IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 15, 2020, accepted July 10, 2020, date of publication July 17, 2020, date of current version July 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010059

Load Balancing for 5G Integrated Satellite-Terrestrial Networks

SYED MAAZ SHAHID[1], YEMANE TEKLAY SEYOUM[1], SEOK HO WON[2], AND SUNGOH KWON[1], (Senior Member, IEEE)
[1]School of Electrical Engineering, University of Ulsan, Ulsan 44610, South Korea
[2]Future Mobile Communication Division, ETRI, Daejeon 34129, South Korea

Corresponding author: Sungoh Kwon (sungoh@ulsan.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2018R1D1A3B07050215, and in part by Institute of Information and Communications Technology Planning and Evaluation funded by the Korean Government (Ministry of Science and ICT) (No. 2018-0-00175, 5G AgiLe and fLexible integration of SaTellite And cellulaR).

ABSTRACT We propose a load balancing algorithm for a multi-RAT (radio access technology) network including a non-terrestrial network (NTN) and a terrestrial network (TN). Fifth generation (5G) and beyond-5G networks consider NTNs to provide connectivity and data delivery to large numbers of user equipments (UEs). However, previous load balancing algorithms do not consider the coexistence of NTNs and TNs and ignore the different resource allocation units in a multi-RAT network. Hence, we define a radio resource utilization ratio (RRUR) as a common load metric to measure the cell load of each RAT and employ an adaptive threshold to determine overloaded cells. The proposed algorithm consists of two steps to overcome the uneven load distribution across 5G cells: intra-RAT load balancing and inter-RAT load balancing. Based on the RRUR of a cell, the algorithm first performs intra-RAT load balancing by offloading the appropriate edge UEs of an overloaded cell to underutilized neighboring cells. If the RRUR of the cell is still higher than a predefined threshold, then inter-RAT load balancing is performed by offloading the delay-tolerant data flows of UEs to a satellite link. Furthermore, the algorithm estimates the impact of moving loads to the target cell load to avoid unnecessary load balancing actions. Simulation results show that the proposed algorithm not only distributes the load across terrestrial cells more evenly but also increases network throughput and the number of quality of service satisfied UEs more than previous load balancing algorithms.

INDEX TERMS 5G, cellular network, satellite, NTN, radio access network, multi-RAT, QoS, load balancing, data flows, load measurement.

I. INTRODUCTION

Fifth generation (5G) technology is expected to provide high-speed broadband, low-latency services and many devices connected to the Internet at one time. The 5G use cases are classified in terms of requirements for different types of communication. One of the use cases is enhanced mobile broadband (eMBB) which needs to support high bandwidth and high throughput [1], [2]. Furthermore, according to a Cisco forecast, demand for wireless data is expected to reach 77 exabytes per month and online video will make up 82% of internet traffic in 2022 [3], [4]. The amount of bandwidth consumed will grow as more and higher-quality videos are watched. To satisfy the high data rate demand and high bandwidth requirements, there is a need to redesign the cellular network. This leads to the use of non-terrestrial networks (NTNs) in cellular networks. The role of the NTN in 5G networks leads to a heterogeneous global system, and increases the available spectrum and coverage area by providing services in underserved areas [5].

NTNs use spaceborne vehicles, i.e., satellites, to host access nodes, which are already deployed and can be integrated to 5G terrestrial system to support 5G key performers indicators. In the past, terrestrial and satellite networks evolved independently of each other. The 5G paradigm provides a unique opportunity for terrestrial and other radio access technologies (RATs) communities to define a harmonized, full-fledged architecture [6]. Different RATs, including 5G and NTNs, are integrated to guarantee seamless coverage, and to support high data-rate transmissions and data offloading [7]. It is expected that satellite systems will provide radio access networks (RANs), called satellite RANs,

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

with more than 100 high-throughput satellite systems using a geostationary earth orbit (GEO) by 2020-2025 [8]. The integration of terrestrial networks (TNs) with GEO satellite support would be beneficial for global, large-capacity coverage [6]. Moreover, satellites can deliver very high data rates (100 Mbps to 1 Gbps) in broadcast mode to outdoor radio access points [9], and can be used to support the eMBB usage scenarios of 5G [10]. Thus, integration of the satellite into 5G systems will increase the quality of service (QoS) of the user equipments (UEs) by intelligently routing traffic between multiple RAT [11]. Furthermore, this integration provides a larger spectrum to the 5G network and broadband connectivity in rural and remote areas. The 3rd Generation Partnership Project (3GPP) also included NTN in 5G systems to support many services in Release 17 work items [12].

In 5G RAT, the network of cells is densely deployed to provide connectivity to a large number of users. The mobility of UEs causes a load imbalance across the cells in the network [13]. The imbalance in the network affects the QoS of UEs and is an inefficient utilization of available resources. Furthermore, the requirement for high data rates for UEs and the uneven distribution of UEs in the network lead to overutilization of resources in some cells. To overcome these problems, it is necessary to share the load among the cells so that network resources are utilized efficiently to balance the network. For that purpose, intra-RAT load balancing is performed to balance load distributions in order to maintain an appropriate end-user experience and good network performance.

With intra-RAT load balancing, the load from an overloaded cell moves to underloaded neighboring cells. The source and target cells are part of the same RAT. However, sometimes UEs cannot move to neighboring cells due to a scarcity of resources and limited coverage. This affects efficient load balancing among cells, and decreases QoS of the users. The combination of multiple RATs, referred to as a multi-RAT network, is considered for wireless networks to increase resource availability as well as coverage. The multi-RAT network enhances the QoS of UEs, because different RATs can support different services. Furthermore, the UEs access the radio resources of multiple RATs and dynamically route particular traffic to a RAT to satisfy QoS. In the multi-RAT network, it is necessary to determine which RAT should serve which UEs to increase network performance and satisfy QoS of the UEs. Both intra-RAT and inter-RAT load transfers from overloaded cells in the multi-RAT network lead to a well-balanced network and increase network throughput. Moreover, a common load metric is also necessary to measure the load of each RAT for the load balancing in a multi-RAT network. Based on the load metric, radio resources utilization of RATs can be determined and used to divide the network load among the cells of different RATs.

Several research works have studied the problem of mobility load balancing in a cellular network. In [14], the authors resolved the mismatch between the distribution of network resources and traffic demand by handing over UEs of an overloaded cell to a neighboring cell. A utility-based mobility load balancing algorithm in [13] considered operator utility and user utility for the handover process in 5G networks. A load balancing efficiency factor was introduced to consider the load of neighbouring cells and the edge UEs of an overloaded cell. An adaptive algorithm for mobility load balancing in a Long Term Evolution (LTE) small-cell network was proposed in [15]. An adaptive threshold is employed to identify overloaded cells and the UE handovers to candidate target cells from overloaded cells. In [16], a cluster-based mobility load balancing algorithm was proposed for heterogeneous LTE networks. The algorithm dynamically constructs clusters of cells by considering overloaded cells and their neighbors, and performs load balancing in those clusters. Previous work considered a single RAT and performed intra-RAT load handover (i.e., terrestrial-terrestrial) for load balancing. whereas in a multi-RAT network, inter-RAT load balancing in conjunction with intra-RAT load balancing is also performed. In the multi-RAT network, it is necessary to determine suitable RATs for UEs in order to provide the required resources. Furthermore, a common load-measure metric is required in the multi-RAT network to measure the resource utilization of each RAT and to compare the loads of multiple RATs. Therefore, these load balancing algorithms are not applicable in a multi-RAT network for balancing the load of terrestrial cells.

In the literature, multiple RATs were also considered in heterogeneous cellular network for enhancing QoS. In [17], the authors proposed an algorithm for traffic-splitting and aggregation in heterogeneous networks. In the algorithm, the UEs' traffic is split across multiple RATs that constitute terrestrial cells and wireless LANs. In [18], the authors proposed a probabilistic RAT selection approach in 5G heterogeneous networks that included Wi-Fi and cellular networks. The previous work used a multi-RAT network to increase capacity and coverage of the TNs, but did not consider load balancing in terrestrial RAT. Further, previous work did not devise a common metric to measure RAT traffic loads, which is necessary in a multi-RAT network because different RATs use different time frequency resource units. Furthermore, load balancing in TNs using multiple RATs increases convergence as well as satisfying-QoS of the UEs providing resource availability to UEs. Thus, the integration of NTNs and 5G networks would balance the terrestrial cells by increasing spectrum availability and the coverage area.

In this paper, we propose a load balancing algorithm to balance the 5G RAT in a multi-RAT network, with NTNs and 5G networks assumed for the multi-RAT network. For load balancing in terrestrial cells, we consider intra-RAT and inter-RAT offloading of the UEs from the overloaded cells. For that purpose, we introduce the radio resource utilization ratio (RRUR), a common metric to represents the load of each RAT. Based on the RRUR of the cells, the algorithm offloads UEs from overloaded terrestrial cells to neighboring cells, as well as to a satellite cell, considering the data flows

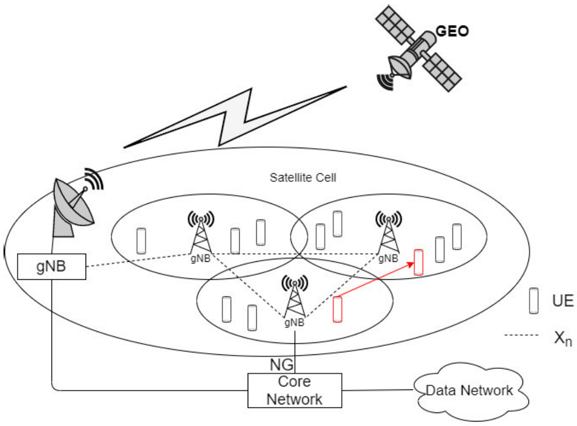


FIGURE 1. Access network architecture.

of the UEs. To offload the UEs, the algorithm estimates the load status of the currently overloaded cells and the candidate target cells and chooses the UEs for offloading in order to effectively distribute the load to avoid candidate target cells that might become overlaoded. An adaptive threshold is used to adopt the network traffic and measure the overload status of a cell. Furthermore, a 5G QoS model is exploited to maintain different queues for delay-sensitive and delay-tolerant data flows. Simulation results show that the proposed algorithm ensures a balanced load among terrestrial cells.

The remainder of this paper is organized as follows. Section II presents the details of the network architecture, load measurement, and the problem formulation. Section III presents the proposed load balancing algorithm aimed at balancing 5G cells. Section IV describes the simulation environment and results, and Section V concludes the paper.

II. SYSTEM MODEL

This section defines the network architecture to be used throughout the paper. Furthermore, the section explains how to measure cell load, and discusses the load balancing problem in 5G multi-RAT network.

A. NETWORK ARCHITECTURE

In this paper, we consider the coexistence of TNs and NTNs, as shown in Figure 1. The TN includes a set of 5G cells, $\mathcal{T}$, with next-generation node B (en-gNB or gNB). An Xn interface is considered for direct communication between the neighboring gNBs. For the NTN, we consider a GEO satellite that is connected to the NTN gNB through a ground station. The GEO satellite is always in the same relative position and therefore, inter-satellite handoff is unnecessary, and there is no Doppler shift. The terrestrial gNB connects with the NTN gNB via Xn to share control information. Management of traffic loads is provided over the Xn interface. For the core network (CN) connection, an NG interface is considered between gNBs and the 5G CN. The multi-connectivity feature for UEs is adopted in which a terrestrial gNB acts as an anchor and the satellite as a slave node. We consider the 3C configuration for the control plane and the 1A configuration for the user plane [19]. The 3C configuration splits the bearer in the anchor, which is the control plane only at the cellular gNBs, whereas the 1A configuration has a separate radio bearer for each of the UEs, and splitting of the user plane occurs in the CN. Figure 2 shows the 1A and 3C configurations for the radio protocol architecture.

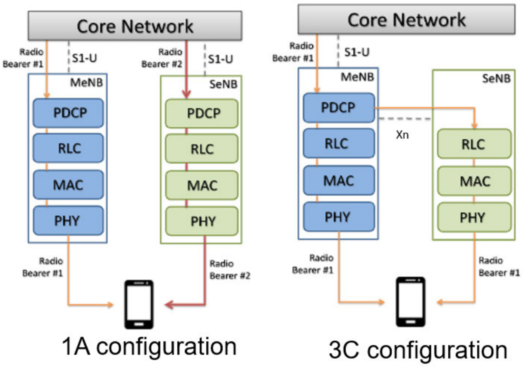


FIGURE 2. Radio protocol architecture for multi-connectivity [23].

TABLE 1. Supported 5G transmission numerology [21].

Subcarrier Spacing (KHz)	Slot Duration (ms)	Max. Bandwidth (MHz)
15	1	50
30	0.5	100
60	0.25	200
120	0.125	400

There are two classes of the UEs' data flow; one class has a delay-tolerant flow, and the other class has a delay-sensitive flow. The packet delay budget (PDB) is defined by 3GPP for data flows in 5G system [20]. The PDB of flows greater than the satellite propagation delay are considered delay-tolerant flows, and flows with a PDB less than the satellite propagation delay are considered delay-sensitive flows. To support multiple data flows, different numerologies are introduced in 5G [21]. Based on the data flows, each UE uses different 5G numerologies, i.e., carrier spacing (CS). Multiple numerologies for 5G New Radio (NR) are shown in Table 1. A physical resource block (PRB) is the smallest unit of a resource block allocated to UEs by a gNB. Each 5G terrestrial cell has some available PRBs based on the system bandwidth and the CS. Furthermore, the PRB bandwidth depends on the CS, and one PRB occupies bandwidth equal to the number of consecutive sub-carriers into the CS. For the NTN, satellite bandwidth is assigned to UEs according to their required data rates using the Shannon capacity formula [22].

B. MEASUREMENT REPORT TRIGGERING

The purpose of the measurement report is to transfer measurement results from the UEs to the network. In 5G, reference signal received power (RSRP) measurements are important for mobility management. A network lets UEs report the signal quality of the current cell, i.e., serving cell, and the target cell. The 3GPP defined several sets of predefined measurement report mechanisms to be executed by UEs and these predefined measurement report types are called event. For 5G, there are six events (A1, A2, A3, A4, A5 and A6) for intra-RAT measurements and two events (B1 and B2) for inter-RAT measurements were specified and discussed in [24]. We consider both intra-RAT and inter-RAT offloading for load balancing among 5G cells. However, events for intra-RAT measurements are used in this work to determine edge UEs and target neighboring cells for intra-RAT offloading. UEs are in the coverage area of a satellite cell, therefore, there is no need to determine edge UEs and target neighboring cells for inter-RAT offloading. Data flows of the UEs and traffic loads of the serving cells are considered for inter-RAT offloading.

Two events (A3 and A4) are considered for intra-RAT load balancing in this paper. Event A3 is the most suitable for finding the best neighboring cells for handover of UEs [15], and A3 is widely used for inter-RAT handovers in wireless networks [25]. Event A3 is triggered when the signal of a neighboring cell is offset better than the serving cell, and UEs report measurements to the serving cell. The following equation shows the trigger condition of the A3 event

$$M_n + Ofn + Ocn - Hys > M_p + Ofp + Ocp + Off$$

where M_n and M_p are the RSRP of the neighboring cell and the current cell, respectively. *Ofn* and *Ofp* are the frequency-specific offsets, and *Ocn* and *Ocp* are the cell individual offsets for the target and serving cells, respectively; *Hys* is the hysteresis parameter; and *Off* is the A3 event offset between the serving cell and the target neighboring cell. The frequency-specific offsets are used for inter-frequency handover, and therefore, we forgo *Ofn* and *Ofp* in this paper. The intra-RAT handover decision changes based on the values of *Ocn*, *Ocp*, and *Off*. Based on the load status of a cell load, the A3 variables (*Ocn*, *Ocp* and *Off*) are changed to intentionally delay or hasten the handovers of UEs.

Consider Figure 3a, where cell 1 is overloaded with five UEs, and neighboring cell 2 has less of a load. There are two edge UEs in cell 1, i.e., UE 1 and UE 2, which can be moved to a neighboring cell to reduce the cell 1 load. Either by decreasing *Ocn* and increasing *Ocp*, the range of cell 1 decreases and UE 1 can be offloaded to the cell 2 gNB to balance the network. For offloading UEs to a particular neighboring cell, only the *Ocn* parameter is adjusted, based on the RRUR of the serving cell. Hence, event A3 will be used to find a suitable target cell for offloading UEs of overloaded cells for intra-RAT load balancing. Moreover, information on the edge UEs of the overloaded cells is also needed prior to handover. For that purpose, event A4 is used to sort the outskirt UEs of the cell. Since event A4 is triggered when the RSRP of neighboring cell M_n becomes better than a provided threshold, *Thresh*. So, event A4 is defined as

$$M_n + Ofn + Ocn - Hys > Thresh \tag{1}$$

Measurement reports by UEs after triggering event A3 are used to determine the threshold for A4 events, as done in [15].

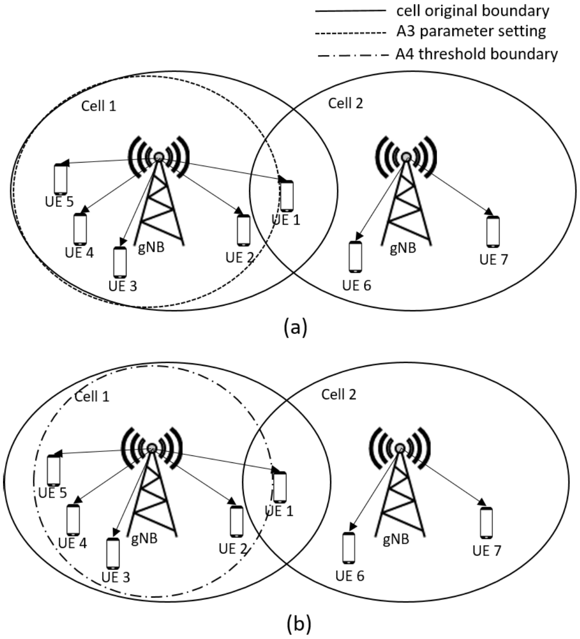


FIGURE 3. Events A3 and A4 for the algorithm: (a) A3 event parameter settings for load balancing, and (b) getting candidate edge UEs and target cell information using A4 event parameters.

UEs that satisfy condition (1) will report the RSRP for the serving cell as well as neighboring cells. For example, in Figure 3b, UE 1 reports measurements to cell 1 because it is outside the A4 event boundary of serving cell 1. Hence, cell 1 reduces the *Ocn* of target cell 2 to offload UE 1 to the target cell. Based on the event A4 boundary, a cell will obtain edge UEs' information and will list candidate UEs, $\boldsymbol{E} = \{e_1, \ldots, e_n\}$ where e_i is the edge UE i for $0 \leq i \leq n$, for intra-RAT load balancing.

C. FLOW CLASSIFICATION IN 5G

To exploit multi-RAT connectivity in 5G networks, it is necessary to steer traffic across the available access networks optimally. A delay incurred by satellite access is orders of magnitude higher than its terrestrial counterpart. That is, in addition to achieving balanced radio resource utilization, we need to guarantee that delay-sensitive traffic is forwarded only through terrestrial access, whereas delay-tolerant traffic can be served through a satellite when the terrestrial network load surpasses a given threshold. To do so, it is necessary to classify data flows into different QoS classes.

In the 5G CN, a session management function (SMF) is introduced for the 5G QoS model [20]. The SMF manages the protocol data unit (PDU) session, which is a logical connection between UEs and the data network (DN), and the related QoS flows in the CN. The SMF assigns a QoS flow identifier (QFI) and a QoS profile to a flow based on information provided by the policy control function. A QFI value corresponds to a particular QoS flow, and each QoS flow is identified by the QFI. The service data flows (SDFs), which are groups of IP flows/packets, are classified based on

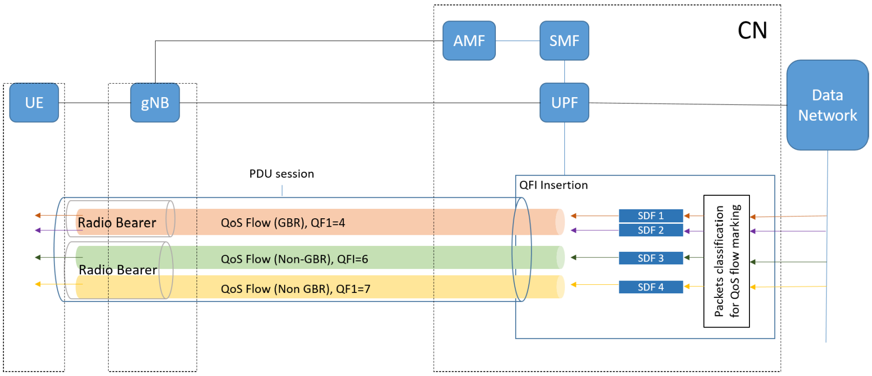


FIGURE 4. **The 5G QoS model.**

IP flows received from a DN. Later, SDFs with the same QoS characteristics are grouped together in the 5G QoS flow and are marked with the same QFI. The SMF provides the user plane function (UPF) with the packet detection rules (PDRs) for mapping SDFs to the QoS flows. Each QoS flow is defined by a QoS profile, and the QoS profile identifies the 5G QoS characteristics with a 5G QoS Identifier (5QI). Based on the 5QI value, 5QI-to-QoS characteristic mapping is provided [20]. Furthermore, the PDB is defined for a QoS flow based on the 5QI value of the flow [20]. For example, the PDB is 150 ms for a 5QI value equal to 2, and the flow is considered delay-sensitive. The QoS flow model based on [26] is shown in Figure 4.

For the multi-RAT network, SDFs with same QoS flow can be directed to a particular RAT, and then, the SMF sends the QoS profile to the gNB via the access and mobility management function (AMF). Our work exploits the SMF service to maintain different queues for delay-sensitive and delay-tolerant flows by offloading flows to different RATs. Based on the QoS flows of the UEs, the data planes of the UEs switch to different RATs using the 5G QoS model.

D. LOAD MEASUREMENT IN 5G MULTI-RATs

Proper load measurement of cells is crucial for optimizing the performance of a network through load balancing. For that purpose, a common load measurement metric is needed to measure the load of each RAT in a multi-RAT network. For LTE networks, PRB allocation information, called the resource block utilization ratio (RBUR), is mainly used to determine overloaded cells. For any given time, T, the average RBUR of a cell n, $\overline{RB}_n$, is expressed in [15] as

$$\overline{RB}_n = \frac{1}{T \cdot N_{PRB}} \sum_{\tau \in (t-T,t)} RB_n \ , \tag{2}$$

where RB_n and N_{PRB} are the number of allocated resource blocks and the total number of resource blocks in the cell, respectively. Similarly, PRB allocation information can also be used to measure the load of 5G RAT. However, the total number of PRBs, N_{PRB}, in 5G changes dynamically with changes in subcarrier spacing [21]. Therefore, the RBUR cannot be directly used to measure the cell load in 5G RAT. Furthermore, radio resources are not allocated in terms of the PRBs in an NTN. Since, we need a common metric/parameter to measure the radio resources utilization of different RATs for a 5G multi-RAT network.

In this paper, we introduce the radio resource usage ratio (RRUR) as a load measurement metric for the multi-RAT network. We defined RRUR as the ratio of bandwidth used by RAT to the total RAT bandwidth. For 5G RAT, the RRUR is calculated based on PRB allocation information and resource block bandwidth. For any given time, T, the RRUR of cell n in 5G RAT is calculated as

$$\beta_n = \frac{1}{T \cdot \omega_n} \sum_{\tau \in (t-T,t)} \gamma_\tau \cdot \varsigma_\tau \ , \tag{3}$$

where ω_n is the total bandwidth of 5G cell n, and γ_τ and ς_τ are the allocated PRBs and resource block bandwidth at time τ, respectively. The resource block bandwidth depends on the numerologies.

In NTN RAT, bandwidth utilization by the satellite determines the satellite load. The RRUR of satellite cell $\mathcal{S}$ is calculated as

$$\beta_{\mathcal{S}} = \frac{1}{T \cdot \omega_{sat}} \sum_{\tau \in (t-T,t)} \Omega_\tau \ , \tag{4}$$

where Ω_τ is the bandwidth allocated to UEs based on the Shannon formula and ω_{sat} is the total bandwidth of the satellite at time τ.

Based on the common load measure metric, i.e., RRUR, load distribution among cells of different RATs is determined. A higher RRUR of a cell indicates that the cell has a higher load to serve and fewer available resources. If RRUR is more than a predefined threshold, the cell is overloaded, and UEs moving to that cell will either be dropped or will experience low data rates. Hence, new UEs in an overloaded cell will reduce the per UE data rates. Therefore, it is necessary to reduce the load of the overloaded cell by switching the data plane of some UEs to a lightly loaded cell or another RAT. Furthermore, the RRUR overcomes the different physical layer channels properties of each RAT in a 5G multi-RAT network. Hence, the physical layer channel of each RAT does not affect the problem formulation of load balancing in 5G integrated satellite-terrestrial networks.

E. PROBLEM FORMULATION

In a network, if the RRUR of a RAT cell is close to 1, a user that moves into the cell will either be dropped or will experience a low data rate. Hence, a new user in an overloaded cell will reduce the per-user data rate, which affects the QoS of the UEs. To reduce the RRUR of a cell, load balancing among cells is necessary. In load balancing, the total network load is shared among the cells. For that purpose, loads from overloaded cells offload to underloaded neighboring cells in the same RAT, referred to as intra-RAT load balancing. Another option is inter-RAT load balancing in which UEs of the overloaded cell move to another RAT to balance the cellular network.

We formulate the problem of load balancing as one of reducing the RRUR of the terrestrial cells to a target RRUR, $\overline{\beta}$, such that the square distance between the cell RRUR and $\overline{\beta}$ is minimized. A multi-RAT network consists of a set of cells, $\mathcal{N}$, in which there is a set of terrestrial cells, $\mathcal{T}$, and a satellite cell, $\mathcal{S}$, i.e., $\mathcal{N} = \mathcal{T} \cup \mathcal{S}$, and $\mathcal{I}$ users. The problem can be expressed as

$$\begin{aligned} &\min \sum_{\forall_n \in \mathcal{T}} |\overline{\beta} - \beta_n|^2 \\ &\text{subject to: } \beta_{\mathcal{S}} \leq Thr_{adp}, \\ &\qquad \beta_\kappa^i \geq \rho_i, \quad \kappa \in \mathcal{N} \end{aligned} \tag{5}$$

where β_n is the RRUR of terrestrial cell n, $\beta_{\mathcal{S}}$ is the RRUR of a satellite cell $\mathcal{S}$, Thr_{adp} is the adaptive threshold, β_κ^i is the resource allocated to user i by cell κ, and ρ_i is the resources required by user i, from which ρ_i is calculated based on the minimum data rate required by UE i. The cell allocates resources to UEs based on the UEs' required data rates and the channel quality.

To estimate $\overline{\beta}$, mean square estimation of β_n can be phrased as was done in [27]. Consider random variable y and the mean square estimation of y by constant c as follows:

$$E[(c-y)^2] = \int_{-\infty}^{\infty} (c-y)^2 f(y) dy$$

The difference, $|c-y|$, is minimum if

$$\frac{de}{dc} = 0$$

Because the difference depends on c, constant c is equal to

$$c = \int_{-\infty}^{\infty} y f(y) dy$$

and $E[y] = \int_{-\infty}^{\infty} y f(y) dy$, and thus

$$c = E[y] \tag{6}$$

Considering (5) and (6), $\overline{\beta}$ is equal to

$$\overline{\beta} = E[\beta_n] \tag{7}$$

Hence, $\overline{\beta}$ is expected RRUR of terrestrial cells.

Algorithm 1 Proposed Load Balancing Algorithm

```
function Load_Balance (void)
info_gather ()
for all $o \in \mathcal{O}$ do
    intRAlb ($\beta_o$, $Thr_{adp}$)
    Determine $\beta_{\mathcal{S}}$ using (4)
    if $\beta_{\mathcal{S}} \leq Thr_{adp}$ and $\beta_o \geq Thr_{adp}$ then
        intERlb ($\beta_o$, $\beta_{\mathcal{S}}$, $Thr_{adp}$)
    end if
end for
```

TABLE 2. Definitions of notations used in the proposed algorithm.

Notations	Definitions
$\overline{\beta}$	Average RRUR of 5G cells
β_n	RRUR of terrestrial cell n
$\beta_{\mathcal{S}}$	RRUR of the satellite cell
$\hat{\beta}_n^i$	Estimated resource utilization of cell n by UE i
ρ_i	Required PRBs by UE i
Ω_i	Allocated bandwidth assigned to UE i by the satellite
Thr_{adpt}	Adaptive threshold to find an overloaded cell
thr_{init}	Initial threshold

III. THE PROPOSED ALGORITHM

The proposed algorithm balances the load in 5G RAT based on data flows of UEs and by considering cell load status in a 5G multi-RAT network. The algorithm runs in each 5G gNB and initiates load balancing when terrestrial cells are overloaded. The proposed algorithm consists of three parts: information gathering, intra-RAT load balancing, and inter-RAT load balancing. For load balancing in 5G cells, the algorithm first gathers information on the load status of the cells using a function call *info_gather*. After that, loads from overloaded 5G cells are released to underloaded cells by calling a function called *intRAlb*. At the end, based on the load status of the cells, the algorithm calls a function called *intERlb* to transfer terrestrial loads to NTN RAT. Each part of the proposed algorithm is described in the subsections below. Algorithm 1 shows the proposed algorithm's process and Table 2 defines the notations used in the algorithm.

A. INFORMATION GATHERING

For gathering the information, the function, info_gather (), measures the load of terrestrial cells, i.e., the RRUR, using (3), and then, the average load of 5G cells is calculated using (7). To estimate the overload status of a cell, adaptive threshold Thr_{adpt} is determined as follows

$$Thr_{adpt} = max(\overline{\beta}, thr_{init}) \tag{8}$$

where thr_{init} is the fixed initial threshold used to determine whether there is a need for load balancing in the network. The adaptive threshold, Thr_{adpt}, is used to adopt the network load. The network load can vary over time because of user

Algorithm 2 Information Gathering

```
function info_gather ()
Get RRUR of terrestrial cells $\mathcal{T}$
Compute average 5G cell load $\overline{\beta}$
Determine $Thr_{adpt}$
Establish a set of overloaded cells, $\mathcal{O} \subsetneq \mathcal{T}$
$Thr_{adpt}$, $\mathcal{O}$
```

mobility and variances in required data rates of the UEs. After that, the algorithm estimates the overload status of a cell by using the following condition

$$\beta_n > Thr_{adpt}, \quad n \in \mathcal{T} \tag{9}$$

and establishes a set, $\mathcal{O}$, of terrestrial cells that satisfy the above condition, where $\mathcal{O} \subsetneq \mathcal{T}$. The process of information gathering is summarized in Algorithm 2.

B. INTRA-RAT LOAD BALANCING

In intra-RAT load balancing, UEs from an overloaded cell, $o \in \mathcal{O}$, move to underloaded neighboring cells. The function gathers information on the edge UEs that are moved from overloaded cell o. For that purpose, the function establishes a set, $\boldsymbol{\nu_o}$, of edge UEs that report measurements to serving cell o based on the A3 event measurement reports. Then, another set of UEs is created, $\boldsymbol{E_o} \subseteq \nu_o$, which report the RSRPs of neighboring cells to the serving cell o during an event A4. The UEs in $\boldsymbol{E_o} = \{e_1, .., e_n\}$ are then sorted in ascending order of serving cell RSRPs and the UEs are arranged according to data flow type. For intra-RAT load balancing, first the UEs of $\boldsymbol{E_o}$ with delay-sensitive flows, and then UEs with delay-tolerant flows, move to underloaded neighboring cells one by one based on the load status of cell o.

Based on event A3, the target neighboring cell is determined in order to offload UE $e_1 \in \boldsymbol{E_o}$. from overloaded cell o. The set $\boldsymbol{\Gamma_{e_1}} = \{\Gamma_1, \Gamma_2, \ldots, \Gamma_m\}$ denotes the neighboring cells reported by UE e_1 to serving cell o under event A4. The neighboring cells are listed in descending order of RSRP values, i.e., the RSRP for Γ_1 is greater than Γ_2. To offload UE e_1, the algorithm estimates $\hat{\beta}^{e_1}_{\Gamma_k}$, the resource utilization of target cell by UE e_1. $\hat{\beta}^{e_1}_{\Gamma_k}$ is calculated based on (3) as follows:

$$\hat{\beta}^{e_1}_{\Gamma_k} = \frac{\rho_{e_1}\varsigma}{\omega_{\Gamma_k}} \tag{10}$$

where ρ_{e_1} is the PRB of cell Γ_k required by UE e_1, ς is the bandwidth of the resource block, and ω_{Γ_k} is the total bandwidth of target cell Γ_k. Before offloading UE e_1 to cell Γ_k, the algorithm checks the following conditions in order to restrict the target cell load to below overload status and to avoid unnecessary offloading of UEs to neighboring cells, i.e., to avoid ping-pongs:

$$\beta_{\Gamma_k} + \hat{\beta}^{e_1}_{\Gamma_k} < Thr_{adpt} \tag{11}$$

$$\beta_o - \hat{\beta}^{e_1}_o > \beta_{\Gamma_k} + \hat{\beta}^{e_1}_{\Gamma_k}. \tag{12}$$

Algorithm 3 Intra-RAT Load Balancing

```
function intRAlb ($\beta_o$, $Thr_{adp}$)
Get candidate edge UEs, $\boldsymbol{E_o}$
Sort $\boldsymbol{E_o}$ in ascending order of RSRP and arrange
    according to data flow type.
for $i \leftarrow 1 : |\boldsymbol{E_o}|$ do
    Determine set $\Gamma_{e_i}$ of target cells for UE $e_i$
    for $k \leftarrow 1 : |\boldsymbol{\Gamma_{e_1}}|$ do
        Estimate $\hat{\beta}^{e_i}_{\Gamma_k}$ using (10)
        if (11) and (12) are satisfied then
            Offload flow of UE $e_i$ to the target cell $\Gamma_k$
            Update RRUR information
            $\beta_o \leftarrow \beta_o - \hat{\beta}^{e_i}_o$
            $\beta_{\Gamma_k} \leftarrow \beta_{\Gamma_k} + \hat{\beta}^{e_i}_{\Gamma_k}$
            Update $\overline{\beta}$ and $Thr_{adpt}$
            break;
        end if
    end for
    if $\beta_o \leq Thr_{adpt}$ then
        break;
    end if
end for
return $\beta_o$, $Thr_{adpt}$
```

If the above conditions are satisfied, UE e_1 moves to target cell Γ_k. After offloading UE e_1, the RRURs of the previous and current serving cells are updated as follows:

$$\beta_o = \beta_o - \hat{\beta}^i_o, \text{ and}$$
$$\beta_{\Gamma_k} = \beta_{\Gamma_k} + \hat{\beta}^{e_1}_{\Gamma_k}.$$

Then, $\overline{\beta}$ and Thr_{adpt} are updated. The same process repeats for each UE in $\boldsymbol{E_o}$ based on the cell loads. Algorithm 3 summarizes the function *intRAlb* (β_o, Thr_{adp}).

C. INTER-RAT LOAD BALANCING

After intra-RAT load balancing, the algorithm again checks the load status of the cell o. If the cell is still overloaded, i.e., $\beta_o > Thr_{adp}$, the algorithm performs inter-RAT load balancing by transferring the load of cell o to satellite cell $\mathcal{S}$ by offloading the delay-tolerant flows of UEs if

$$\beta_{\mathcal{S}} < Thr_{adp} \tag{13}$$

To release the load of 5G cells to a satellite cell, the function generate a set of UEs $\boldsymbol{\mathcal{E}_o} = \{\varepsilon_1, \ldots, \varepsilon_n\}$, where $\boldsymbol{\mathcal{E}_o}$ denotes the UEs of cell o with delay-tolerant data flows. After that, UEs in $\boldsymbol{\mathcal{E}_o}$ are sorted in ascending order of RSRPs from cell o and data flows of UEs in $\boldsymbol{\mathcal{E}_o}$ are offloaded to a satellite link one by one. Before offloading UE ε_1, the function first estimates $\hat{\beta}^{\varepsilon_1}_{\mathcal{S}}$, i.e., the resource utilization of the satellite by UE ε_1. Then, $\hat{\beta}^{\varepsilon_1}_{\mathcal{S}}$ is calculated using the Shannon formula based on the data rate required by the UE:

$$\hat{\beta}^{\varepsilon_1}_{\mathcal{S}} = \frac{\Omega_{\varepsilon_1}}{\omega_{sat}}$$

Algorithm 4 Inter-RAT Load Balancing

```
function intERlb ($\beta_o$, $\beta_{\mathcal{S}}$, $Thr_{adp}$)
Get list of UEs with delay-tolerant flows, $\boldsymbol{\mathcal{E}_o}$
Sort UEs in ascending order of RSRP
for $i \leftarrow 1 : |\boldsymbol{\mathcal{E}_o}|$ do
    if $\beta_{\mathcal{S}} < Thr_{adpt}$ then
        Estimate $\hat{\beta}_{\mathcal{S}}^{\varepsilon_i}$ using (14)
        if (14) is satisfied then
            UPF offloads flow of UE to satellite gNB
            Update RRUR
            $\beta_{\mathcal{S}} \leftarrow \beta_{\mathcal{S}} + \hat{\beta}_{\mathcal{S}}^{\varepsilon_i}$
            $\beta_o \leftarrow \beta_o - \hat{\beta}_o^{\varepsilon_i}$
            Update $\overline{\beta}$ and $Thr_{adp}$
        end if
    else
        break;
    end if
    if $\beta_o \leq Thr_{adp}$ then
        break;
    end if
end for
```

where Ω_{ε_1} is the bandwidth allocated to UE ε_1. Then, the algorithm checks the following condition to offload UE ε_1 to NTN user plane:

$$\beta_{\mathcal{S}} + \hat{\beta}_{\mathcal{S}}^{\varepsilon_1} < Thr_{adpt} \tag{14}$$

The above condition prevents the satellite from being overloaded. For the offloading of data flows, the UPF directs the flow of UE ε_1 to NTN gNB as we considered the separate user plane for each RAT. And the SMF sends QoS policy information based on the 5QI to the NTN gNB through AMF as described II-C. The proposed algorithm offloads the UEs to the satellite cell irrespective of the position of UEs in the cell, since all UEs are within the coverage area of the satellite. After offloading of UE ε_1, the algorithm updates the RRURs of terrestrial cell o and satellite cell $\mathcal{S}$ as follows:

$$\begin{aligned} \beta_{\mathcal{S}} &= \beta_{\mathcal{S}} + \hat{\beta}_{\mathcal{S}}^{\varepsilon_1}, \text{ and} \\ \beta_o &= \beta_o - \hat{\beta}_o^{\varepsilon_1}. \end{aligned} \tag{15}$$

Then the algorithm updates $\overline{\beta}$ and Thr_{adp}. The algorithm again checks the RRURs of the satellite and cell o and repeats the process for each UE of $\boldsymbol{\mathcal{E}_o}$. Algorithm 4 summarizes the function intERlb (β_o, $\beta_{\mathcal{S}}$, Thr_{adp}).

When UEs moves to a satellite, they will experience a long delay. However, offloading UEs with delay-tolerant data flows will not affect the QoS of the UEs, whereas UEs with delay-sensitive data are served by the 5G RAT. Similar to NTNs, the proposed algorithm can be extended to other RATs, i.e., unmanned aerial vehicle (UAV) communication systems [28]. Based on the RRUR, the load status of a RAT can be determined and UEs from an overloaded cell move to the RAT, taking into account the minimum QoS requirements of the users.

We analyzed the computational complexity of the proposed algorithm using big O notation.[1] For load balancing of a terrestrial cell network, the considered number of cells in a multi-RAT network under the proposed algorithm is $|\mathcal{N}|$, which represents the number of cells in set $\mathcal{N}$. Set $\mathcal{N}$ consists of $|\mathcal{T}|$ terrestrial cells and a satellite cell $\mathcal{S}$. Therefore, the maximum numbers of cells to be considered for intra-RAT and inter-RAT load balancing are $|\mathcal{T}|$ and $|\mathcal{N}|$, respectively. Similarly, the maximum numbers of target cells in intra-RAT and inter-RAT offloading are limited by the $|\mathcal{T}|$ terrestrial cells and satellite cell $\mathcal{S}$, respectively. In addition to the number of cells for load balancing, the algorithm also considers UEs in the network, and the number of considered UEs under the algorithm is $\mathcal{I}$.

Since there are, at most, $|\mathcal{T}|$ serving and target cell pairs and $\mathcal{I}$ UEs involved in intra-RAT offloading, the loop in the intra-RAT offloading function of Algorithm 3 should take $O(|\mathcal{T}|) + O(\mathcal{I})$. In the case of inter-RAT offloading, there are, at most, $\mathcal{I}$ UEs, and only one pairing of a terrestrial serving cell and a target satellite cell involved. Hence, the loop in the inter-RAT offloading function of Algorithm 4 should take $O(\mathcal{I})$. Furthermore, the number of overloaded cells is bounded by the number of terrestrial cells, $|\mathcal{T}|$. So, the overall computational complexity of the proposed load balancing algorithm becomes $O(|\mathcal{T}|^2) + O(\mathcal{I}|\mathcal{T}|)$. Generally, $\mathcal{I} \gg |\mathcal{T}|$, so we can say that the computational complexity for the proposed load balancing algorithm is $O(\mathcal{I}|\mathcal{T}|)$.

IV. PERFORMANCE EVALUATION

A. SIMULATION ENVIRONMENTS

We considered a 5G multi-RAT network including a satellite RAT and a 5G RAT. In the satellite RAT, a GEO satellite was connected to an NTN gNB through a ground station. The gNB was connected with a 5G CN that provided access to the public data network. There were seven 5G small cells deployed in a hexagonal pattern. A single satellite cell covered the whole terrestrial network. We considered 110 UEs in the network, and the required data rates for each UE were 5 Mbps to 15 Mbps. Regarding the UEs' distribution over the network area, UEs were randomly distributed among the cells. Half of them were static, and half were in random motion.

In the network, 70% of the UEs had delay-tolerant traffic, while the remaining UEs had delay-sensitive traffic. The UEs with delay-tolerant data flows had carrier spacing of 15KHz, and UEs with delay-sensitive data flows had either 15 KHz or 30 KHz carrier spacing. Transmission power was set to 46 dBm for 5G cells, and the bandwidth was 20 MHz. For the satellite, the C band was used for communications, and bands of frequencies from 3.7 to 4.2 GHz were used for downlink. The satellite had a channel bandwidth of 500 MHz and 12 transponders. Each transponder had a bandwidth of 36 MHz and a guard band of 4 MHz between

[1] Big O is a notation for asymptotic behavior of functions. Suppose f and g are real valued functions; therefore, $f(x) = O(g(x))$ if and only if there exists a positive integer, N, and a positive constant, c, such that $|f(x)| \leq c|g(x)|$, $\forall_x > N$.

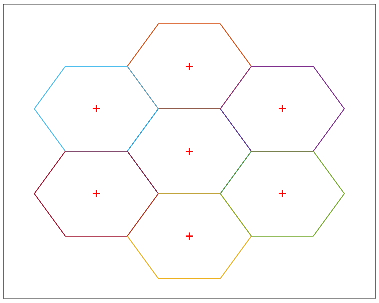


FIGURE 5. Uniformly deployed 5G cell network.

TABLE 3. Simulation parameters.

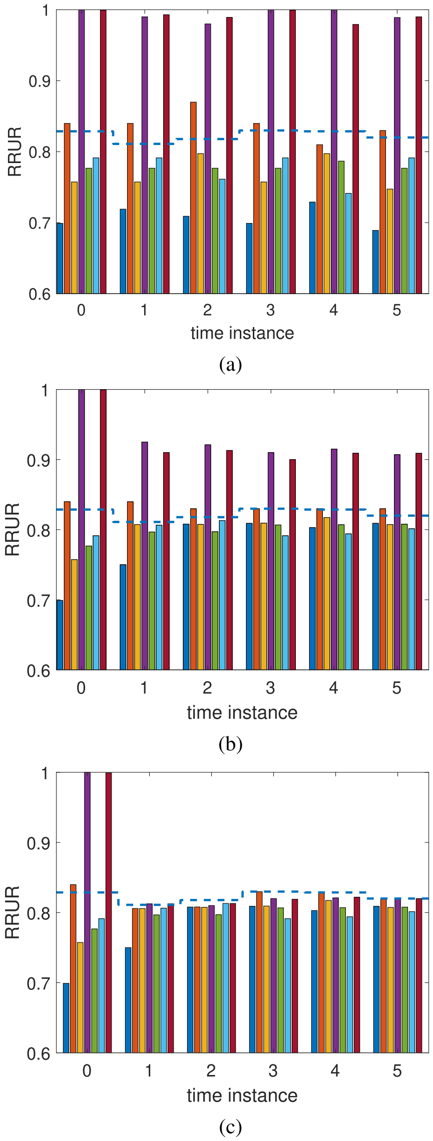
Parameters	Values
Number of terrestrial cells	7
Tx power of terrestrial RAT	46 dBm
Terrestrial RAT bandwidth	20 MHz
Terrestrial path loss	$PL = 147.4 + 43.3log_{10}(d)$
Satellite bandwidth	500 MHz (C band)
Satellite altitude	35780 Km
Number of transponders	12
Number of UEs	110
UEs data rates	5-15 Mbps
Initial threshold	75%

adjacent transponders to avoid interference. The simulation parameters are summarized in Table 3.

For the performance evaluation, we investigated the effect of the proposed algorithm on load distribution across the network and on network throughput. RRUR, which is defined in equation (3), was used to check load distribution among the cells. To validate the performance of the proposed algorithm, which is based on intra-RAT and inter-RAT load balancing, we compared it with an adaptive mobility load balancing algorithm [15]. Further scenarios with various numbers of UEs and cell bandwidths were simulated to show the effectiveness of the proposed algorithm. For the sake of simplicity, we denote the proposed mobility load balancing (MLB) algorithm as adaptive multi-RAT MLB, the adaptive mobility load balancing algorithm as adaptive intra-RAT MLB, and simulations without an MLB algorithm are denoted no MLB.

B. IMPACT OF THE PROPOSED ALGORITHM ON LOAD DISTRIBUTION

The algorithm's impact on load distribution across the network cells in terms of RRUR was compared with adaptive intra-RAT MLB and no MLB algorithms. The scenario with the initial setting was simulated without the MLB algorithm as well as with the MLB algorithms, and the RRUR of the terrestrial cells are shown in Figure 6. Each time instance shows the RRUR of seven 5G cells. Figure 6a shows the RRUR of the cells when no MLB was considered, and some terrestrial cell loads were more than the threshold, showing the cells were overloaded. The blue dotted line in each plot of Figure 6 shows the adaptive threshold, which changed with the network load. As we can see in Figure 6a, some cells had an RRUR greater than the threshold, i.e.,0.82, and some cells were underloaded, with an RRUR of less than 0.7. Consider time instance 2, cell 4 shows a maximum RRUR of 0.99, whereas cell 1 shows a minimum RRUR of 0.71, and the gap is 0.28. The RRURs of the cells with the adaptive intra-RAT MLB are shown in Figure 6b. As we can see in the figure, load from the overloaded cell moves to the underutilized cell to balance the network, and the gap between the maximum RRUR and the minimum RRUR was reduced to 0.10 in time instance 5. Although the adaptive intra-RAT MLB algorithm reduced the RRUR of the overloaded cells, cells had an RRUR greater than the threshold.



(a)

(b)

(c)

FIGURE 6. RRUR of terrestrial cells in the network (a) without the MLB algorithm (b) with the adaptive intra-RAT MLB algorithm, and (c) with the adaptive multi-RAT MLB algorithm.

The RRURs of 5G cells were reduced to defined threshold under the adaptive multi-RAT MLB, as shown in Figure 6c. With the adaptive multi-RAT MLB, first the load from overloaded cells was released to underloaded neighboring cells, which increased the resource utilization of the underloaded

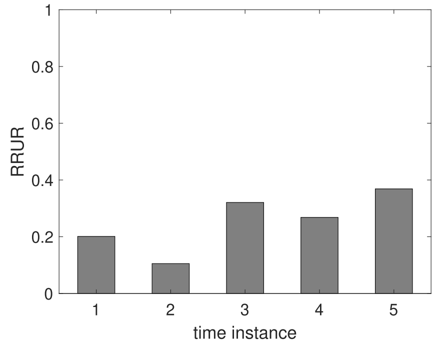


FIGURE 7. RRUR of the satellite.

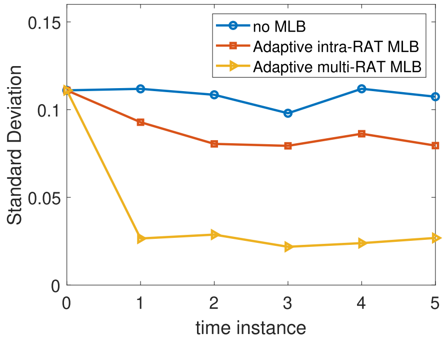


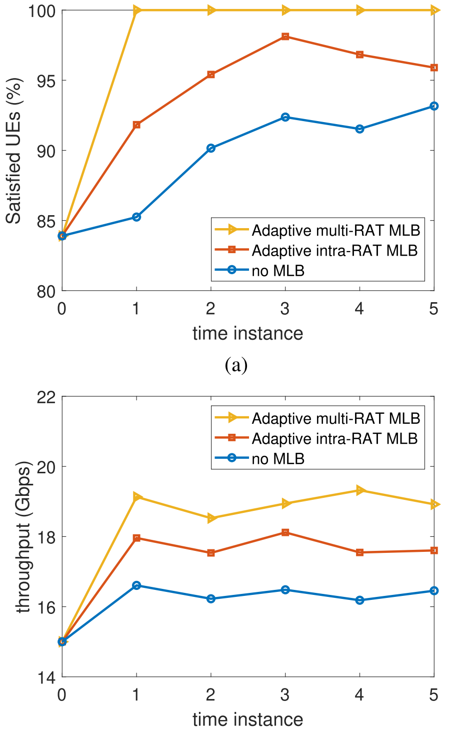
FIGURE 8. Standard deviation of RRUR among the cells of the 5G RAT.

cells and decreased the load on highly utilized cells. After that, the excess load from the overloaded cell, i.e., center UEs with a delay-tolerant data flow, moved to the satellite cell, which further reduced the load of the overloaded cells to the defined threshold. This eventually reduced the gap between the maximum RRUR and minimum RRUR until it reached 0.019. The RRUR of each terrestrial cell decreased to the threshold and the terrestrial cells network was evenly balanced under the adaptive multi-RAT MLB, as shown in Figure 6c. The satellite serves the UEs with delay-tolerant flows by keeping the RRUR at less than the threshold, which is shown in Figure 7. Considering the load status of the satellite, new users can easily be accommodated in the network and the satellite can assign more resources to satisfy the QoS of the users.

Figure 8 shows the standard deviation of 5G cell loads with and without load balancing algorithms. The standard deviation of the RRUR under the adaptive multi-RAT MLB algorithm is close to zero, and less than the adaptive intra-RAT MLB due the fact that the data flows of the center UEs in the overloaded cell can be offloaded to the satellite. Hence, the adaptive multi-RAT MLB performs load balancing considering 5G RAT and NTN RAT resources together and effectively released the load to balance the terrestrial network. Furthermore, the proposed algorithm considers the limitations of adaptive MLB as well as QoS of the UEs.

C. IMPACT OF THE MLB ALGORITHM ON NETWORK THROUGHPUT AND QoS

The network performance in terms of average throughput and QoS of the UEs is shown in Figure 9. Without MLB, the resources of some cells were underutilized, whereas the UEs in overloaded cells could not have the required resources due to the scarcity of available resources. Therefore, the network had minimum throughput and fewer UEs getting the required resources, compared to load balancing algorithms. The adaptive intra-RAT MLB increased both the number of satisfied UEs and network throughput, but it was still less than the adaptive multi-RAT MLB, as shown in Figure 9. Considering the intra-RAT and inter-RAT offloading of UEs, the adaptive multi-RAT MLB allocated enough resources to all the UEs. More resources were available to UEs from multiple RATs that fulfilled the UEs' required data rates. The offloading of UEs from the overloaded cell to the neighboring cells, as well as to the satellite cell decreased the cell load and released more resources of the cells. This allowed the cells to allocate more resources to satisfy QoS of the UEs, and offloaded UEs got their required resources from the underloaded cells of different RATs, which satisfied the QoS of all UEs in the network, as shown in Figure 9a. These factors eventually led to an increase in overall network throughput, as shown in Figure 9b. Thus, from Figures 6, 8 and 9, we can say that the adaptive multi-RAT MLB not only increased network capacity but also satisfied the QoS of the UEs Furthermore, the adaptive multi-RAT MLB balanced the terrestrial cells efficiently by keeping the RRURs of the cells of each RAT to less than the defined threshold.



(a)

(b)

FIGURE 9. (a) The number of satisfied UEs in the network (b) Average throughput of the network.

D. IMPACT OF VARIOUS NUMBERS OF USERS

We studied the impact of various numbers of UEs in the network on the different approaches to load balancing.

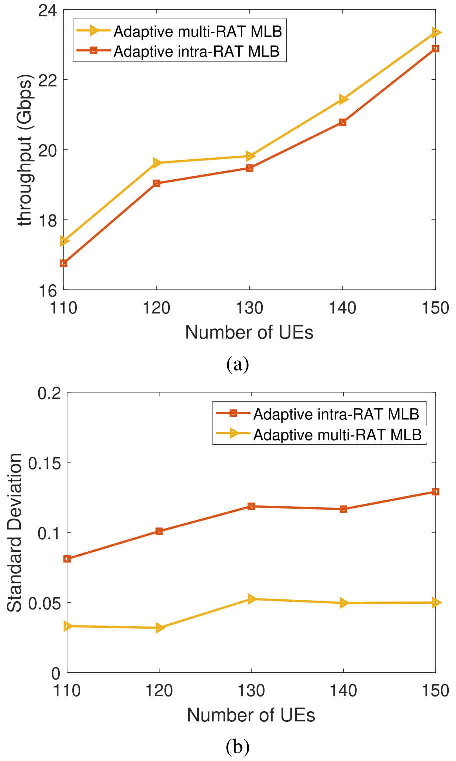


FIGURE 10. **Performance of the load balancing algorithms with different numbers of UEs: (a) average throughput of the network, and (b) standard deviation of RRUR among the cells of 5G RAT.**

The network throughput and the standard deviation of the RRURs among terrestrial cells were observed. Network throughput increased under both of the MLB algorithms by increasing the number of UEs, as shown in Figure 10a. The adaptive multi-RAT MLB had more throughput as the resources of multiple RATs were efficiently utilized to satisfy the QoS flows of the UEs. However, the standard deviation in RRURs among terrestrial cells increased with the increasing numbers of UEs, as shown in Figure 10b. The standard deviation of the RRUR increased by a very small amount under the adaptive multi-RAT MLB, and by less than the adaptive intra-RAT MLB. The gap between maximum RRUR and minimum RRUR increased more under the adaptive intra-RAT MLB, compared to the adaptive multi-RAT MLB with the increasing numbers of UEs. The adaptive multi-RAT MLB with intra-RAT and inter-RAT offloading transferred loads that cannot move to terrestrial neighboring cells to the satellite cell. The RRUR of the satellite is shown in Figure 11, and the utilized resources of the satellite were less than half of the available resources with large number of UEs in the network. Thus, the proposed algorithm keeps the network balanced with a large number of UEs, keeping the RRUR of the satellite minimal.

E. IMPACT OF DIFFERENT CHANNEL BANDWIDTH

We changed the terrestrial cell bandwidth to observe the impact on the load balancing algorithms. The standard deviation of RRURs among terrestrial cells with different 5G cell bandwidth is shown in Figure 12. The standard deviation keeps decreasing when increasing the channel bandwidth in the adaptive intra-RAT MLB above the 30MHz bandwidth, and came close to matching the adaptive multi-RAT MLB. The available resources were increasing in the 5G RAT when increasing the channel bandwidth, which reduced the gap between maximum RRUR and minimum RRUR. The network throughput increased with increasing bandwidths under both MLB algorithms. Network throughput under the adaptive intra-RAT MLB increases more rapidly, compared to the adaptive multi-RAT MLB, as shown in Figure 13. However, the adaptive multi-RAT MLB had more throughput because racecourses of multiple RATs were available to more UEs at the same time. Hence, the proposed algorithm was able to achieve more even load balancing, and increased the capacity of the network at the same time.

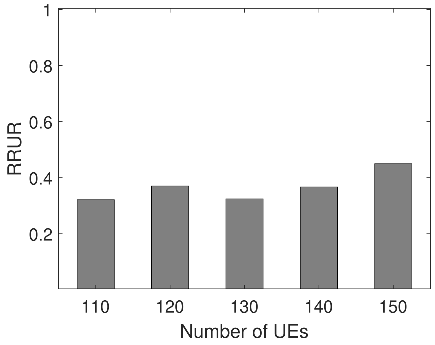


FIGURE 11. **RRUR of the satellite cell.**

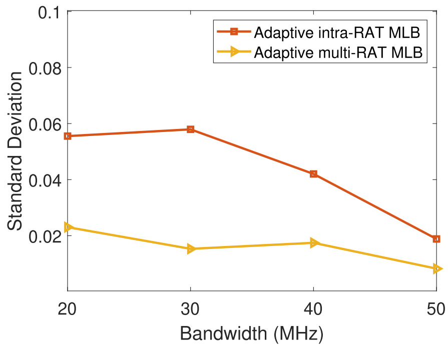


FIGURE 12. **Standard deviation of RRURs among the terrestrial cells with varied 5G cell bandwidths.**

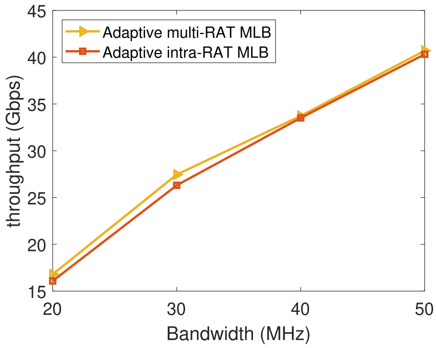
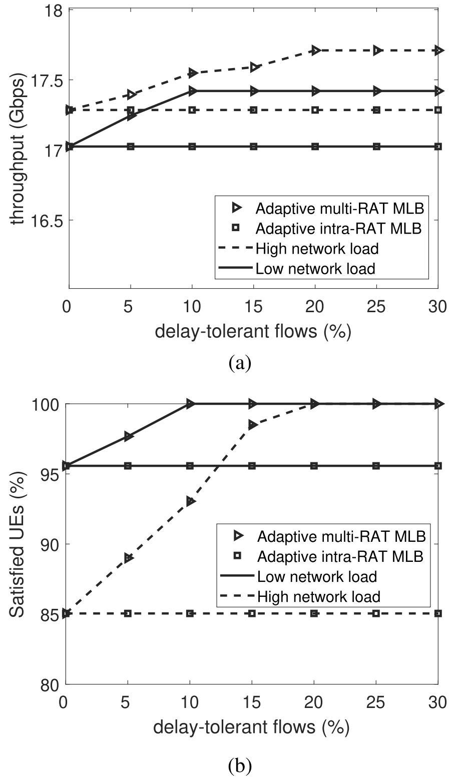


FIGURE 13. **Average throughput of the network with varied terrestrial RAT bandwidths.**



(a)

(b)

FIGURE 14. (a) Average throughput of the network, and (b) the number of satisfied UEs in the network with different delay-tolerant and delay-sensitive traffic ratios.

F. IMPACT OF DELAY-TOLERANT FLOWS WITH DIFFERENT NETWORK LOAD

We varied the delay-tolerant flow ratio in the network to observe the impact on the proposed algorithm. For a given total number of UEs, the percentage of UEs with delay-tolerant flows was changed from 0 to 30 for different network loads. For the different network load, we changed the required data rate of each UE. The required data rates for each UE were 5-10 Mbps and 10-15 Mbps for low and high network load, respectively. Figures 14a and 14b show the network throughput and the number of satisfied UEs, respectively, for different delay-tolerant flow ratios under different network loads. The adaptive multi-RAT MLB has better performance than the adaptive intra-RAT MLB when there are UEs with delay-tolerant flows in the network. When there is no delay-tolerant traffic, i.e., all UEs have delay-sensitive flows, the adaptive multi-RAT MLB only performs intra-RAT offloading. So, the performance of the adaptive multi-RAT MLB returns to the adaptive intra-RAT MLB when there is no UE with a delay-tolerant flow for inter-RAT offloading.

The performance of the adaptive intra-RAT MLB remains constant for different delay-tolerant and delay-sensitive ratios, as shown in Figure 14. The reason is that the adaptive intra-RAT MLB performs terrestrial to terrestrial offloading of the UEs irrespective of the data flow type to balance cell loads, whereas, the performance of the adaptive multi-RAT MLB increases with increases in delay-tolerant traffic. By increasing delay-tolerant traffic, the adaptive multi-RAT MLB finds more UEs with delay-tolerant flows, and offloads the UEs from overloaded cells to a satellite to balance the network. As a result, more UEs get the required resources from multiple RATs, and the network throughput and percentage of satisfied UEs increases. After a required minimum amount of delay-tolerant flows, the network throughput and number of satisfied UEs become constant under the adaptive multi-RAT MLB under different network load conditions. When the network load is high, the adaptive multi-RAT MLB requires a higher ratio of delay-tolerant flows to balance the terrestrial cells. Hence, we can say that the adaptive multi-RAT MLB depends on the availability of delay-tolerant flows for inter-RAT offloading to achieve better performance.

V. CONCLUSION

In this paper, we proposed a load balancing algorithm for a multi-RAT network that consisted of an NTN and a TN. The uneven distribution of the UEs in cells of the 5G network led to imbalanced load distribution across the cells and degraded network performance such as throughput and QoS of UEs. A multi-RAT network uses different time frequency resource units for resource allocation, and therefore, to develop a load balancing algorithm, we the defined RRUR as a common load measurement metric, and employed an adaptive threshold to determine the overload status of the cell based on the network load. To avoid unnecessary offloading of UEs, the proposed algorithm estimates the impact of moving loads on the RRUR of the target cells. Based on intra-RAT and inter-RAT offloading, the load across terrestrial cells became more balanced and the number of satisfied UEs increased in the network. UEs of an overloaded cell that cannot move to neighboring cells are offloaded to a satellite cell, and the cell load is reduced to the defined threshold. Simulation results showed that the proposed algorithm balances terrestrial cell networks and increases the throughput as well as QoS of the UEs better than previous load balancing algorithm. Furthermore, the proposed algorithm assigns enough resources to all UEs from multiple RATs, and 100% of the UEs get their required data rate. The proposed algorithm depends on the availability of delay-tolerant flows to achieve better performance.

REFERENCES

[1] T. Doukoglou, V. Gezerlis, K. Trichias, N. Kostopoulos, N. Vrakas, M. Bougioukos, and R. Legouable, "Vertical industries requirements analysis & targeted KPIs for advanced 5G trials," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2019, pp. 95–100.

[2] *Study on Scenarios and Requirements for Next Generation Access Technologies*, document TS 38.913, Version 15.0.0, 3GPP, Jun. 2018. [Online]. Available: https://portal.3gpp.org/

[3] T. Barnett, S. Jain, U. Andra, and T. Khurana, "Cisco visual networking index (VNI), complete forecast update, 2017–2022," Amer./EMEAR Cisco Knowl. Netw. (CKN) Presentation, Dec. 2018.

[4] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022*, Cisco, San Jose, CA, USA, 2019.

[5] K. Liolis, A. Geurtz, R. Sperber, D. Schulz, S. Watts, G. Poziopoulou, B. Evans, N. Wang, O. Vidal, B. T. Jou, M. Fitch, S. S. Diaz, P. S.Khodashenas, and N. Chuberre, "Satellite use cases and scenarios for 5G eMBB," in *Satellite Communications in the 5G Era*. Edison, NJ, USA: IET, 2018, pp. 25–60.

[6] A. Guidotti, A. Vanelli-Coralli, M. Conti, S. Andrenacci, S. Chatzinotas, N. Maturo, B. Evans, A. Awoseyila, A. Ugolini, T. Foggi, L. Gaudio, N. Alagha, and S. Cioni, "Architectures and key technical challenges for 5G systems incorporating satellites," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2624–2639, Mar. 2019.

[7] V. W. Wong, *Key Technologies for 5G Wireless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[8] G. Giambene, S. Kota, and P. Pillai, "Satellite-5G integration: A network perspective," *IEEE Netw.*, vol. 32, no. 5, pp. 25–31, Sep. 2018.

[9] K. Liolis, A. Geurtz, R. Sperber, D. Schulz, S. Watts, G. Poziopoulou, B. Evans, N. Wang, O. Vidal, B. T. Jou, M. Fitch, S. D. Sendra, P. S. Khodashenas, and N. Chuberre, "Use cases and scenarios of 5G integrated satellite-terrestrial networks for enhanced mobile broadband: The SaT5G approach," *Int. J. Satell. Commun. Netw.*, vol. 37, no. 2, pp. 91–112, Mar. 2019.

[10] E. Zeydan and Y. Turk, "On the impact of satellite communications over mobile networks: An experimental analysis," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11146–11157, Nov. 2019.

[11] B. Evans, O. Onireti, T. Spathopoulos, and M. A. Imran, "The role of satellites in 5G," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 2756–2760.

[12] *Solutions for NR to Support Non-Terrestrial Networks (NTN)*, document WI 860046, Rel-17, 3GPP, Dec. 2019.

[13] K. M. Addali, S. Y. B. Melhem, Y. Khamayseh, Z. Zhang, and M. Kadoch, "Dynamic mobility load balancing for 5G small-cell networks based on utility functions," *IEEE Access*, vol. 7, pp. 126998–127011, 2019.

[14] J. Park, Y. Kim, and J.-R. Lee, "Mobility load balancing method for self-organizing wireless networks inspired by synchronization and matching with preferences," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2594–2606, Mar. 2018.

[15] M. M. Hasan, S. Kwon, and J.-H. Na, "Adaptive mobility load balancing algorithm for LTE small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2205–2217, Apr. 2018.

[16] M. M. Hasan and S. Kwon, "Cluster-based load balancing algorithm for ultra-dense heterogeneous networks," *IEEE Access*, vol. 8, pp. 2153–2162, 2020.

[17] S. Singh, S.-P. Yeh, N. Himayat, and S. Talwar, "Optimal traffic aggregation in multi-RAT heterogeneous wireless networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, May 2016, pp. 626–631.

[18] B. Soleymani, A. Zamani, S. H. Rastegar, and V. Shah-Mansouri, "RAT selection based on association probability in 5G heterogeneous networks," in *Proc. IEEE Symp. Commun. Veh. Technol. (SCVT)*, Nov. 2017, pp. 1–6.

[19] C. Rosa, K. Pedersen, H. Wang, P.-H. Michaelsen, S. Barbera, E. Malkamaki, T. Henttonen, and B. Sebire, "Dual connectivity for LTE small cell evolution: Functionality and performance aspects," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 137–143, Jun. 2016.

[20] *System Architecture for the 5G System; Stage 2*, document TS 23.501, Version 16.3.0, 3GPP, Dec. 2019. [Online]. Available: https://portal.3gpp.org

[21] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. New York, NY, USA: Academic, 2018.

[22] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

[23] R. Antonioli, G. Parente, C. Silva, D. Sousa, E. Rodrigues, T. Maciel, and F. Cavalcanti, "Dual connectivity for LTE-NR cellular networks: Challenges and open issues," *J. Commun. Inf. Syst.*, vol. 33, no. 1, pp. 282–294, 2018.

[24] *Radio Resource Control (RRC) Protocol Specification*, document TS 38.331, Version 15.8.0, 3GPP, Dec. 2019. [Online]. Available: https://portal.3gpp.org/

[25] R. Kwan, R. Arnott, R. Paterson, R. Trivisonno, and M. Kubota, "On mobility load balancing for LTE systems," in *Proc. IEEE 72nd Veh. Technol. Conf. (Fall)*, Sep. 2010, pp. 1–5.

[26] *NR; NR and NG-RAN Overall Description*, document TS 38.300, Version 16.0.0, 3GPP, Dec. 2019. [Online]. Available: https://portal.3gpp.org

[27] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 2002.

[28] Y. Sun, D. Xu, D. W. K. Ng, L. Dai, and R. Schober, "Optimal 3D-trajectory design and resource allocation for solar-powered UAV communication systems," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4281–4298, Jun. 2019.

SYED MAAZ SHAHID received the B.E. degree in electrical engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2015. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Ulsan, South Korea. His research interests include cellular networks and applications of machine learning in signal processing and wireless networks.

YEMANE TEKLAY SEYOUM received the B.Sc. degree in computer science from Bahir Dar University, Ethiopia, in 2008. He is currently pursuing the integrated master's and Ph.D. degrees in electrical engineering with the University of Ulsan, South Korea. His research interests include cellular networks, edge computing, and artificial intelligence for 5G and beyond networks.

SEOK HO WON received the B.S. degree in clinical pathology and electrical engineering from Kwangwoon University, in 1985 and 1990, respectively, and the Ph.D. degree in electrical engineering from Chungnam National University, South Korea, in 2002. Since 1985, he has been a Medical Technician with the Sin-chon General Hospital, South Korea. Since 1990, he has been a Principal Engineer with ETRI, South Korea. He was a Research Faculty Member with Virginia Tech, USA, in 2005, where his duty was developing cognitive radios. His research interests include physical, MAC, and application layers of LTE based on 5G new radio with an emphasis on machine learning.

SUNGOH KWON (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from KAIST, Daejeon, South Korea, in 1994 and 1996, respectively, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2007. From 1996 to 2001, he was a Research Staff Member with Shinsegi Telecomm Inc., Seoul, South Korea. From 2007 to 2010, he was a Principal Engineer with Samsung Electronics Company Ltd., South Korea, where he developed LTE schedulers. Since 2010, he has been with the School of Electrical Engineering, University of Ulsan, South Korea, as an Assistant Professor, where he is currently a Professor. His research interest includes wireless communication networks.