize the probability of handoff. Nonetheless, the priority can also be based on other factors, such as the use-case that yields more revenue to the operator.

For any given gene $\mu^k_{o_j,i}$ of a chromosome μ_k, if $x_{i,k} = 1$, then the user i is allocated its required number of BBUs $\mathcal{N}_i^{BBU}$ given by (21), else, the user is not allocated any resources. We assume that L^{υ}_{thres} is the data rate request from a user demanding a service in the service group $\upsilon \in \{\mathcal{R}, \mathcal{E}, \mathcal{D}\}$.

$$\mathcal{N}_i^{BBU} = \frac{L^{\upsilon}_{thres}}{\mathcal{L}_{o_j,i,c_j}},\quad \forall\, \upsilon \in \{\mathcal{E},\mathcal{R},\mathcal{D}\}, \forall i \in \mathcal{I},\quad \forall o_j \in \mathcal{O}, \forall j \in \{\mathcal{B},\mathcal{U},\mathcal{H},\mathcal{S}\} \tag{21}$$

The overall fitness value of the chromosome μ_k in any given iteration is the summation of all fitness values for the different genes in the chromosome given by:

$$\Theta(\mu_k) = \sum_{i\in\mathcal{I}} \Theta_{\upsilon}(\mu^k_{o_j,i}),\quad \forall\, \upsilon \in \{\mathcal{R},\mathcal{E},\mathcal{D}\} \tag{22}$$

According to (22), all users in the network cooperate and contribute to the fitness value of the chromosome, ultimately leading to fairness between users since the goal is to admit as many users as possible to maximize the fitness value.

The pseudo-code of the proposed GA is given in Algorithm 1. The selection of parents in line 7 is achieved using the roulette wheel technique. In this technique, each chromosome in the population is assigned a probability P_{μ_k} of being selected depicted by (23) that is proportional to its fitness value. The chromosomes with higher values of P_{μ_k} have higher chances of contributing to the creation of the next generation.

$$P_{\mu_k} = \frac{\Theta(\mu_k)}{\sum\limits_{\mu_k \in \mathcal{M}} \Theta(\mu_k)} \tag{23}$$

In lines 8-9, the two selected parents recombine through the crossover operator. In this work, the two-point crossover technique is used [62]. In this technique, two points are randomly selected on both parents, and genes are exchanged between them to create two different chromosomes, otherwise called children. For each created child, if a randomly generated number in the range [0, 1] is less than P_c, then the child is inserted into the new population; otherwise, the parent is. The new chromosomes are then mutated as per line 10 of Algorithm 1. For each gene in the chromosomes created in lines 8-9, if a random number in the range [0,1] is less than P_m, the gene is mutated by replacing it with a random gene $\mu^k_{o_j,i} \in \{1, 2, \ldots, |\mathcal{O}|\}$, otherwise, the gene is not mutated. Line 12 performs the elitism process, which replaces E_r of the chromosomes in the new population with the same fraction of best performing chromosomes in the old population. In line 15, if the fitness value of the best chromosome remains unchanged for a given number of consecutive iterations Q, the algorithm breaks out of the for loop and returns the best/optimal user association solution μ^*_k. Line 17 returns the optimal user association, which is the chromosome in the population with the best fitness value. This solution indicates which ANs the users should be associated with but does not give the number of BBUs that should be allocated to the users to meet their QoS requirements. Therefore, line 18 of Algorithm 1 inputs the optimal user association decision into Algorithm 2, and this returns a list $Assoc^i_{o_j}$ containing the AN o_j (which is the gene $\mu^k_{o_j,i}$) serving user i, and the number of BBUs $\mathcal{N}^{BBU}_i$ determined according to (21) that are allocated to the user.

C. THE OPTIMAL SOLUTION

The optimal solution to the problem in (18) is obtained to validate the optimality of the proposed GA using the Gurobi solver. However, it is important to note the non-linearity of problem (18) introduced by the product of the variables $\mu_{o_j,i}$ and ω_{o_j,i,c_j} in the objective function and constraints. Such a non-linearity hinders our usage of the Gurobi solver. To resolve this issue, the problem is linearised, as detailed in Appendix A. The reformulated problem in (28) is an integer linear programming (ILP) problem whose optimal solution can now be derived using the Gurobi solver via linear programming (LP) relaxation, Branch and Bound (BnB), and other advanced mixed integer programming techniques [63].

Algorithm 1 Genetic Algorithm for Service-Aware User Association and Resource Allocation

```
Input: Size of population $M$, Number of genes $|\mathcal{I}|$, Number of iterations $G$, Number of consecutive iterations $Q$ for stopping criterion, $P_c$, $P_m$, $E_r$, achievable data rate user statistics to the different ANs determined using (10)
Output: User association and resource allocation set, $Assoc^i_{o_j}$
 1: procedure User association and resource allocation
 2: Generate the initial population set $\mathcal{M}$ containing $M$ chromosomes each of length $|\mathcal{I}|$
 3: Calculate the fitness value of each chromosome in $\mathcal{M}$ using (22)
 4:     for $iteration = 1 : G$ do
 5:         Create empty new population set $\mathcal{M}_{new}$
 6:         for $w = 1 : M/2$ do
 7:             Select two parents from $\mathcal{M}$ using the roulette wheel technique
 8:             Carry out crossover on the two parents to create two children
 9:             Insert either the children or parent chromosomes in $\mathcal{M}_{new}$ using $P_c$
10:             For each of the just inserted chromosomes in $\mathcal{M}_{new}$, carry out mutation based on $P_m$
11:         end for
12:         Carry out Elitism using $E_r$
13:         $\mathcal{M} = \mathcal{M}_{new}$
14:         Determine the fitness of each chromosome in $\mathcal{M}$
15:         Break the for loop if fitness value of best chromosome does not change for $Q$ consecutive iterations
16:     end for
17:     return the user association solution $\mu^*_k$
18:     Input $\mu^*_k$ in algorithm 2 to determine the number of BBUs allocated to user $i$ for each association decision $\mu^k_{o_j,i}$ in $\mu^*_k$
19:     Return $Assoc^i_{o_j}$
20: end procedure
```

Consequently, we utilize the Gurobi solver to obtain the optimal solution to the problem defined in (28).

D. COMPLEXITY ANALYSIS

In this section, we analyze the time complexity of the proposed GA versus the optimal solution, the baseline association, and the random user association used as benchmark solutions. The big Omicron (big-O) is employed to characterize the time complexity of the algorithms. The big-O is a mathematical notation that gives a measure of an algorithm's worst-case execution time or required memory in relation to the problem size. A detailed description of the big-O can be found in [64] and [65].

1) COMPUTATIONAL COMPLEXITY OF THE PROPOSED GA

The GA performs the selection, crossover, and mutation operators in each generation. Similar to many roulette wheel

Algorithm 2 Resource Allocation Algorithm

```
Input: Chromosome $\mu_k$, AN available number of BBUs $\mathcal{N}_{oj}^{BBU}$, achievable data rate user statistics to the different ANs determined using (10)
Output: User association and resource allocation set, $Assoc_{oj}^{i}$
procedure Resource allocation
Initialise: $Assoc_{oj}^{i} = \emptyset$
    for Each gene $\mu_{o_j,i}^{k} \in \mu_k$ do
        $i = 1$
        Determine validity status $x_{i,k}$ of $\mu_{o_j,i}^{k}$
        if $x_{i,k} == 1$ then
            Allocate $\mathcal{N}_{i,k}^{BBU}$ BBUs to user $i$ according to (21)
            Deduct $\mathcal{N}_{i}^{BBU}$ from $\mathcal{N}_{oj}^{BBU}$
            Append [$i$, $\mu_{o_j,i}^{k}$, $\mathcal{N}_{i}^{BBU}$] to $Assoc_{oj}^{i}$
        else
            $\mathcal{N}_{i}^{BBU} = 0$
            Append [$i$, $\mu_{o_j,i}^{k}$, $\mathcal{N}_{i}^{BBU}$] to $Assoc_{oj}^{i}$
        end if
        $i = i + 1$
    end for
    Return $Assoc_{oj}^{i}$
end procedure
```

selection routines, a chromosome is selected for reproduction in this work using a search algorithm. Hence, the time complexity of the selection operator is of the order $O(M)$ [66], where M is the population size. The time taken to execute the crossover is proportional to the population size M; hence its time complexity is bounded by $O(M)$. On the other hand, mutation requires that a random number in the range [0,1] is generated for every gene of every chromosome. Therefore, the time complexity of mutation in any given generation is $O(M \times |\mathcal{I}|)$, where $|\mathcal{I}|$ is the number of genes in a chromosome, which also corresponds to the number of users in the ITNTN. Also, the summation of all genes' fitness values gives a chromosome's fitness value. Therefore, the time complexity due to the evaluation of fitness values of all chromosomes in a generation is $O(M \times |\mathcal{I}|)$. The overall time complexity of the proposed GA is therefore given by $O(G \times (M + M + M \times |\mathcal{I}| + M \times |\mathcal{I}|)) = O(G \times M \times (2 + 2 \times |\mathcal{I}|))$, where G is the number of generations. This time complexity can be reduced to $O(G \times M \times |\mathcal{I}|)$. Consequently, the proposed GA has a polynomial time complexity of the order $O(G \times M \times |\mathcal{I}|)$.

2) COMPUTATIONAL COMPLEXITY OF THE OPTIMAL SOLUTION

The problem defined by (28) represents a variant of the well-known knapsack problem, which is NP-complete [67]. The optimal solution to such a problem can be obtained via LP relaxation and BnB, but this requires exponential upper bound time complexity in tandem with the problem size [67], [68]. Since the decision variables are binary, the problem's search space has a size of two to the power of the number of binary variables. Therefore, the time complexity of the optimal solution is given by $O(2^{|\mathcal{I}| \times (|\mathcal{B}| \times |\mathcal{C}_{\mathcal{B}}| + |\mathcal{U}| \times |\mathcal{C}_{\mathcal{U}}| + |\mathcal{H}| \times |\mathcal{C}_{\mathcal{H}}| + |\mathcal{S}| \times |\mathcal{C}_{\mathcal{S}}|)})$, where $|\mathcal{B}|$, $|\mathcal{U}|$, $|\mathcal{H}|$, $|\mathcal{S}|$, denote the number of ANs in the MBS, LAP, HAP, and SatComs RAN respectively. On the other hand, $|\mathcal{C}_{\mathcal{B}}|$, $|\mathcal{C}_{\mathcal{U}}|$, $|\mathcal{C}_{\mathcal{H}}|$, and, $|\mathcal{C}_{\mathcal{S}}|$ represent the number of BBUs owned by an AN in the MBS, LAP, HAP, and SATCOMs RAN respectively. Since the worst-case time complexity of the optimal solution is exponential, algorithms that yield near-optimal solutions but with polynomial complexity should be considered; hence the GA proposed in this work.

3) COMPUTATIONAL COMPLEXITY OF THE BASELINE AND RANDOM USER ASSOCIATION

In this work, we also analyze the baseline and random user association (RUA) schemes as benchmark solutions.

The baseline association, also referred to as the greedy algorithm in this work, associates users with ANs based on maximum SINR. The description of the greedy algorithm, together with its pseudo code, is given in our work [32]. The time complexity for the computation of the SINR values from users to all ANs is of the order $O(|\mathcal{I}| \times |\mathcal{O}|)$ where $|\mathcal{O}|$ is the total number of ANs in the ITNTN. For each user, the SINR values to the different ANs must be sorted such that if the AN with the highest SINR does not have sufficient resources to meet the user's QoS requirements, then the user is associated with the next best AN. The time complexity due to sorting is $O(|\mathcal{O}| \log |\mathcal{O}| \times |\mathcal{I}|)$. Therefore, the overall complexity of the greedy algorithm is given as $O(|\mathcal{I}| \times |\mathcal{O}| + |\mathcal{O}| \log |\mathcal{O}| \times |\mathcal{I}|)$ which can be reduced to $O(|\mathcal{I}| \times |\mathcal{O}| \times \log |\mathcal{O}|)$. As the population size M and the number of generations G of the GA are usually greater than the number of access nodes $|\mathcal{O}|$, it is clear that the greedy algorithm has a much shorter worst-case running time than the proposed GA. However, as the results will show in section V, the GA achieves a better performance as far as maximizing the objective function in (18) is concerned.

On the other hand, the RUA approach associates users randomly with any available AN. Such an algorithm has the shortest worst-case running time, which is proportional to the number of users. Hence, the time complexity of the RUA algorithm is given by $O(|\mathcal{I}|)$. Important to note is that, like the GA, both the greedy and RUA algorithms prioritize (1) the euRLLC use-case and (2) the use of NTNs over the TNs for service provisioning of the mobile LDHMC users.

V. RESULTS AND PERFORMANCE EVALUATION

In this section, the performance of the proposed GA is compared to the optimal, the greedy [32] and random user association (RUA) solutions. First, the main simulation parameters are presented, and after, the results and their discussion.

A. SIMULATION ASSUMPTIONS

We consider a circular urban region of 3 km radius that is within the coverage of a LEO satellite and a HAP AN and

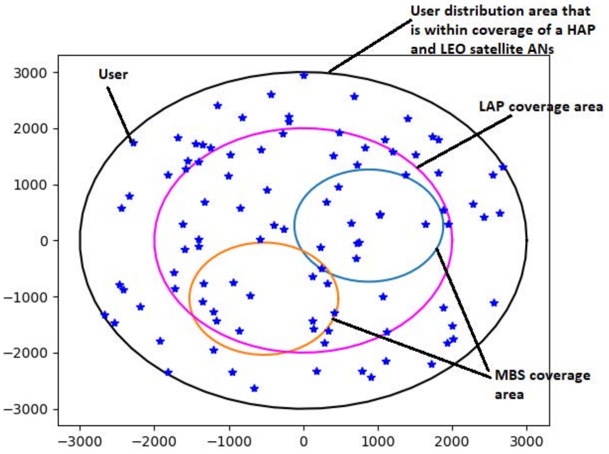


FIGURE 5. Network deployment.

TABLE 2. Simulation parameters and values.

Parameter	Value
$f_{o_B}, f_{o_U}, f_{o_H}, f_{o_S}$	[4, 2, 3, 5] GHz
$\mathcal{T}_{c_\mathcal{B}}, \mathcal{T}_{c_\mathcal{U}}, \mathcal{T}_{c_\mathcal{H}}, \mathcal{T}_{c_\mathcal{S}}$	[0.18, 0.18, 1, 2] MHz]
Number of BBUs for $[o_B, o_U, o_H, o_S]$	[10, 10, 20, 20]
$[P^{thres}_{o_B}, P^{thres}_{o_U}, P^{thres}_{o_H}, P^{thres}_{o_S}]$	[8, 5, 20, 25] Watts
$x,\ y,\ \eta_{LoS},\ \eta_{NLoS}$	[10.39, 0.05, 1, 20]
Shadow fading $[\mathcal{B}, \mathcal{S}]$	[8, 4] dB
$[CL, PL^k,\ PL^y]$	[0, 0, 23 dB]
AN height $[z_b, z_u, z_h, z_s]$	[40m, 2km, 17km, 600km]
Service group user ratio $[\lvert\mathcal{I}_\mathcal{R}\rvert : \lvert\mathcal{I}_\mathcal{D}\rvert : \lvert\mathcal{I}_\mathcal{E}\rvert]$	[0.3:0.1:0.6]
$L^v_{thres} v \in [\mathcal{R}, \mathcal{D}, \mathcal{E}]$	[500,1000,1000] kbps
Noise spectral density	-174 dBm

also contains 1 LAP AN and MBSs with a radius of 2 km and 1 km respectively. Fig (5) is an example of the network deployment with 2 MBSs in the considered user distribution area. Users within the considered region are uniformly and randomly distributed. Table 2 [9], [24], [43], [44] gives the radio environment parameters used to validate the proposed solution. Given that the performance of the GA is highly dependent on the probability of crossover P_c, probability of mutation P_m and population size $\mathcal{M}$, the values of these parameters are determined before results analysis.

B. GA PARAMETER SETTING

The appropriate parameters used in the proposed GA solution are identified in this sub-section. Fig. 6 shows the effect of P_c on the GA convergence. P_m, M and the number of users $|\mathcal{I}|$ are fixed at 0.1, 50 and, 80 respectively, while P_c is varied from 0.2 to 1, with increments of 0.2. It is observed that the higher the value of P_c, the larger the fitness value; hence, the better the solution found by the GA is at satisfying the objective function. Since $P_c = 0.8$ performed well as per Fig. 6, it is chosen to be used in this work.

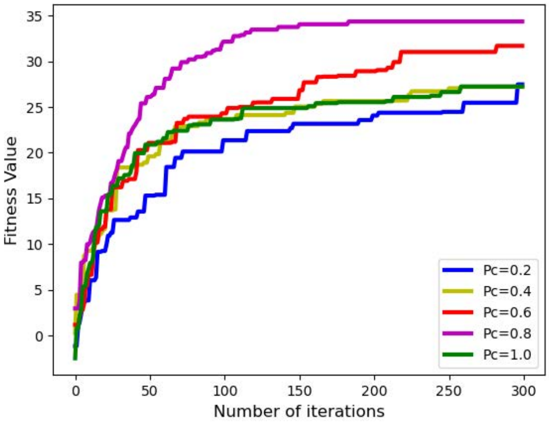


FIGURE 6. Effect of probability of crossover on GA convergence.

Next, the effect of P_m on convergence is analysed by setting $P_c = 0.8$, $M = 50$ and $|\mathcal{I}| = 80$. Fig. 7 shows that the higher the value of P_m, the worse the performance of the GA, as the algorithm is transformed into random search. P_m is set to 0.1 since its fitness value and convergence rate is much better than any other value of P_m as can be observed in Fig. 7.

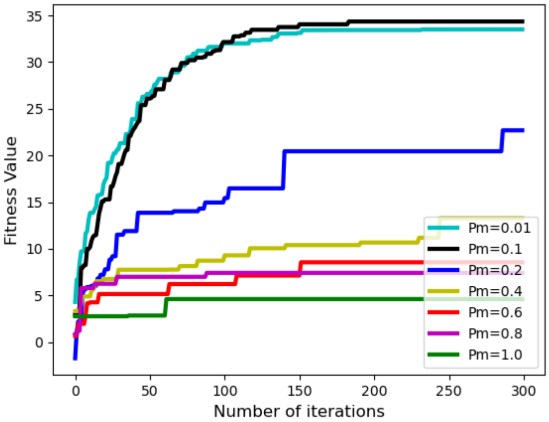


FIGURE 7. Effect of probability of mutation on GA convergence.

In Fig. 8, the effect of population size M on the GA convergence is analysed, with $P_c = 0.8$, $P_m = 0.1$ and $|\mathcal{I}| = 80$. The figure shows that convergence speed increases with the population size M. Also, we observe that convergence is achieved by the 200th iteration for all population sizes. In the following section, we set the population size M to 50 and the number of iterations G to 150. These parameters are chosen to strike a balance between the accuracy and computational complexity of the GA, as both increase with M and G. Table 3 gives the parameters used for the GA.

C. SIMULATION RESULTS

To validate the performance of the proposed GA, we simulate the optimal solution based on the gurobi solver of the problem

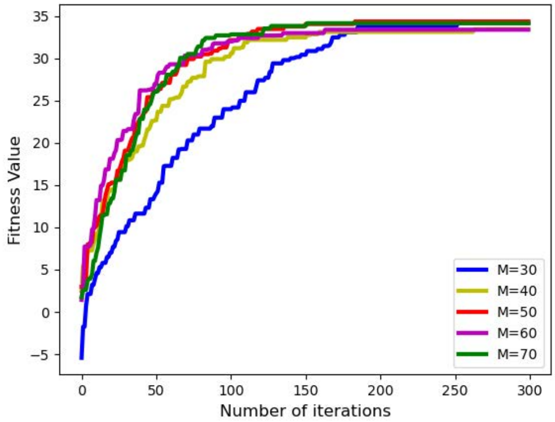


FIGURE 8. Effect of population size on GA convergence.

TABLE 3. GA parameters.

Parameter	Value
Probability of crossover P_c	0.8
Probability of mutation P_m	0.1
Population size M	50
Number of iterations G	150

described in (28). In addition, we compare the proposed GA with the greedy algorithm [32] and the random user association (RUA) scheme.

Performance evaluation is based on three main metrics; the acceptance ratio (AR), the spectrum efficiency (SE), and the handoff probability. In this work, the user AR quantifies the ratio of served users to the total number of users in the network. On the other hand, the SE is the ratio of the overall system data rate to the total network bandwidth [69], while we define the probability of handoff as the ratio of the number of users that experienced a handoff (and are thus served by another AN) to the total number of mobile users during a given TTI.

1) IMPACT OF TRADE-OFF FACTOR α

First, we analyze the effect of the trade-off term α on data rate maximization (objective 1) and handoff minimization (objective 2) in (18). As α affects only mobile users, for this analysis, we consider 20 LDHMC users in a network comprising of 5 ANs, i.e., 2 MBSs, 1 LAP, 1 HAP, and 1 SatComs AN.

Fig. 9 depicts that in solving the MOOP in (17) as a SOOP in (18) for varying values of α, a set of Pareto-optimal solutions exist. These solutions are generated using Algorithm 1 for values of α ranging from 0 to 1 with an increment size of approximately 0.0526. As Fig 9 shows, the generated Pareto-optimal solutions form a Pareto-optimal front below which the region comprises suboptimal solutions, and above which are infeasible solutions. From the figure, the points are concentrated at both ends of the curve. This shows that for the mobile users, α acts in a manner as to either maximize data rate or minimize mobility-induced handoff. When the value of α is lower than 0.5, function two of (18) is maximized, which ultimately minimizes the handoff probability, and as α increases beyond 0.5, then function one is maximized consequently maximizing data rate. This observation is further supported by Figs. 10 and 11. In Fig. 10, the data rate is low for low values of α, and a step to higher values of data rate is observed at $\alpha \approx 0.5$. In the same manner, in Fig. 11, the probability of handoff is approximately 0 for $\alpha < 0.5$ when objective two is prioritized, and once α increases above 0.5, the handoff probability increases since the priority becomes data rate maximization, and the nodes that maximize data rate are not necessarily the same as those that minimize mobility-induced handoff. The instability observed in both Figs. 10 and 11 for $\alpha > 0.5$ is caused by motion of the users. The users keep moving at different velocities out and into coverage of different ANs resulting in unstable achieved total data rate and handoffs experienced. Since α either maximizes the data rate or minimizes the mobility-induced handoff, in all the following simulations, we assume that the objective of the mobile users is to minimize the handoff probability and, as such, set $\alpha = 0$ for the LDHMC service group.

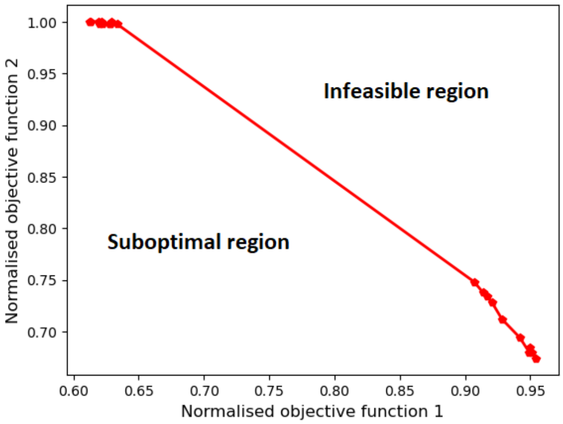


FIGURE 9. Pareto-optimal front of the MOOP in (17).

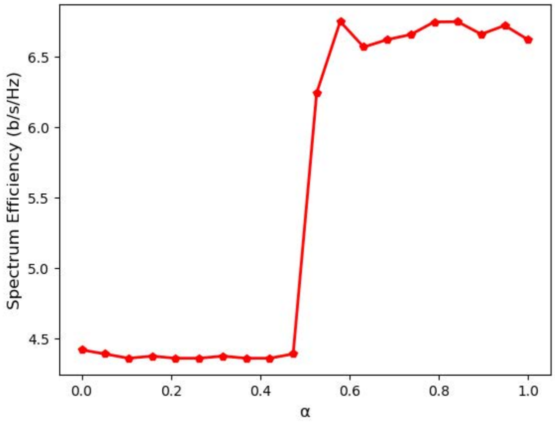


FIGURE 10. Spectrum efficiency with varying α.

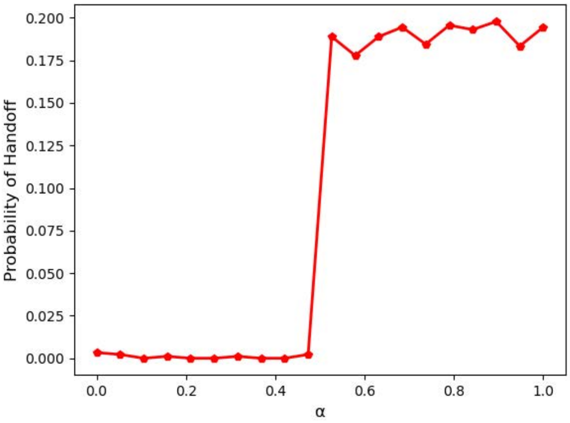


FIGURE 11. Probability of handoff with varying α.

2) IMPACT OF USER DENSITY

We then evaluate the proposed algorithm's performance while varying the number of users in the network. We maintain the number of ANs at 5, with 2 MBSs, 1 LAP, 1 HAP, and 1 SatComs AN.

In Fig. 12, we analyze the user AR performance of all algorithms. Generally, as the number of users in the network increases, the AR reduces due to resource scarcity. On average, the AR achieved by the GA, optimal, greedy, and RUA algorithms is 0.87, 0.86, 0.85, and 0.84, respectively. The proposed GA achieves an AR that is better than the optimal, greedy, and RUA solutions by 0.71%, 2.02%, and 2.75%, respectively. The GA performs better because, unlike the greedy and RUA algorithms, it is optimized to consider all different user association possibilities, thereby serving users with fewer ANs within their coverage first. Also, since the fitness value of the GA increases with the number of users admitted to the network, the proposed algorithm performs slightly better than the optimal algorithm that focuses on maximizing the data rate without regard to the AR. The greedy and RUA algorithms have more or less the same performance since we consider a network with a small number of nodes. Therefore, there is a high chance of selecting the same node within a user's coverage, whether by random or through the use of maximum SINR.

Fig. 13 depicts the performance of the different algorithms with respect to SE. As the number of users in the network increases, the total data rate increases, thus increasing the achieved SE. However, at about 60 users and above, the SE remains constant since the available resources become insufficient to meet the QoS requirements of all users. Ultimately, all algorithms saturate, as the achieved total network data rate remains constant irrespective of the number of users in the network. The SE achieved on average by the GA, optimal, greedy, and RUA is 10.79, 10.79, 10.28, and 10.25 b/s/Hz respectively. The performance of the GA closely follows that of the optimal solution, outperforming the greedy and RUA algorithms by 4.81% and 5.094% on average, respectively. It is important to note that the RANs in the ITNTN have BBUs of different bandwidths. Therefore, a RAN may have a higher SINR, but because of a smaller sized BBU, it achieves less data rate than another RAN with a bigger sized BBU. Therefore, an association based on maximum SINR does not guarantee the maximum data rate in the ITNTN, thereby leading to a lower SE when compared to the proposed GA, whose value function is based on maximizing the achieved data rate. The performance of the RUA is also lower than the GA, as this algorithm associates users randomly to any available capable RAN without any regard for the achieved data rate. The excellent performance of both the GA and the optimal solutions is because both these solutions are based on the maximization of the data rate of the ITNTN.

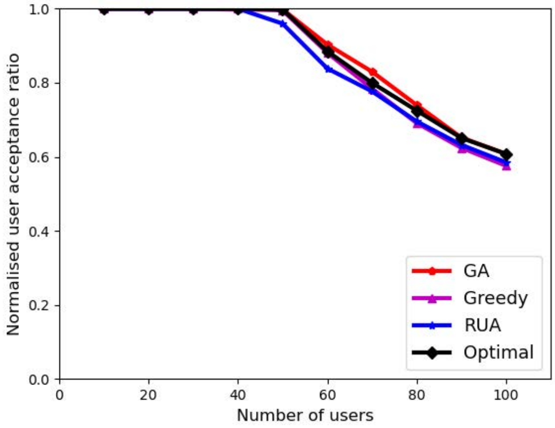


FIGURE 12. User acceptance ratio with varying number of users.

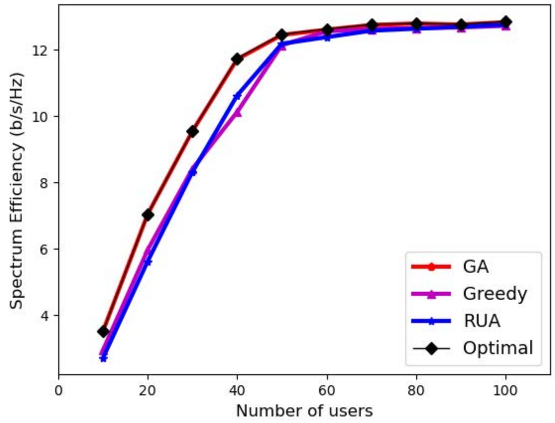


FIGURE 13. Spectrum efficiency with varying number of users.

In Fig. 14, the euRLLC user acceptance ratio performance is depicted. The euRLLC users are prioritized over other users for all four algorithms, as it is vital to mitigate call blocks for this use-case. Consequently, for the number of users from 10 to 70, all euRLLC users are accepted into the network since resources are still available to meet their needs. For the number of users beyond 70, a fraction of euRLLC users is dropped due to limited resources to serve all mission-critical users. The GA and optimal algorithms have the same performance with an average AR of 0.99, performing better than the greedy and RUA, with average euRLLC AR of 0.98 and 0.97, respectively. The GA and optimal solutions perform better than the greedy and RUA due to their intelligence in considering all the options necessary to mitigate euRLLC user call blocks.

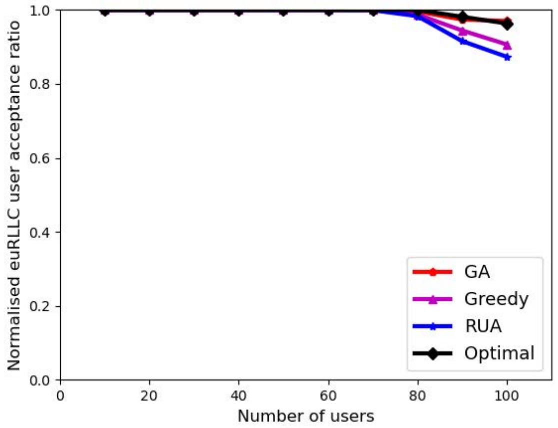


FIGURE 14. euRLLC acceptance ratio with varying number of users.

The large coverage NTNs were prioritized in all four algorithms to serve the mobile LDHMC service group. Therefore, as is depicted by Fig. 15, the AR for this service group is 1 for all algorithms and all number of users. This is because of available resources in the network to serve this service class. On the other hand, Fig. 16 shows the acceptance ratio of the feMBB use-case. In this work, the priority of this use-case is lower than other use-cases, hence the steep decline in AR beyond 50 users when the resources in the network are no longer enough to serve all users. Because of its intelligence in considering the different available AN association options, the GA has a better feMBB of 0.79 on average than the optimal, greedy, and RUA, which have an average of 0.78, 0.76, and 0.76, respectively.

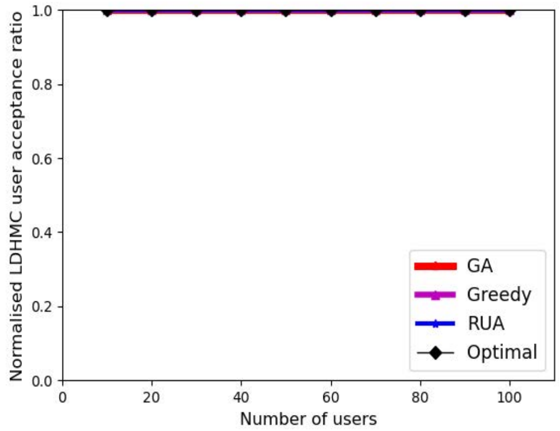


FIGURE 15. LDHMC acceptance ratio with varying number of users.

We analyze the handoff performance of all algorithms in Fig. 17. The optimal and GA algorithms can associate the LDHMC mobile users to the ANs with the largest cell radius and thus achieve a handoff probability of 0 for all numbers of users in the network. On the other hand, the handoff probability for the greedy and RUA algorithms is worse by 16.7% and 41.3% on average, respectively. This is because the LDHMC user association keeps alternating between the NTN ANs depending on the maximum SINR for the greedy algorithm and randomly for the RAU algorithm. The handoff probability reduces with the number of users for these two algorithms because resources of smaller radius ANs are depleted, and users are now forced to associate with the large coverage cells.

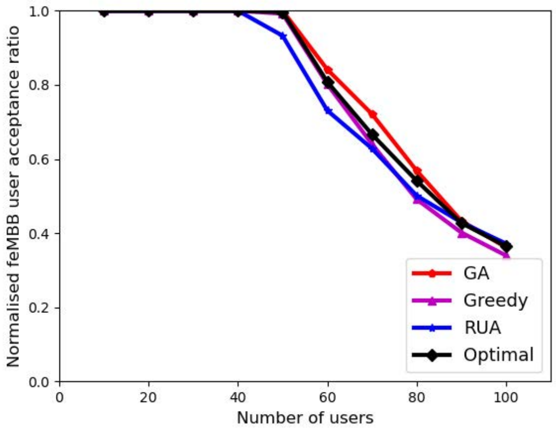


FIGURE 16. feMBB acceptance ratio with varying number of users.

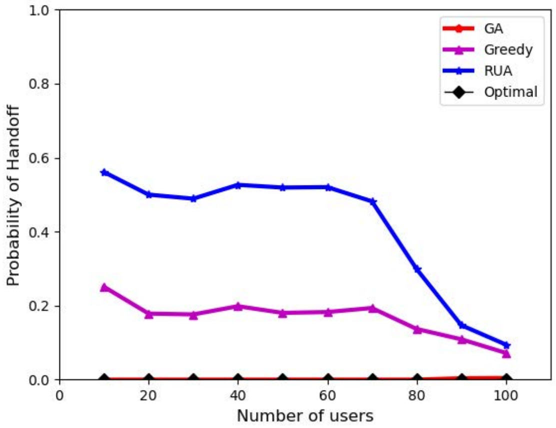


FIGURE 17. Probability of handoff with varying number of users.

3) IMPACT OF ACCESS NODES DENSITY

Next, we analyze the impact of AN density on the four different algorithms. We vary the number of MBS from 1 to 6 while maintaining the number of LAPs, HAPs, and SatComs to 1 AN each. The number of users in the network is maintained at 80.

Fig. 18 shows that as the number of ANs in the network increases, the SE also increases since resources to support the users' data rate requirements keep increasing. The SE performance of the GA is within an average of 0.4% of the optimal SE and outperforms the greedy and RUA algorithms by 1.23% and 0.97% on average, respectively. Fig. 19 shows that the GA outperforms all the other three algorithms in terms of AR, with an average AR of 0.867 compared to 0.855, 0.774, 0.801 of the optimal, greedy, and RUA, respectively, translating to 1.41%, 10.8%, and 7.6% better performance respectively. Since all users in the network cooperatively contribute to the fitness value of the GA algorithm, this solution maximizes the number of admitted users in the network and hence achieves a higher AR. The greedy algorithm has the worst performance in terms of SE and AR since the association is based on maximum SINR without regard for the user and AN distribution, and hence lacks the intelligence of first associating users that are within the coverage of few ANs.

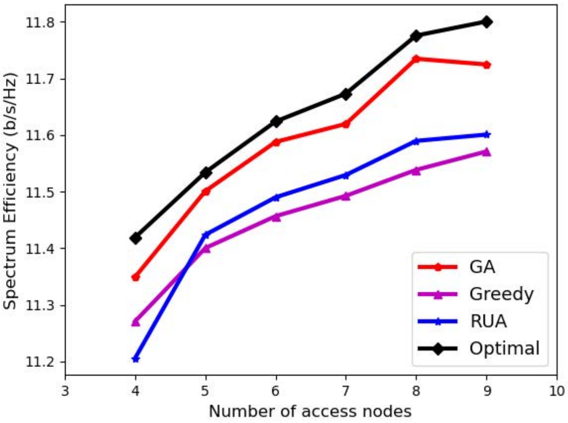


FIGURE 18. Spectrum efficiency with varying number of access nodes.

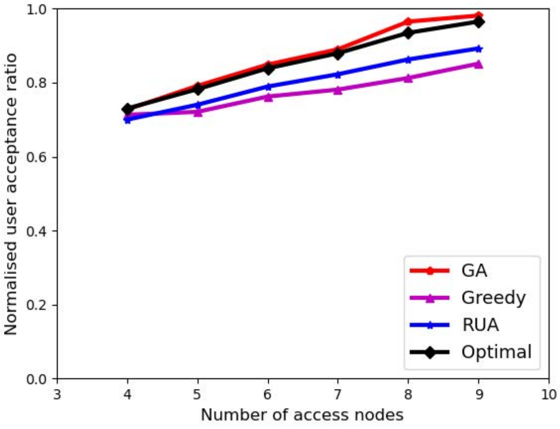


FIGURE 19. User acceptance ratio with varying number of access nodes.

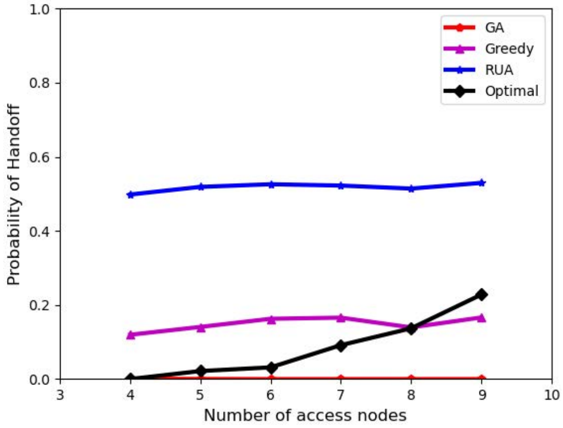


FIGURE 20. Probability of Handoff with varying number of access nodes.

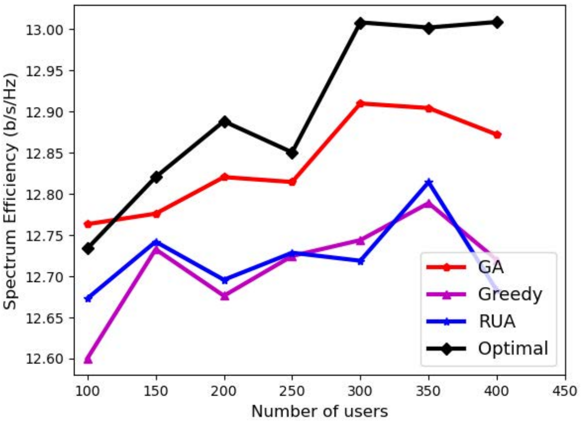


FIGURE 21. Spectrum efficiency with varying number of users.

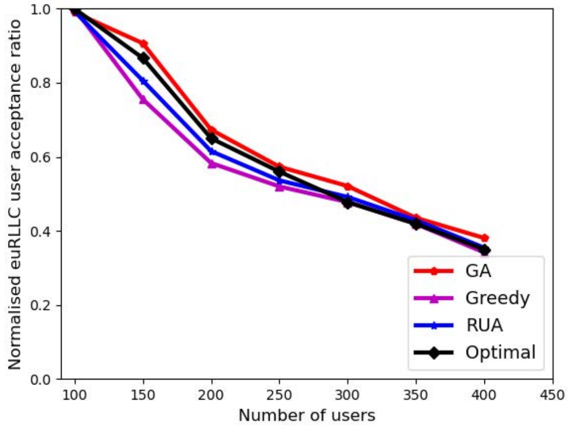


FIGURE 22. euRLLC acceptance ratio under network overload conditions.

Moreover, Fig. 20 also shows that as the number of ANs in the network increases, the GA achieves superior performance in terms of handoff probability. The handoff performance of the optimal algorithm falls below that of the GA by 8.4% on average since this algorithm chooses to maximize the total data rate in the network at the expense of minimizing handoff. On the other hand, the GA performs better than the greedy and RUA in terms of handoff probability by 14.9% and 51.8% on average, respectively. This is because the greedy chooses an NTN AN with the highest SINR in each TTI, while the RUA chooses any NTN AN at random. Consequently, the ANs chosen for the association by the greedy and RUA algorithms keep changing in each iteration, while the GA fitness value is formulated so that the mobile LDHMC users are associated with the AN with the largest cell radius in each TTI. These results demonstrate that the proposed GA is well suited for future scenarios characterized by highly mobile users for which increased handoff implies increased delays and a high probability of call drops due to handoff failure, consequently degrading the QoS of the users.

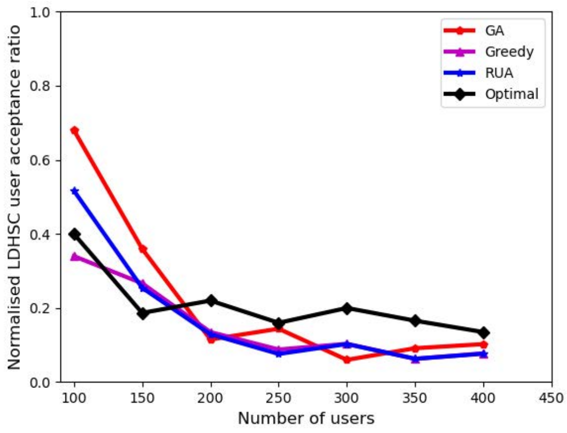


FIGURE 23. LDHMC acceptance ratio under network overload conditions.

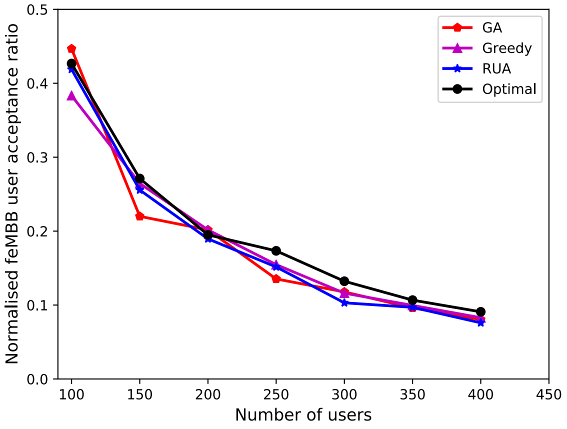


FIGURE 24. feMBB acceptance ratio under network overload conditions.

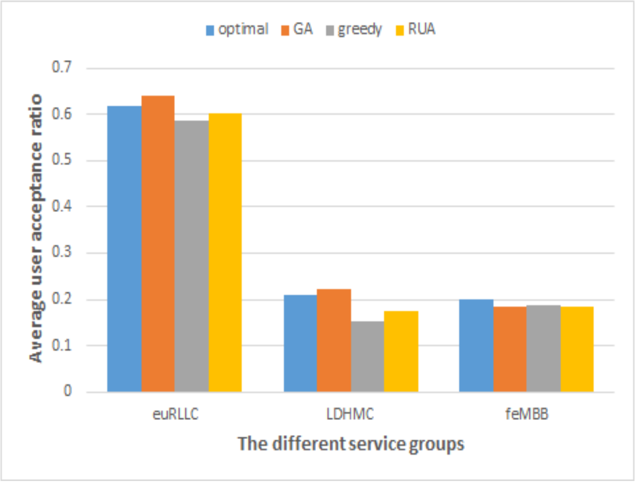


FIGURE 25. Average AR of the different use-cases under network overload conditions.

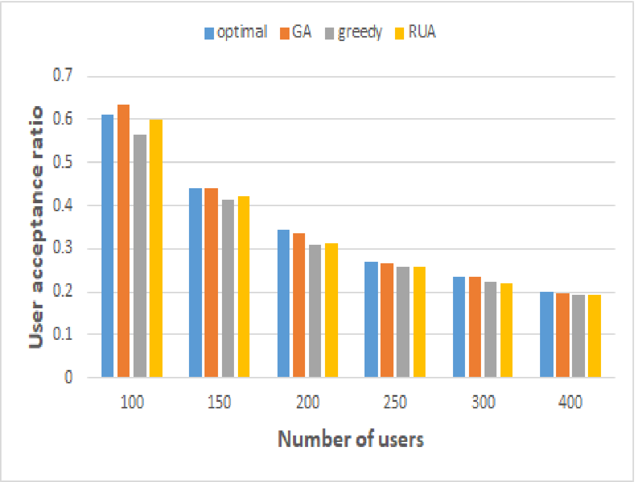


FIGURE 26. User AR under network overload conditions.

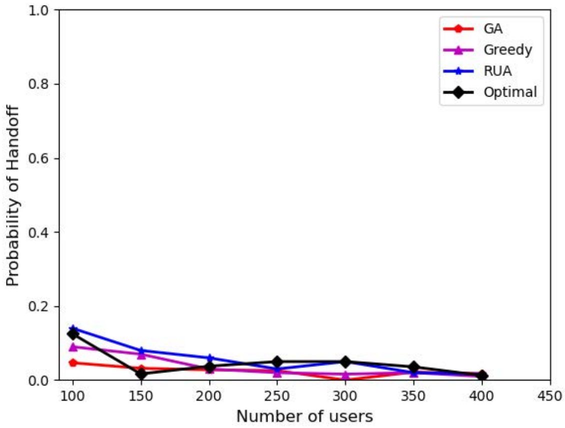


FIGURE 27. Mobility-induced handoff under network overload conditions.

4) IMPACT OF NETWORK OVERLOAD

In the previous simulations, the argument is that the nodes with the broadest coverage should be prioritized to serve the LDHMC service group to reduce mobility-induced handoff. However, in this section, we analyze the performance of the proposed algorithm in a network experiencing overload conditions. To best evaluate this scenario, we decided to give the same priority to both the feMBB and LDHMC service groups so as to have a fair comparison in overloading conditions for both use-cases. The euRLLC use-case is still prioritized, and the simulation is performed considering 5 ANs, i.e., 2 MBS, 1 LAP, 1 HAP, and 1 Satcoms AN. The objective is to analyze how the system responds to the distribution of resources to the different use-cases.

First, Fig. 21 shows that for all algorithms, there is a small change in the SE as the number of users increases in the overloading conditions. In this condition, the limit on the number of users the network can support has already been reached, beyond which the network observes only a slight variation in SE. Nonetheless, on average, the GA still outperforms the greedy and RUA algorithms by 0.97% and 0.9% and is within 0.5% of the optimal SE.

The Figs. 22, 23, and 24 depict the AR of the euRllC, LDHMC, and feMBB use-cases respectively. In all figures, the AR of the respective use-cases decreases as the number of users keeps increasing beyond the threshold that the network can support. Fig. 25 is a combination of Figs. 22, 23, and 24, showing the average AR of the different algorithms, for the different use-cases.

As can be observed in Fig. 25, All algorithms prioritize the euRLLC use-case over the other two, with the GA being better than the optimal, greedy, and random algorithms by 3.5%, 8.55%, and 5.68% respectively. This continues to show the strength of the formulated fitness function of the GA in prioritizing mission-critical users.

Since the LDHMC and feMBB have the same priority, we observe an almost similar AR for the two use-cases for all algorithms in Fig. 25. Fig 26 shows the AR of all algorithms in overloading conditions. The GA is able to still outperform the optimal, greedy, and RUA algorithms by 0.37%, 7%, and 4.78%, respectively, and yet still maintain a low mobility-induced handoff probability as observed by Fig. 27.

VI. CONCLUSION

In this paper, we have formulated a user association and resource allocation problem in the ITNTN as a MOOP that maximized the total network data rate and minimized the mobility-induced handoff. Moreover, the mission-critical euRLLC service group provisioning was prioritized over other service groups. The MOOP was transformed into a weighted sum SOOP, which was solved using the GA. In the GA, service group-dependent fitness functions were formulated to determine the near-optimal user association and resource allocation solution. The Simulation results showed that for an increasing number of ANs, the proposed GA's SE is within 0.4% of the optimal solution. At the same time, the GA's user AR and handoff probability outperformed the optimal, the greedy SINR association-based, and the random user association solutions. While the greedy and RUA algorithms are characterized by a shorter running time compared to the GA, the above results show that the GA achieves a better SE and lower probability of handoff. In future work, we plan to investigate the proposed service-aware user association and resource allocation in the ITNTN based on reinforcement learning (RL). In RL, once the training is done, the agent can make decisions in real-time, a considerable advantage for wireless networks, especially those serving mission-critical users.

APPENDIX A

The multiplication of the two decision variables $\mu_{o_j,i}$ and ω_{o_j,i,c_j} introduces a non-linearity in the objective function and constraints of the optimization problem formulated in (18). Similar to [70], a linearisation term is introduced that replaces the product with a single binary variable as depicted by (24) to avoid this non-linearity.

$$\Psi_{o_j,i,c_j} = \mu_{o_j,i}\omega_{o_j,i,c_j}, \quad \forall\, o_j \in \mathcal{O}, \ \ \forall\, i \in \mathcal{I}, \ \ \forall\, c_j \in \mathcal{C}_j \tag{24}$$

$\Psi_{o_j,i,c_j} \in \{0, 1\}$ in (24) is a binary decision variable that is one when a user i is associated with AN o_j and allocated a BBU c_j, while it is zero otherwise. Subsequently, the product $\mu_{o_j,i}\omega_{o_j,i,c_j}$ in the objective function and constraints of (18) is replaced with Ψ_{o_j,i,c_j}. Furthermore, the linearised optimization problem must include additional constraints that establish the relationship between Ψ_{o_j,i,c_j}, $\mu_{o_j,i}$, and ω_{o_j,i,c_j}; defined as

$$\Psi_{o_j,i,c_j} \leqslant \mu_{o_j,i}, \quad \forall\, o_j \in \mathcal{O}, \ \ \forall\, i \in \mathcal{I}, \ \ \forall\, c_j \in \mathcal{C}_j \tag{25}$$

$$\Psi_{o_j,i,c_j} \leqslant \omega_{o_j,i,c_j}, \quad \forall\, o_j \in \mathcal{O}, \ \ \forall\, i \in \mathcal{I}, \ \ \forall\, c_j \in \mathcal{C}_j \tag{26}$$

$$\Psi_{o_j,i,c_j} \geqslant \mu_{o_j,i} + \omega_{o_j,i,c_j} - 1, \quad \forall\, o_j \in \mathcal{O}, \ \ \forall\, i \in \mathcal{I}, \ \ \forall\, c_j \in \mathcal{C}_j \tag{27}$$

Henceforth, the modified integer linear programming (ILP) problem is formulated as

$$\begin{aligned} \max_{\Psi_{o_j,i,c_j}} \quad & \alpha\, \delta_1 \sum_{o_j\in\mathcal{O}} \sum_{i\in\mathcal{I}} \sum_{c_j\in\mathcal{C}_j} \rho_{i_\upsilon}\ \Psi_{o_j,i,c_j}\ \mathcal{L}_{o_j,i,c_j} \\ & + (1-\alpha)\, \delta_2 \sum_{o_j\in\mathcal{O}} \sum_{i\in\mathcal{I}} \sum_{c_j\in\mathcal{C}_j} \rho_{i_\upsilon}\ \Psi_{o_j,i,c_j}\ \frac{R_{o_j}}{\zeta} \end{aligned} \tag{28}$$

s.t.

$$C1 : \mu_{o_j,i} \leq \pi_{o_j,i}, \quad \forall\, o_j \in \mathcal{O}, \ \ \forall\, i \in \mathcal{I} \tag{28a}$$

$$C2 : \sum_{o_j\in\{\mathcal{B}\,\cup\,\mathcal{U}\,\cup\,\mathcal{H}\}} \pi_{o_j,i}\ \Psi_{o_j,i,c_j} \ \leqslant \ 1, \forall\, i_{\mathcal{R}} \in \mathcal{I}_{\mathcal{R}} \tag{28b}$$

$$C3 : \sum_{o_j\in\{\mathcal{S}\}} \pi_{o_j,i}\ \Psi_{o_j,i,c_j} \ = \ 0, \quad \forall\, i_{\mathcal{R}} \in \mathcal{I}_{\mathcal{R}} \tag{28c}$$

$$\begin{aligned} C4 : & \sum_{o_j\in\mathcal{O}} \pi_{o_j,i}\ \Psi_{o_j,i,c_j} \ \leqslant \ 1, \ \ \forall\, i_{\mathcal{E}} \in \mathcal{I}_{\mathcal{E}}, \\ & \forall\, i_{\mathcal{D}} \in \mathcal{I}_{\mathcal{D}} \end{aligned} \tag{28d}$$

$$\begin{aligned} C5 : & \sum_{o_j\in\mathcal{O}} \sum_{c_j\in\mathcal{C}_j} \pi_{o_j,i}\ \Psi_{o_j,i,c_j} \mathcal{L}_{o_j,i,c_j} \geqslant \mathcal{L}^{\upsilon}_{thres}, \\ & \forall i \in \mathcal{I}_{served}, \ \ \forall\, \upsilon \in \{\mathcal{E}, \mathcal{R}, \mathcal{D}\} \end{aligned} \tag{28e}$$

$$C6 : \sum_{i\in\mathcal{I}} \sum_{c_j\in\mathcal{C}_j} \pi_{o_j,i}\ \Psi_{o_j,i,c_j} \mathcal{T}_{c_j} \leq \Phi_{o_j}, \quad \forall\, o_j \in \mathcal{O} \tag{28f}$$

$$C7 : \sum_{i\in\mathcal{I}} \pi_{o_j,i}\ \Psi_{o_j,i,c_j} \leqslant 1, \ \ \forall\, o_j \in \mathcal{O}, \ \ \forall c_j \in \mathcal{C}_j \tag{28g}$$

$$C8 : \Psi_{o_j,i,c_j} \leqslant \mu_{o_j,i}, \ \ \forall\, o_j \in \mathcal{O}, \ \ \forall\, i \in \mathcal{I}, \ \ \forall\, c_j \in \mathcal{C}_j \tag{28h}$$

$$C9 : \Psi_{o_j,i,c_j} \leqslant \omega_{o_j,i,c_j}, \ \ \forall\, o_j \in \mathcal{O}, \ \ \forall\, i \in \mathcal{I}, \ \ \forall\, c_j \in \mathcal{C}_j \tag{28i}$$

$$\begin{aligned} C10 : & \Psi_{o_j,i,c_j} \geqslant \mu_{o_j,i} + \omega_{o_j,i,c_j} - 1, \\ & \forall\, o_j \in \mathcal{O}, \ \ \forall\, i \in \mathcal{I}, \ \ \forall\, c_j \in \mathcal{C}_j \end{aligned} \tag{28j}$$

$$\begin{aligned} C11 : & \mu_{o_j,i} = \{0, 1\}, \ \omega_{o_j,i,c_j} = \{0, 1\}, \ \Psi_{o_j,i,c_j} = \{0, 1\} \\ & \forall\, j \in \{\mathcal{B}, \mathcal{H}, \mathcal{U}, \mathcal{S}\}, \ \ \forall\, o_j \in \mathcal{O}, \ \ \forall\, c_j \in \mathcal{C}_j, \ \ \forall\, i \in \mathcal{I} \end{aligned} \tag{28k}$$

Important to note is that the data rate $\mathcal{L}_{o_j,i,c_j}$ achieved via a BBU c_j by a user i associated to AN o_j is a constant since this term is computed before hand. Also the terms $\pi_{o_j,i}$ and ρ_{i_υ} are known before hand.

REFERENCES

[1] Y. Cao, S.-Y. Lien, and Y.-C. Liang, "Deep reinforcement learning for multi-user access control in non-terrestrial networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1605–1619, Mar. 2021.

[2] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2334–2360, 3rd Quart., 2019.

[3] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A comprehensive survey of RAN architectures toward 5G mobile communication system," *IEEE Access*, vol. 7, pp. 70371–70421, 2019.

[4] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[5] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.

[6] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, and C. Dehos, "6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 42–50, Sep. 2019.

[7] B. Zong, C. Fan, X. Wang, X. Duan, B. Wang, and J. Wang, "6G technologies: Key drivers, core requirements, system architectures, and enabling technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 18–27, Sep. 2019.

[8] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[9] *Technical Specification Group Radio Access Network; Study on New Radio (NR) to Support Non-Terrestrial Networks (Release 15)*, document (TS) 38.811, Version 15.2.0, 3GPP, Technical Specification, Sep. 2019.

[10] X. Cao, P. Yang, M. Alzenad, X. Xi, D. Wu, and H. Yanikomeroglu, "Airborne communication networks: A survey," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 1907–1926, Sep. 2018.

[11] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[12] O. Kodheli, "Satellite communications in the new space era: A survey and future challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 70–109, 4th Quart., 2021.

[13] Airbus. *Zephyr; Pionering Stratosphere*. Accessed: May 31, 2021. [Online]. Available: https://www.airbus.com/defence/uav/zephyr.html

[14] Google. *Google LOON Project*. Accessed: May 31, 2021. [Online]. Available: https://loon.com/

[15] *Technical Specification Group Radio Access Network Solutions for NR to Support Non-Terrestrial Networks (NTN) (Release 16)*, document (TS) 38.821, Version 16.0.0, 3GPP, Technical Specification, Dec. 2019.

[16] Y. Hu, M. Chen, and W. Saad, "Joint access and backhaul resource management in satellite-drone networks: A competitive market approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3908–3923, Jun. 2020.

[17] B. Deng, C. Jiang, J. Yan, N. Ge, S. Guo, and S. Zhao, "Joint multigroup precoding and resource allocation in integrated terrestrial-satellite networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8075–8090, Aug. 2019.

[18] R. Mijumbi, J. Serrat, J. L. Gorricho, N. Bouten, F. D. Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 236–262, 1st Quart., 2016.

[19] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, May 2017.

[20] M. Series, *IMT Vision—Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, Recommendation ITU, document 2083, 2015.

[21] *Technical Specification Group Radio Access Network;Study on Scenarios and Requirements for Next Generation Access Technologies;(Release 14)*, document (TS) 38.913, Version 14.3.0, 3GPP, Technical Specification, Jun. 2017.

[22] Y. Hu, M. Chen, and W. Saad, "Competitive market for joint access and backhaul resource allocation in satellite-drone networks," in *Proc. 10th IFIP Int. Conf. New Technol., Mobility Secur. (NTMS)*, Jun. 2019, pp. 1–5.

[23] S. M. Shahid, Y. T. Seyoum, S. H. Won, and S. Kwon, "Load balancing for 5G integrated satellite-terrestrial networks," *IEEE Access*, vol. 8, pp. 132144–132156, 2020.

[24] A. Alsharoa and M.-S. Alouini, "Improvement of the global connectivity using integrated satellite-airborne-terrestrial networks with resource optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5088–5100, Aug. 2020.

[25] C. Qiu, Z. Wei, Z. Feng, and P. Zhang, "Joint resource allocation, placement and user association of multiple UAV-mounted base stations with in-band wireless backhaul," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1575–1578, Dec. 2019.

[26] J. Li, K. Xue, D. S. L. Wei, J. Liu, and Y. Zhang, "Energy efficiency and traffic offloading optimization in integrated satellite/terrestrial radio access networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2367–2381, Apr. 2020.

[27] X. Li, W. Feng, Y. Chen, C.-X. Wang, and N. Ge, "Maritime coverage enhancement using UAVs coordinated with hybrid satellite-terrestrial networks," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2355–2369, Apr. 2020.

[28] Y. Li, N. Deng, and W. Zhou, "A hierarchical approach to resource allocation in extensible multi-layer LEO-MSS," *IEEE Access*, vol. 8, pp. 18522–18537, 2020.

[29] W. Abderrahim, O. Amin, M.-S. Alouini, and B. Shihada, "Latency-aware offloading in integrated satellite terrestrial networks," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 490–500, 2020.

[30] S. He, H. Tian, X. Lyu, G. Nie, and S. Fan, "Distributed cache placement and user association in multicast-aided heterogeneous networks," *IEEE Access*, vol. 5, pp. 25365–25376, 2017.

[31] J. Elhachmi and Z. Guennoun, "Cognitive radio spectrum allocation using genetic algorithm," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, pp. 1–11, Dec. 2016.

[32] D. J. Birabwa, D. Ramotsoela, and N. Ventura, "Slice-aware user association and resource allocation in integrated terrestrial and non-terrestrial networks," in *Proc. Southern Afr. Telecommun. Netw. Appl. Conf. (SATNAC)*, 2021, pp. 44–49.

[33] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1647–1650, Jun. 2016.

[34] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 3949–3963, Jun. 2016.

[35] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-D placement of an aerial base station in next generation cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–5.

[36] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 604–607, Mar. 2017.

[37] A. Mohammed, A. Mehmood, F.-N. Pavlidou, and M. Mohorcic, "The role of high-altitude platforms (HAPs) in the global wireless connectivity," *Proc. IEEE*, vol. 99, no. 11, pp. 1939–1953, Nov. 2011.

[38] O. E. Falowo and H. A. Chan, "Joint call admission control algorithm for fair radio resource allocation in heterogeneous wireless networks supporting heterogeneous mobile terminals," in *Proc. 7th IEEE Consum. Commun. Netw. Conf.*, Jan. 2010, pp. 1–5.

[39] M. Setayesh, S. Bahrami, and V. W. S. Wong, "Joint PRB and power allocation for slicing eMBB and URLLC services in 5G C-RAN," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[40] Space Exploration Holdings. *Application for Fixed Satell. Service by Space Exploration Holdings, LLC*. Accessed: Jul. 11, 2022. [Online]. Available: https://fcc.report/IBFS/SAT-MOD-20181108-00083

[41] B. Di, H. Zhang, L. Song, Y. Li, and G. Y. Li, "Ultra-dense LEO: Integrating terrestrial-satellite networks into 5G and beyond for data offloading," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 47–62, Jan. 2019.

[42] H. Tabassum, M. Salehi, and E. Hossain, "Fundamentals of mobility-aware performance characterization of cellular networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2288–2308, 3rd Quart., 2019.

[43] A. Moubayed, A. Shami, and H. Lutfiyya, "Wireless resource virtualization with device-to-device communication underlaying LTE network," *IEEE Trans. Broadcast.*, vol. 61, no. 4, pp. 734–740, Dec. 2015.

[44] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.

[45] T. L. Marzetta and B. M. Hochwald, "Fast transfer of channel state information in wireless systems," *IEEE Trans. Signal Process.*, vol. 54, no. 4, pp. 1268–1278, Apr. 2006.

[46] Z. Liu, L. Zhang, and Z. Ding, "Exploiting bi-directional channel reciprocity in deep learning for low rate massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 889–892, Jun. 2019.

[47] J. Guo, C.-K. Wen, and S. Jin, "Deep learning-based CSI feedback for beamforming in single-and multi-cell massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1872–1884, Jul. 2020.

[48] Y. Liang, J. Tan, H. Jia, J. Zhang, and L. Zhao, "Realizing intelligent spectrum management for integrated satellite and terrestrial networks," *J. Commun. Inf. Netw.*, vol. 6, no. 1, pp. 32–43, 2021.

[49] M. K. Arti, "Channel estimation and detection in hybrid satellite–terrestrial communication systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5764–5771, Jul. 2016.

[50] H. Chaouech and R. Bouallegue, "Channel estimation and detection for multibeam satellite communications," in *Proc. IEEE Asia Pacific Conf. Circuits Syst.*, Dec. 2010, pp. 366–369.

[51] M. Arti, "Imperfect CSI based AF relaying in hybrid satellite-terrestrial cooperative communication systems," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 1681–1686.

[52] S. O. Oladejo and O. E. Falowo, "Latency-aware dynamic resource allocation scheme for multi-tier 5G network: A network slicing-multitenancy scenario," *IEEE Access*, vol. 8, pp. 74834–74852, 2020.

[53] L. Liang, J. Kim, S. C. Jha, K. Sivanesan, and G. Y. Li, "Spectrum and power allocation for vehicular communications with delayed CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 458–461, Aug. 2017.

[54] J. V. Saraiva, I. M. Braga, Jr., V. F. Monteiro, F. Rafael M. Lima, T. F. Maciel, W. C. Freitas, Jr., and F. Rodrigo P. Cavalcanti, "Deep reinforcement learning for QoS-constrained resource allocation in multiservice networks," 2020, *arXiv:2003.02643*.

[55] D. A. Sousa, V. F. Monteiro, T. F. Maciel, F. R. M. Lima, and F. R. P. Cavalcanti, "Resource management for rate maximization with QoE provisioning in wireless networks," *J. Commun. Inf. Syst.*, vol. 31, no. 1, pp. 290–303, 2016.

[56] F. R. M. Lima, T. F. Maciel, W. C. Freitas, and F. R. P. Cavalcanti, "Resource assignment for rate maximization with QoS guarantees in multiservice wireless systems," *IEEE Trans. Veh. Technol.*, vol. 61, no. 3, pp. 1318–1332, Mar. 2012.

[57] M. Emmerich and A. Deutz, "Multicriteria optimization and decision making," Dept. Leiden Inst. Adv. Comput. Sci., Leiden Univ., Leiden, The Netherlands, Tech. Rep., 2006.

[58] H. Pervaiz, L. Musavian, Q. Ni, and Z. Ding, "Energy and spectrum efficient transmission techniques under QoS constraints toward Green heterogeneous networks," *IEEE Access*, vol. 3, pp. 1655–1671, 2015.

[59] J. Tang, D. K. C. So, E. Alsusa, and K. A. Hamdi, "Resource efficiency: A new paradigm on energy efficiency and spectral efficiency tradeoff," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4656–4669, Aug. 2014.

[60] L. Xu, G. Yu, and Y. Jiang, "Energy-efficient resource allocation in single-cell OFDMA systems: Multi-objective approach," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5848–5858, Oct. 2015.

[61] C. A. C. Coello, G. B. Lamont, and D. A. Van Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*, vol. 5. New York, NY, USA: Springer, 2007.

[62] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, and V. Prasath, "Choosing mutation and crossover ratios for genetic algorithms—A review with a new dynamic approach," *Information*, vol. 10, no. 12, p. 390, 2019.

[63] Gurobi Optimization, LLC. (2022). *Mixed Integer Programming Basics*. [Online]. Available: https://www.gurobi.com/resource/mip-basics/

[64] D. E. Knuth, "Big omicron and big Omega and big theta," *ACM SIGACT News*, vol. 8, no. 2, pp. 18–24, Apr./Jun. 1976.

[65] P. E. Black, Ed., "Ω," in *Dictionary of Algorithms and Data Structures*, Mar. 2018. Accessed: Sep. 2, 2022. [Online]. Available: https://www.nist.gov/dads/HTML/omegaCapital.html

[66] A. Lipowski and D. Lipowska, "Roulette-wheel selection via stochastic acceptance," *Phys. A, Statist. Mech. Appl.*, vol. 391, no. 6, pp. 2193–2196, 2012.

[67] D. S. Johnson, "The NP-completeness column: An ongoing guide," *J. Algorithms*, vol. 6, no. 3, pp. 434–451, Sep. 1985.

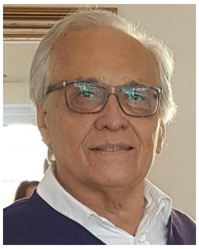
[68] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco, CA, USA: Freeman, 1979.

[69] M. Adedoyin and O. Falowo, "Joint optimization of energy efficiency and spectrum efficiency in 5G ultra-dense networks," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–6.

[70] A. Jain, E. Lopez-Aguilera, and I. Demirkol, "User association and resource allocation in 5G (AURA-5G): A joint optimization framework," *Comput. Netw.*, vol. 192, Jun. 2021, Art. no. 108063.

DENISE JOANITAH BIRABWA (Student Member, IEEE) received the B.Sc. degree in telecommunications engineering from Makerere University, Uganda, in 2011, and the M.Sc. degree in telecommunications engineering from the University of Trento, Italy, in 2014. She is currently pursuing the Ph.D. degree in electrical engineering with the University of Cape Town, South Africa. She is currently an Assistant Lecturer with Kyambogo University, Uganda. Her research interests include resource allocation optimization, heterogeneous networks, non-terrestrial networks, meta-heuristics, and reinforcement learning.

DANIEL RAMOTSOELA (Member, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in computer engineering from the University of Pretoria, in 2013, 2015, and 2020, respectively. He is currently a Senior Lecturer with the Department of Electrical Engineering, University of Cape Town. His research interests include system security, machine learning, and wireless sensor networks, primarily focusing on the IoT applications and cyber-physical systems.



NECO VENTURA (Life Member, IEEE) is currently a Senior Research Scholar, the Head of the Centre for Broadband Networks, and the Director of the Communications Research Group, Department of Electrical Engineering and the Department of Computer Science, University of Cape Town (UCT). He has held positions on several conference organizing committees. He is on the technical program committees of various international conferences. Over the last decade, he has contributed over 100 publications in refereed journals, chapters in books, and refereed international conferences. His research interests include the field of networking, currently centered on next generation mobile networks and architectures, the Internet of Things, machine to machine communications, SDN, NFV fog/edge computing, and 5G. He is a member of the IEEE Computer and the IEEE Communications Societies.