Unicast Multicast Convergence in Cellular Networks : Supplementing FeMBMS with SDN and Dual Connectivity

Indu Yadav*, Pranav Jha*, and Abhay Karandikar *†
*Department of Electrical Engineering, Indian Institute Technology Bombay, India 400076
Email: {indu, pranavjha, karandi}@ee.iitb.ac.in
†Director, Indian Institute Technology Kanpur, India 208016
Email: karandi@iitk.ac.in

Abstract—**Handling the ever-growing data traffic in mobile networks is a tremendously challenging problem, primarily due to the limited availability of radio resources. Supplementing unicast transmissions with the multicast transmission can significantly improve radio resource utilization in mobile networks and help address this problem. Usage of multicast transmission may especially be useful for live streaming and on-demand video services, which constitute a significant % of the burgeoning data traffic. In this paper, we propose a novel SDN based architecture for unified control and management of unicast and multicast transmissions in the Fourth Generation (4G), Fifth Generation (5G) and beyond mobile networks. It brings significant flexibility to the selection of transmission mechanism (unicast or multicast) for individual users (UEs). We also propose an integrated scheme for radio resource allocation for unicast and multicast transmissions. The proposed scheme enables efficient utilization of radio resources in the network while ensuring the required Quality of Service (QoS) to users.**

I. INTRODUCTION

Explosive growth of multimedia content over cellular networks has been observed in recent years. According to [1], on-demand video constitutes 80% of the data traffic generated in the mobile networks today. A large fraction of such content includes streaming traffic, such as live sports events, video conferencing, news, concerts. An efficient and dynamic radio resource allocation scheme is required to address the problem of exponentially increasing mobile data traffic. Multicast/Broadcast transmission enables a group of User Equipments (UEs) access the same multimedia content over shared radio resources and is a promising solution to the inefficient resource utilization problem in cellular networks. Third Generation Partnership Project (3GPP) has also standardized broadcast & multicast transmission. One of the early attempts in this direction was "Multimedia Broadcast and Multicast Services (MBMS)", introduced in 3GPP Release 6. MBMS evolved into "enhanced MBMS (eMBMS)" as part of the Fourth Generation Long Term Evolution (4G LTE) standards and to "Further Enhanced MBMS (FeMBMS) or Enhanced Television Services over eMBMS (EnTV)" recently under LTE Advanced and the Fifth generation (5G) standards [2] [3]. Though FeMBMS is not part of 3GPP 5G Release 15, some of the enhancements, such as support for dedicated eMBMS carriers (cells), MBMS offload, and larger inter-site distances are likely to play important roles in the future evolution of multicasting/broadcasting services in 5G and beyond networks.

FeMBMS defines three types of cells in wireless networks, i.e., dedicated eMBMS cells (carrier providing only Multicast/Broadcast service), unicast only cells, and mixed-mode cells. In "MBMS Offload", also called "MBMS operation on Demand" (MOOD) [4], content being delivered as unicast service to UE may be converted into a broadcast/multicast[1] service in order to conserve network resources when the demand for the content increases.

In addition to MBMS features, 3GPP LTE Advanced and 5G standards have also introduced several other enhancements in recent years. Dual Connectivity is one of the key features introduced in the standards. Dual Connectivity allows a UE to concurrently connect to two Base Stations (BSs) to receive the services. These BSs may belong to the same Radio Access Technology (RAT) or two different RATs, such as LTE and 5G.

FeMBMS enhancements, along with dual connectivity offer an opportunity for improved resource utilization in networks by dynamically switching data flows from one transmission mode to another. The transmission mode switching may be done based on different factors, e.g., demand for specific content, radio condition experienced by UEs across multicast and unicast carriers (cell), load on different network nodes, etc. For example, in a dense heterogeneous network (HetNet) environment with both unicast and multicast cells, a UE may experience dis-similar radio conditions in different cells. This radio link diversity along with the dual-connectivity capability can be utilized to dynamically arrive at an appropriate transmission mode for individual UEs, enabling improved radio resource utilization in the network. However, the existing 3GPP architecture does not provide any framework to utilize the FeMBMS features together with dual connectivity to improve the network performance.

Figure 1 illustrates a high-level MOOD architecture in a 3GPP network. The MOOD architecture enables improved radio resource utilization in the network by sharing radio-resources (via multicast transmission) when the demand for specific content increases. In case multiple UEs concurrently ask for the same content, the Broadcast Multicast Service

---

[1] We use broadcast and multicast interchangeably throughout the paper.

Center (BM-SC) triggers the flow transition from unicast to multicast mode. Thus, a UE may receive content via eNB/gNB either through Packet Data Network Gateway (PDN-GW) in unicast mode or over Multicast Broadcast Multimedia Service Gateway (MBMS-GW) in multicast mode.

While the MOOD [4] architecture in the 3GPP standards allows for switching between unicast and multicast services, it may solely be based on the demand for specific content. It does not provide any mechanism to utilize factors, such as UE specific radio link quality across different multicast and unicast cells or the load on different network nodes, on eNB/gNB for improved resource utilization and network performance. This becomes all the more limiting when the radio nodes responsible for unicast and multicast transmissions may be separate and the dual connectivity feature is available.

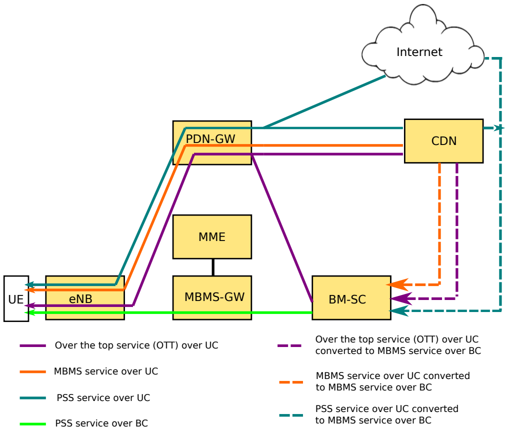

Fig. 1: Illustration of existing 3GPP architecture for MBMS services (courtesy 3GPP).

This paper proposes a simple and novel framework to integrate unicast and multicast transmissions in mobile networks and achieve the objectives mentioned above. We have developed a Software Defined Networking (SDN) based architecture for unified management of unicast and multicast services in 4G, 5G & beyond mobile networks. The architecture enables integration of dual connectivity with MBMS services and provides a mechanism to utilize dual connectivity for improved resource utilization and network performance. We also propose resource allocation algorithms aligned with the proposed architecture. While this paper focuses on the optimal utilization of radio resources, the proposed architectural framework is very generic and can enhance other performance parameters through the convergence of unicast and multicast transmission. For example, the framework enables switching from unicast to multicast transmission to reduce the load on unicast BSs (eNB/gNB) if they are overloaded. As another example, switching can be done based on the load on the core network elements/functions, such as User Plane Function (UPF) or PDN-GW.

A. Related Work

The authors in [5], emphasize the necessity of integration of unicast and multicast services under one framework in 5G networks. However, no specific framework has been discussed in the paper. The authors in [6], analyze the use case where multicast, broadcast and unicast transmissions share resources in 5G New Radio (NR). The authors claim that 5G NR results in better coverage to cell-edge UEs as compared to eMBMS.

In [7], the authors propose and analyze architectures for 5G mobile core network to provision multicast and broadcast services. The proposed architectures are applicable for Digital Terrestrial Television, Public Warning, Internet of Things, Vehicle to Everything (V2X) and Mission Critical Communications (MCC).

The authors in [8], propose a multicast resource allocation scheme where the transmission rate is limited by the worst channel condition experienced by the UE requesting content in the network. A detailed survey on multicasting in wireless access networks has been presented by the authors in [9]. The authors in [10], present the requirement of handling hybrid unicast-multicast approaches for efficient utilization of radio resources in the network. An approach that considers channel conditions experienced by UEs has been proposed by the authors in [11]. Here, UEs with good channel conditions receive content via unicast transmission, whereas UEs experiencing poor channel conditions are delivered content via multicast transmission. However, the selection of the transmission mode is made by individual UEs and may not be efficient due to the unavailability of network wide resource utilization information with the UEs.

In [12], a resource allocation algorithm has been proposed to maximize the Quality of Experience (QoE) of all UEs in an LTE MOOD system. The authors consider physical resource block allocation to each live stream individually, based on the UE demand. The authors in [13], discuss the enhancements made to LTE eMBMS for TV services and MOOD. Furthermore, use cases for each of the enhancements are also described.

The authors in [14], have proposed a mechanism to address the trade-off between fairness and efficiency in resource allocation. The game-theoretic bargaining approach has been used in modeling the fairness and efficiency of the system.

The authors in [15], have proposed joint content delivery of unicast and e-MBMS services to UEs in LTE networks. In [16], the authors jointly optimize the content delivery of unicast and multicast in the network for the given set of UEs in the system. They focus on maximizing the sum-rate of the best effort UEs by adaptive power and subcarrier allocation across UEs.

Various algorithms have been proposed for determining optimal grouping of UEs into different multicast groups in [17], [18]. The authors in [19], have proposed a scheme for grouping UEs into different multicast groups, considering the time varying channel conditions. An efficient and optimal grouping mechanism is proposed by the authors such that the UEs with good channel conditions are not grouped with the UEs experiencing poor channel conditions [20]. However, dynamic traffic in the wireless network has not been considered.

To the best of our knowledge, none of the available works propose a unified framework for the management of unicast and multicast delivery that leverages dual connectivity and

SDN, with utilization of network-wide information for decision making as has been proposed in this work.

Additionally, most of the existing literature, while focusing on objectives, such as efficient resource utilization, maximization of throughput or achievement of fairness across UEs in multicast delivery, consider a fixed number of UEs. This may be particularly limiting as network traffic is typically dynamic in practice. In order to address this limitation, we have considered the dynamic arrival and departure of UEs in the network.

B. Contributions

The key contributions of the paper as follows:

- We propose an SDN based architecture for unified control and management of unicast and multicast services in 4G/5G and beyond mobile wireless networks. It is a generic architecture and can be applied to 4G, 5G and future networks with suitable but small adjustments. It can also be used to integrate other broadcast technologies, such as, Advanced Television Systems Committee (ATSC), with 4G/5G mobile networks.
- We identify relevant network functions as part of the proposed architecture.
- We propose a set of low complexity algorithms for dynamic radio resource allocation for unicast and multicast services in the system. The algorithms consider dynamic UE arrival and departure. We also provide proof of the optimality of the algorithms. The proposed architecture facilitates an efficient utilization of network resources, as demonstrated through these algorithms.
- Finally, an evaluation of algorithms through simulation has been presented.

The rest of the paper is organized as follows: In Section II, we present the proposed SDN based unicast-multicast convergence architecture. The detailed system model is presented in Section III. In Section IV, we propose network assisted dynamic radio resource allocation schemes across unicast and MBMS cells. Simulations results are discussed in Section V. We conclude in Section VI.

II. Proposed System Architecture

In this section, we propose a converged network architecture to deliver unicast and multicast services in a cellular mobile network. Even though the focus of this work is on the efficient use of radio resources, the proposed framework can be used for improved utilization of other network resources as well. The architecture has been proposed in the context of FeMBMS, which enables integration of any unicast service delivery framework with any multicast service delivery framework.

The fundamental concept behind SDN is the separation of control and forwarding (data) planes through a standardized protocol interface [21]. In SDN based architectures, an additional application plane is also present. However, for the sake of simplicity, the application plane has been omitted from the discussion here. SDN provides flexibility to dynamically handle the resource requirements (e.g., radio resources) in a network as per real-time usage patterns and enables efficient utilization.

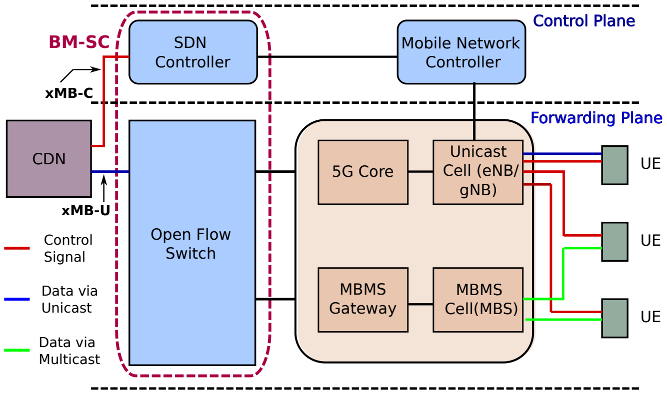

Fig. 2: SDN based converged mobile network architecture for 5G and beyond.

In alignment with the SDN paradigm, network entities in the proposed architecture can be classified as forwarding plane and control plane entities as illustrated in Figure 2. Some of the 4G/5G network elements, i.e., Unicast Core (UPF, etc.), Multicast Core (MBMS Gateway, etc.), eNB/gNB (managing unicast radio transmission), and Multicast Base Station (MBS) (managing dedicated MBMS Cells) are part of the proposed architecture and constitute the forwarding plane. An MBS may also be an eNB/gNB capable of MBMS (multicast/broadcast) transmission. Content Delivery Network (CDN) is also a part of the forwarding plane in the proposed architecture.

BM-SC is a network entity in the existing 3GPP architecture, responsible for receiving content from the CDN and delivering them to the UEs utilizing MBMS delivery framework [4]. We propose a novel SDN based architecture for BM-SC comprising an SDN Controller and OpenFlow (OF) switches. The proposed architecture enables enhanced forwarding capabilities in BM-SC, allowing dynamic and flexible switching of flows across unicast and multicast modes for individual UEs in order to utilize the radio resources efficiently.

The BM-SC OF switches are connected to the mobile core (both multicast and unicast core) on one side and CDN on the other side and are part of the forwarding plane in the proposed architecture. They receive data from the CDN and forward it to the unicast core (UPF) for delivery to the UEs via eNB/gNB (as unicast transmission) or to the multicast core (MBMS Gateway) for delivery via MBMS cells (as multicast transmission).

In addition to the SDN (BM-SC) Controller, an additional control plane function called Mobile Network Controller (MNC) has also been defined as part of the proposed architecture. While SDN Controller is responsible for setting up the data path through underlying OF switches, MNC collects the radio channel quality reports (both for the multicast as well as the unicast cells) from UEs through eNBs/gNBs and decides about the mode for data transmission (unicast or multicast) for individual UEs. This is done with an aim to achieve efficient radio resource utilization. Once decided, MNC also communicates the transmission mode (multicast or unicast) to the SDN controller for each UE. As explained in Section IV later, MNC also needs to allocate radio resources for multicast transmission at the dedicated MBMS Cells (MBS).

Upon receiving the transmission policy or mode (unicast or multicast) for individual UEs from the MNC, the SDN Controller establishes flow rules (data flow) through the underlying OF switches, i.e., for UEs expected to receive transmission in the unicast mode, unicast flows are established and for UEs expected to receive transmission in the multicast mode, multicast flows are set up over the OF switches. The flow rules are configured such that for each unicast flow, the switch performs content replication with the individual UE IP address as the destination address and forwards the replicated content to the unicast core (UPF) to be finally delivered to the UE via eNB/gNB. For multicast UEs, the switch forwards a single copy of the content with a multicast IP address as the destination to the MBMS Gateway to be delivered via the MBMS cell. The separation of control plane functionality in two distinct functions, an SDN Controller and an MNC, improves the scalability and modularity of the proposed architecture.

Dual Connectivity allows a UE to receive data from two BSs concurrently. Here, we propose the use of dual connectivity by UEs to receive data either through an MBS providing multicast service or an eNB/gNB providing unicast service. As shown in Figure 2, a UE is always connected to an eNB/gNB for control communication over a unicast channel. It can also receive data from that eNB/gNB when required. At the same time, UE can also receive data over a multicast channel through an MBS. UE uses its dedicated unicast connectivity to eNB/gNB to provide relevant information, e.g., the link quality of both its unicast and multicast channels. This radio link related information is sent by the eNB/gNB to the MNC for further action, as mentioned above. The information is used by the control plane functions (MNC and SDN Controller) to direct the data to the UE either through the unicast channel or the multicast channel.

Though desirable, due to the advantages of potential implementation simplicity with a separate scheduler and hardware platform, a physically separate dedicated MBMS cell is not a necessity in the proposed framework. It can also be seen as a logically separate dedicated MBMS cell wherein the resources on a physical cell can be divided into separate unicast and multicast resources. Keeping a logically separate MBMS cell (MBS) enables a simple framework for integration of dual connectivity with MBMS services where a user can simultaneously be connected to a unicast BS and receive data from a separate dedicated multicast BS. It also facilitates a flexible convergence of unicast and multicast services, enabling the integration of any multicast technology with any unicast technology.

The proposed SDN based architecture for BM-SC (with a clear separation between the control plane and the data plane functionality) is also better aligned with the 3GPP FeMBMS standard, with the xMB-C (control plane interface between BM-SC and CDN) terminating at the SDN Controller, i.e., at the BM-SC control plane entity and xMB-U (data plane interface between BM-SC and CDN) terminating at the OpenFlow switches, the proposed data plane of BM-SC.

III. System Model

We consider a scenario where UEs are interested in multimedia content (typically live streaming). While the proposed architecture is applicable to both 5G NR and 4G LTE, we consider LTE cells in the system model. The system model considers a dedicated MBMS cell in a region providing multicast service to the UEs inside its coverage area. In addition, one or more LTE cells supporting unicast transmission are also present in that region, overlapping with the coverage area of the multicast cell, as illustrated in Fig 3. We assume multicast cells have larger coverage area than that of a unicast cell. This is also aligned with the larger inter-site distance for MBMS cells, as proposed under FeMBMS [2].

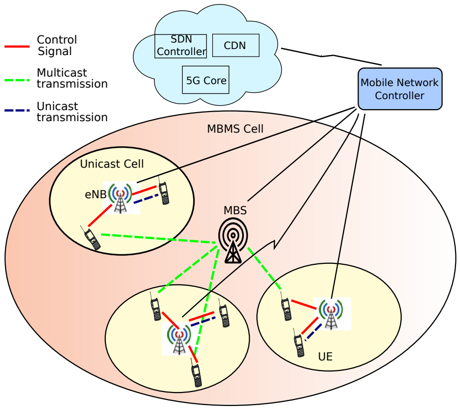

Fig. 3: System Model.

We assume that each UE is connected to a unicast cell in its vicinity called the “anchor cell” here. The unicast cell with the best signal strength for a UE is the “anchor cell” and the eNB controlling the “anchor cell” is the “anchor eNB” for that UE. A UE can receive its desired streaming content either via unicast transmission from its anchor cell or via multicast transmission from an MBMS cell. UE also uses the anchor cell for the exchange of control information with the network.

Each UE is capable of dual-connectivity, i.e., it is connected to the anchor cell for signalling control communication and data can be received either from unicast cell or MBMS cell. Further, we assume that a UE requests only one streaming content at a time. This can be generalized to a UE accessing multiple streams simultaneously where individual content streams require orthogonal radio resources for delivery.

The system model considers dynamic UE arrival, which follows a Poisson process. We also assume that UEs are not mobile and hence their channel conditions do not vary (static users). Although a single dedicated MBMS cell supporting multicast service and multiple LTE cells supporting unicast transmission have been considered in the system model, it can be generalized to include multiple dedicated MBMS cells if we assume that out of all the MBMS cells in the vicinity, UE is served by the one providing the best signal strength. Each UE can be treated as a single point in the considered geographical region, which (the point) can be mapped to one of the MBMS cells (the one with the best signal strength at that point).

We consider that each UE ℓ has a minimum rate requirement R_ℓ to guarantee the required Quality of Service (QoS). When a new UE arrives in the network, it associates itself with an "anchor eNB" and reports the Channel Quality Indicator (CQI) as observed by it, both for the best MBMS cell and the "anchor (unicast) cell" to the "anchor eNB". The channel reports are forwarded by the eNB to the MNC. Based on the collected CQI reports and the required data rate, MNC computes the number of Resource Blocks (RBs) (W_ℓ^{u}, W_ℓ^{m}) needed to serve the UE via unicast and multicast transmissions, respectively as per 3GPP standard [22].

Depending on the objective function, MNC decides the transmission mode (unicast or multicast) for the new UE and directs the SDN Controller to set up the data flow accordingly. In certain scenarios, MNC may also need to change the transmission modes of some existing UEs and hence it also needs to direct the SDN Controller to modify the data flows for such UEs. Once it decides the transmission modes for individual UEs, the MNC also guides the unicast and the MBMS cells in radio resource allocation, taking into account the CQI values and the rate requirements of individual UEs. After receiving the direction from the MNC, the SDN Controller sets up/modifies the unicast and multicast data flows over the OpenFlow switches, which are finally delivered to the UEs either via the unicast or the MBMS cells.

In Table I, we present the notations and their significance, which have been used throughout the paper.

Notations	**Significance**
$\mathcal{N}$	Set of UEs in the system
$\mathcal{U}$	Set of UEs served via unicast transmission
$\mathcal{M}$	Set of UEs served via multicast transmission
W_ℓ^{u}	RBs required to serve UE ℓ via unicast transmission
W_ℓ^{m}	RBs required to serve UE ℓ via multicast transmission
L	Sorted list of all UEs in the ascending order of W_ℓ^{m}
$L[\mu]$	UE stored at index μ in list L
$W_{L[\mu]}^{\mathrm{u}}$	RBs required to serve UE at index μ in L via unicast transmission
$W_{L[\mu]}^{\mathrm{m}}$	RBs required to serve UE at index μ in L via multicast transmission
W^{u}	Total number of RBs required to serve all UEs in set $\mathcal{U}$
W^{m}	Total number of RBs required to serve all UEs in set $\mathcal{M}$
W^{s}	Total number of RBs required to serve all UEs i.e., $\ell \in \mathcal{U} \cup \mathcal{M}$
W^{a}	Additional RBs required to include UE ℓ in multicast set
R_ℓ	Minimum rate requirement of UE ℓ
R_ℓ^{u}	Rate UE ℓ receives via unicast transmission
R_ℓ^{m}	Rate UE ℓ receives via multicast transmission

Table I: Notations and their significance

When a new UE arrives in the system, 3 attributes are assigned to the UE: a unique UE ID $\ell \in \mathbb{Z}_+$, W_ℓ^{m} and W_ℓ^{u} as shown in Figure 4.

UE ID ℓ	W_ℓ^{u}	W_ℓ^{m}

Fig. 4: Illustration of UE structure.

When UE $\ell \in \mathcal{U}$ is served content via unicast transmission, a dedicated set of RBs are allocated to the UE. Thus, the required RBs to serve all UEs in set $\mathcal{U}$ denoted by W^{u} is the aggregate sum of the RBs required by each UE in $\mathcal{U}$. Formally, W^{u} can be defined as follows:

$$W^{\mathrm{u}} = \sum_{\ell \in \mathcal{U}} W_\ell^{\mathrm{u}}. \tag{1}$$

Let W^{m} denote the RBs required for multicast transmission in the system. Unlike UEs in unicast transmission, a common set of RBs are used to serve the UEs in $\mathcal{M}$ for multicast transmission. Therefore, the number of RBs required for multicast transmission in the system is equal to the maximum number of the RBs required by UEs in the set $\mathcal{M}$. Thus, W^{m} is obtained as

$$W^{\mathrm{m}} = \max_{\ell \in \mathcal{M}} W_\ell^{\mathrm{m}}. \tag{2}$$

From Equations (1) and (2), the overall RBs required in the system W^{s} to serve all the UEs i.e., $\mathcal{U} \cup \mathcal{M}$ can be obtained as follows:

$$W^{\mathrm{s}} = W^{\mathrm{u}} + W^{\mathrm{m}}. \tag{3}$$

A. Problem Formulation

All UEs associated with a specific MBMS cell and receiving a particular multimedia content simultaneously constitute a set (or a group). Along with the MBMS cell, each UE is also associated with a unicast cell (anchor cell). As mentioned earlier, we consider the problem of efficient delivery of content to a set of UEs either through unicast or multicast delivery modes wherein any one of the two modes may be utilized for a particular UE.

The system aims to allocate all UEs in $\mathcal{N}$ to the unicast ($\mathcal{U}$) and the multicast ($\mathcal{M}$) sets so that W^{s} required to serve UEs in $\mathcal{N}$ is minimized provided the individual rate requirement of each UE is satisfied.

$$\begin{aligned} \mathbb{A}: \ \min_{\chi} \quad & W^{\mathrm{s}} = W^{\mathrm{u}} + W^{\mathrm{m}} \\ \text{s.t.} \quad & \chi_\ell^{\mathrm{u}} + \chi_\ell^{\mathrm{m}} = 1, \quad \forall \ell \in \mathcal{N}, \\ & R_\ell^{\mathrm{u}}.\chi_\ell^{\mathrm{u}} + R_\ell^{\mathrm{m}}.\chi_\ell^{\mathrm{m}} \geq R_\ell, \ \forall \ell \in \mathcal{N}. \end{aligned} \tag{4}$$

The objective is to determine the optimal allocation χ that minimizes the total number of RBs (or resources) required to serve all UEs in the system. The first constraint states that a UE can be served either via unicast or multicast transmission mode. Here, $\chi_\ell^{\mathrm{u}} \in \{0, 1\}$ denotes UE ℓ is served via unicast cell if $\chi_\ell^{\mathrm{u}} = 1$ otherwise not. Similarly, $\chi_\ell^{\mathrm{m}} \in \{0, 1\}$ denotes UE ℓ is served via MBMS cell if $\chi_\ell^{\mathrm{m}} = 1$. The last constraint reflects that the individual rate requirement of each UE is satisfied.

IV. RESOURCE ALLOCATION ALGORITHM

In this section, we propose a resource allocation mechanism for problem $\mathbb{A}$ mentioned in Equation (4). As mentioned earlier, we consider the dynamic arrival and departure of UEs in the system.

A. User (UE) Arrival

Algorithm 1 is designed to achieve the objective in problem $\mathbb{A}$ upon arrival of a new UE in the system. To achieve the optimal radio resource allocation, the algorithm may (re)distribute existing UEs ($\mathcal{N}$) and the new arrival (UE ID ℓ'), in two disjoint sets, a set of unicast UEs ($\tilde{\mathcal{U}}$) and a set of multicast UEs ($\tilde{\mathcal{M}}$). Algorithm 1 uses a sorted list L of all existing UEs for processing, where UEs are sorted based on their W_ℓ^{m} values. Upon arrival, the new UE (with ID ℓ') is inserted in list L at the appropriate position (based on $W^{\mathrm{m}}_{\ell'}$). Let μ' be the index of UE ℓ' in list L. Note that UE ID and UE index in list L are independent values in $\mathbb{Z}_{+}$. If $W^{\mathrm{m}} \geq W^{\mathrm{m}}_{\ell'}$ (or $W^{\mathrm{m}}_{L[\mu']}$), i.e., the RBs allocated to existing set $\mathcal{M}$ are greater than that of the RBs required for multicast transmission by new UE. Thus, the UE allocation that achieves optimal RB utilization is $\tilde{\mathcal{U}} \leftarrow \mathcal{U}$ and $\tilde{\mathcal{M}} \leftarrow \mathcal{M} \cup \{\ell'\}$ (or $\tilde{\mathcal{M}} \leftarrow \mathcal{M} \cup \{L[\mu']\}$) (line 6). By $L[\mu']$, we denote the UE stored at index μ' in list L. Therefore, W^{s} remains unchanged even after the inclusion of new UE ℓ' in the system.

However, if $W^{\mathrm{m}} < W^{\mathrm{m}}_{L[\mu']}$, optimal allocation may require re-assignment of UEs in the unicast $\mathcal{U}$ and the multicast $\mathcal{M}$ sets. First, include new UE $L[\mu']$ (or ℓ') to unicast set $\mathcal{U}$. Then, set index ν to $|\mathcal{M}| + 1$ in L (which indicates that the set of UEs served via multicast transmission appear before the set of UEs served via unicast transmission mode in list L. This has been discussed in detail in Lemma 1). Next, we check the condition in line 11 iteratively till the last entry in L. If the condition is true then serve UEs via multicast transmission instead of unicast transmission.

We illustrate Algorithm 1 using an example: Consider 6 UEs in the system $\mathcal{N} = \{1, \ldots, 6\}$, where $\ell \in \mathcal{N}$ is unique ID assigned to each UE on arrival. As described in Algorithm 1, UEs are sorted in list L in increasing order of W^{m}_{ℓ} shown in Table II. As is apparent, the optimal allocation of UEs (in $\mathcal{N}$) is $\mathcal{U} = \{3, 6, 4\}$, $\mathcal{M} = \{2, 1, 5\}$ with $W^{\mathrm{m}} = 5$ and $W^{\mathrm{u}} = 5$.

TABLE II :List L of UEs

UE Index in $L[\mu]$	1	2	3	4	5	6
UE ID (ℓ)	2	1	5	3	6	4
$W^{\mathrm{m}}_{L[\mu]}$	3	4	5	7	10	14
$W^{\mathrm{u}}_{L[\mu]}$	2	2	2	1	3	1

$\underbrace{\hphantom{1\quad 2\quad 3}}_{\mathcal{M}}\underbrace{\hphantom{4\quad 5\quad 6}}_{\mathcal{U}}$

When a new UE with ID 7 arrives, let the RBs required to serve the new UE via multicast and unicast transmission be 8 and 3, respectively. Next, new UE is inserted in list L (based on the value of W^{m}_{ℓ}) at index $\mu' = 5$, as shown in Table III. Since $W^{\mathrm{m}}_{L[5]} > W^{\mathrm{m}}$, the UE with ID 7 cannot be served via multicast transmission (i.e., by the MBMS cell) without increasing the required RBs at the MBMS cell. Therefore, it is not added to set $\mathcal{M}$ and instead added to set $\mathcal{U}$ (lines $5-8$) initially. This leads to an initial value of $W^{\mathrm{m}} = 5$, $W^{\mathrm{u}} = 5 + 3 = 8$ and $W^{\mathrm{s}} = 13$ RBs. However, this may not be the optimal allocation of radio resources, as we will observe shortly. In order to achieve the optimal utilization of resources in the system, the new arrival may require the shifting of the new UE from the set $\mathcal{U}$ to $\mathcal{M}$ along with some existing UEs. This is the key insight into the algorithm. The decision of shifting of UEs is performed using the for loop at line 10 and hence list L is traversed from index $\lambda = \mu'$ till the end.

In iteration 1, the condition (at line 11) happens to be true (i.e., $8 - 5 \leq (1 + 3)$) which implies that RBs required to serve UEs (at $L[4]$ and $L[5]$) can be reduced further if served via MBMS cell. Hence, $L[4]$ and $L[5]$ (i.e., UEs with IDs 3 and 7) are shifted to $\mathcal{M}$. Now, update W^{m}, ν and λ to 8, 6, and 6, respectively. Next in iteration 2, condition $W^{\mathrm{m}}_{L[6]} - W^{\mathrm{m}} \leq W^{\mathrm{u}}_{L[6]}$ i.e., $10 - 8 \leq 3$, satisfies. Therefore, $L[6]$ (or UE 6) is also shifted to $\mathcal{M}$, and $W^{\mathrm{m}} = 10$ and $\nu = 7$ are updated. In last iteration, condition $14-10 \leq 1$ false and hence $L[7]$ (or UE 4) continues to remain in set $\mathcal{U}$, to be served via unicast cell as before the arrival of UE 7. Thus, we get optimal allocation ($\tilde{\mathcal{U}} = \{4\}, \tilde{\mathcal{M}} = \{2, 1, 5, 3, 6, 7\}$) on arrival of UE 7 with $W^{\mathrm{m}} = 10$, $W^{\mathrm{u}} = 1$ and $W^{\mathrm{s}} = 11$ (shown in Table III).

Algorithm 1 UE arrival in the network

```
1:  Input: New UE ID $\ell'$, $\mathcal{U}$, $\mathcal{M}$, List $L$
    Precondition: Disjoint sets $\mathcal{U}$ and $\mathcal{M}$ provide optimal RB
    utilization s.t. $\mathcal{N} = \mathcal{U} \cup \mathcal{M}$
2:  Output: Optimal allocation $\tilde{\mathcal{U}}$, $\tilde{\mathcal{M}}$ with new UE $\ell'$
3:  Insert new UE $\ell'$ in sorted list $L$
4:  $\mu'$ = position of UE $\ell'$ in $L$
5:  if $W^{\mathrm{m}} \geq W^{\mathrm{m}}_{L[\mu']}$ then
6:      $\tilde{\mathcal{U}} \leftarrow \mathcal{U}$ and $\tilde{\mathcal{M}} \leftarrow \mathcal{M} \cup \{L[\mu']\}$
7:  else
8:      Update $\mathcal{U} \leftarrow \mathcal{U} \cup \{L[\mu']\}$
9:      $\nu = |\mathcal{M}| + 1$
10:     for $\lambda = \mu', \ldots, length[L]$ do
11:         if $W^{\mathrm{m}}_{L[\lambda]} - W^{\mathrm{m}} \leq \sum_{\mu=\nu}^{\lambda} W^{\mathrm{u}}_{L[\mu]}$ then
12:             $\mathcal{U} \leftarrow \mathcal{U} \backslash \{L[\nu], \ldots, L[\lambda]\}$,
                $\mathcal{M} \leftarrow \mathcal{M} \cup \{L[\nu], \ldots, L[\lambda]\}$
13:             $W^{\mathrm{m}} \leftarrow W^{\mathrm{m}}_{L[\lambda]}$
14:             $\nu \leftarrow \lambda + 1$
15:         end if
16:     end for
17:     $\tilde{\mathcal{U}} \leftarrow \mathcal{U}, \tilde{\mathcal{M}} \leftarrow \mathcal{M}$
18:     Update $W^{\mathrm{u}}$, $W^{\mathrm{s}}$
19: end if
```

TABLE III: Updated list L with new UE

UE Index in $L[\mu]$	1	2	3	4	5	6	7
UE ID (ℓ)	2	1	5	3	7	6	4
$W^{\mathrm{m}}_{L[\mu]}$	3	4	5	7	8	10	14
$W^{\mathrm{u}}_{L[\mu]}$	2	2	2	1	3	3	1

$\underbrace{\hphantom{1\quad 2\quad 3\quad 4\quad 5\quad 6}}_{\tilde{\mathcal{M}}}\underbrace{\hphantom{7}}_{\tilde{\mathcal{U}}}$

We observe from Tables II and III, that the UE with maximum RB requirement in multicast set ($\mathcal{M}$) is always less than that of the UE with the minimum RB requirement in unicast set ($\mathcal{U}$). We give the formal proof of the same in Lemma 1.

Lemma 1. *Suppose allocation* $(\mathcal{U}, \mathcal{M})$ *is the optimal solution to problem* $\mathbb{A}$, *then* $\max_{\ell \in \mathcal{M}} W^{\mathrm{m}}_{\ell} < \min_{\ell \in \mathcal{U}} W^{\mathrm{m}}_{\ell}$.

Proof. Let us consider that UE $\ell^{\star} \in \mathcal{M}$ requires maximum number of RBs for multicast transmission. This implies that $W^{\mathrm{m}} = W^{\mathrm{m}}_{\ell^{\star}}$ using Equation (2). Let us assume that there exists a UE $\ell' \in \mathcal{U}$ such that $W^{\mathrm{m}}_{\ell'} \leq W^{\mathrm{m}}$.

As $W^{\mathrm{m}}_{\ell'} \leq W^{\mathrm{m}}$, therefore if we shift UE ℓ from $\mathcal{U}$ to $\mathcal{M}$ the RBs required for unicast transmission in the system W^{u} are reduced by $W^{\mathrm{u}}_{\ell'}$, with no change in W^{m}. Thus, the RBs required to serve UEs in the system W^{s} can be reduced further when UE ℓ' is served via multicast transmission. However, ($\mathcal{U}$, $\mathcal{M}$) is optimal solution such that $\mathcal{U} \cup \mathcal{M} = \mathcal{N}$, hence W^{s} cannot be reduced further. This leads to contradiction. Therefore, UE ℓ' cannot have $W^{\mathrm{m}}_{\ell'} \leq W^{\mathrm{m}}$, if the allocation is optimal. Thus, $\max\limits_{\ell \in \mathcal{M}} W^{\mathrm{m}}_{\ell} < \min\limits_{\ell \in \mathcal{U}} W^{\mathrm{m}}_{\ell}$, always hold. □

Lemma 2. *Suppose the overall system level resource requirement W^{s} is optimal for UEs $\ell \in \mathcal{U} \cup \mathcal{M}$. Let $\tilde{W}^{\mathrm{s}}$ be the optimal number of RBs after arrival of a new UE ℓ' in the system, i.e., for UEs $\ell \in \mathcal{U} \cup \mathcal{M} \cup \{\ell'\}$. Then, $\tilde{W}^{\mathrm{s}} \geq W^{\mathrm{s}}$.*

Proof. There can be two cases on arrival of a new UE ℓ' in the system: (i) UE ℓ' is served via unicast, or (ii) UE ℓ' is served via multicast.

Case (i): When new UE ℓ' is served via unicast transmission, W^{u} increases to $\tilde{W}^{\mathrm{u}} = W^{\mathrm{u}} + W^{\mathrm{u}}_{\ell'}$. However, W^{m} does not change, i.e., $\tilde{W}^{\mathrm{m}} = W^{\mathrm{m}}$. Thus, if new UE is delivered content via unicast, system level RB requirement $\tilde{W}^{\mathrm{s}}$ is always greater than the previous system level RB requirement W^{s}, i.e., $\tilde{W}^{\mathrm{s}} > W^{\mathrm{s}}$.

Case (ii): When new UE ℓ' is served via multicast, then there are two possibilities based on $W^{\mathrm{m}}_{\ell'}$: (a) $W^{\mathrm{m}}_{\ell'} \leq W^{\mathrm{m}}$, then $\tilde{W}^{\mathrm{m}} = W^{\mathrm{m}}$. Therefore, $\tilde{W}^{\mathrm{s}} = W^{\mathrm{s}}$.

(b) $W^{\mathrm{m}}_{\ell'} > W^{\mathrm{m}}$, then W^{m} must increase at least to $\tilde{W}^{\mathrm{m}} = W^{\mathrm{m}}_{\ell'}$ to include UE ℓ' in multicast set. Let us define the additional RBs required in MBMS cell as $\tilde{W}^{a} = W^{\mathrm{m}}_{\ell'} - W^{\mathrm{m}}$. With the increase of RBs in MBMS cell from W^{m} to $\tilde{W}^{\mathrm{m}}$, all UEs in $\mathcal{U}$ with $W^{\mathrm{m}}_{\ell} \leq \tilde{W}^{\mathrm{m}}$ also shift to multicast transmission mode using Lemma 1. Let set $\mathcal{Z}$ contains all UEs (with $W^{\mathrm{m}}_{\ell} \leq W^{\mathrm{m}}_{\ell'}$) which shift to multicast from unicast transmission. Therefore, $\tilde{\mathcal{U}} = \mathcal{U} \backslash \mathcal{Z}$ and $\tilde{\mathcal{M}} = \mathcal{M} \cup \mathcal{Z} \cup \{\ell'\}$, where $\tilde{\mathcal{U}}$ and $\tilde{\mathcal{M}}$ constitute an optimal allocation after arrival of UE ℓ' in the system.

Let $W^{\mathrm{m}}_{\ell^{\star}} = \max\limits_{\ell \in \mathcal{Z}} W^{\mathrm{m}}_{\ell}$. Without UE ℓ' in the system, $(\mathcal{U}, \mathcal{M})$ being the optimal allocation, UEs $j \in \mathcal{Z} \subset \mathcal{U}$ satisfy Equation (5).

$$W^{a} = W^{\mathrm{m}}_{\ell^{\star}} - W^{\mathrm{m}} > \sum_{\ell \in \mathcal{Z}} W^{\mathrm{u}}_{\ell}. \tag{5}$$

Here, W^{a} is the additional RBs required to shift all UEs in $\mathcal{Z}$ from unicast to multicast. Equation (5) reflects the fact that shifting UEs from $\mathcal{U}$ to $\mathcal{M}$ will result in higher RB requirement if the allocation is optimal. As $W^{\mathrm{m}}_{\ell'} \geq W^{\mathrm{m}}_{\ell^{\star}}$, therefore $\tilde{W}^{a} \geq W^{a}$ (using definitions of $\tilde{W}^{a}$ and W^{a}). To serve UEs in $(\mathcal{U} \cup \mathcal{M})$, the RBs required W^{s} are given as follows:

$$W^{\mathrm{s}} = W^{\mathrm{m}} + W^{\mathrm{u}} = W^{\mathrm{m}} + \sum_{\ell \in \{\mathcal{U} \backslash \mathcal{Z}\}} W^{\mathrm{u}}_{\ell} + \sum_{\ell \in \mathcal{Z}} W^{\mathrm{u}}_{\ell}. \tag{6}$$

The overall (unicast + multicast) system level RBs $\tilde{W}^{s}$ for UEs in $(\mathcal{U} \cup \mathcal{M} \cup \ell)$ are given as

$$\tilde{W}^{\mathrm{s}} = \tilde{W}^{\mathrm{m}} + \tilde{W}^{\mathrm{u}} = W^{\mathrm{m}} + \tilde{W}^{a} + \sum_{\ell \in \{\mathcal{U} \backslash \mathcal{Z}\}} W^{\mathrm{u}}_{\ell}. \tag{7}$$

From Equations (6) and (7), we obtain $\tilde{W}^{\mathrm{s}} > W^{\mathrm{s}}$ as $\tilde{W}^{a} > \sum\limits_{\ell \in \mathcal{Z}} W^{\mathrm{u}}_{\ell}$ (using Equation (5)). □

Theorem 1. *Algorithm 1 provides optimal solution on UE arrival in the network.*

Proof. Let $(\mathcal{U}, \mathcal{M})$ be the optimal solution, when UE ℓ' is not in the system. The RBs required to serve UEs in sets $\mathcal{U}$ and $\mathcal{M}$ are W^{u} and W^{m}, respectively. Thus, the total RBs required in the system are $W^{\mathrm{s}} = W^{\mathrm{u}} + W^{\mathrm{m}}$. The required proof can be split into two cases: Case (i) new UE ℓ' has $W^{\mathrm{m}}_{\ell'} \leq W^{\mathrm{m}}$ and Case (ii) new UE ℓ' has $W^{\mathrm{m}}_{\ell'} > W^{\mathrm{m}}$.

Case (i): From Lemma 2, on arrival of a new UE, the best possibility is that the total RBs required in the system remains unchanged, i.e., $\tilde{W}^{\mathrm{s}} = W^{\mathrm{s}}$, where $\tilde{W}^{\mathrm{s}}$ is the total RBs required in the system on inclusion of new UE ℓ'. As we know that $W^{\mathrm{m}} \geq W^{\mathrm{m}}_{\ell'}$, therefore when UE ℓ' is served via multicast transmission, the total RBs required in the system does not change. Thus, the optimal allocation inclusive of UE ℓ' becomes $\tilde{\mathcal{U}} = \mathcal{U}$ and $\tilde{\mathcal{M}} = \mathcal{M} \cup \{\ell'\}$ with $\tilde{W}^{\mathrm{s}} = W^{\mathrm{s}}$. We can see that Algorithm 1 achieves the same allocation, as shown by lines 6 and 7.

Case (ii): If UE ℓ' has $W^{\mathrm{m}}_{\ell'} > W^{\mathrm{m}}$, this implies $\tilde{W}^{\mathrm{s}} > W^{\mathrm{s}}$ (using Lemma 2). Let $(\tilde{\mathcal{U}}, \tilde{\mathcal{M}})$ obtained from Algorithm 1 is not optimal. Then there are two possibilities to achieve the optimal allocation.

- Shift UEs from $\tilde{\mathcal{M}}$ to $\tilde{\mathcal{U}}$ to get the optimal allocation: Suppose set of UE in $\mathcal{Z} \subset \tilde{\mathcal{M}}$ are shifted from $\tilde{\mathcal{M}}$ to $\tilde{\mathcal{U}}$. Let $\mathcal{T} = \{L[\tilde{\mu}], \ldots, L[\hat{\mu}]\}$, where $\tilde{\mu}, \ldots, \hat{\mu}$ are UE indices in L corresponding to UEs in $\mathcal{Z}$, sorted in increasing order of $W^{\mathrm{m}}_{L[\mu]}$. Suppose $W^{\mathrm{m}}_{L[\hat{\mu}]} = \tilde{W}^{\mathrm{m}}$. Note that if $\mathcal{T}$ does not include UE $L[\hat{\mu}]$, then shifting UEs from multicast to unicast mode does not change W^{m}, however, W^{u} would increase. Thus, W^{s} increases further. Therefore, $L[\hat{\mu}]$ must be included in $\mathcal{T}$ to decrease the RBs required to serve UEs. Let $\hat{W}^{\mathrm{u}}$ and $\hat{W}^{\mathrm{m}}$ are RBs required to serve UEs in $\hat{\mathcal{U}}$ and $\hat{\mathcal{M}}$, respectively obtained after UEs shift. Then $W^{\mathrm{m}}_{L[\hat{\mu}]} - \hat{W}^{\mathrm{m}} > \sum_{j=\tilde{\mu}}^{\hat{\mu}} W^{\mathrm{u}}_{L[\mu]}$ to decrease overall required RBs. However, $W^{\mathrm{m}}_{L[\hat{\mu}]} - \hat{W}^{\mathrm{m}} > \sum_{\mu=\tilde{\mu}}^{\hat{\mu}} W^{\mathrm{u}}_{L[\mu]}$ can never be true as per UE shifting strategy mentioned in line 11 of Algorithm 1. This leads to contradiction. Hence, $\tilde{\mathcal{U}}$ and $\tilde{\mathcal{M}}$ constitute the optimal allocation.
- Shift UEs from $\tilde{\mathcal{U}}$ to $\tilde{\mathcal{M}}$ to achieve optimal allocation: Proof is similar to the previous scenario. Suppose set of UEs $\mathcal{Z} \subset \tilde{\mathcal{U}}$ are shifted from $\tilde{\mathcal{U}}$ to $\tilde{\mathcal{M}}$. After shifting of the UEs, let the allocation becomes $(\hat{\mathcal{U}}, \hat{\mathcal{M}})$. Suppose set $\mathcal{T} = \{L[\mu'], \ldots, L[\mu'']\}$ contains UEs sorted in increasing order of $W^{\mathrm{m}}_{L[\mu]}$, for all $L[\mu] \in \mathcal{T}$. The UEs shifted from $\tilde{\mathcal{U}}$ to $\tilde{\mathcal{M}}$ must satisfy $\hat{W}^{\mathrm{m}} - \tilde{W}^{\mathrm{m}} \leq \sum_{\mu=\mu'}^{\mu''} W^{\mathrm{u}}_{L[\mu]}$. However, Algorithm 1 traverses list L till the end to ensure that if shifting of UEs result in reduced RB requirement then UEs are already shifted to achieve $\tilde{\mathcal{U}}$ and $\tilde{\mathcal{M}}$. Thus, $\hat{W}^{\mathrm{m}} - \tilde{W}^{\mathrm{m}} \leq \sum_{\mu=\mu'}^{\mu''} W^{\mathrm{u}}_{L[\mu]}$ condition never holds, hence $(\tilde{\mathcal{U}}, \tilde{\mathcal{M}})$ remains optimal allocation. □

B. User (UE) Departure

As the main focus of the system is optimal utilization of the RBs, we are also required to consider the effect of UE departures from the system. Algorithm 2 presents the (re)allocattion of UEs across unicast and multicast sets, when a

UE departs from the system resulting in optimal RB utilization to the unicast and the MBMS cells.

Algorithm 2 UE departure from the network

```
1:  Input: $\mathcal{U}$, $\mathcal{M}$, UE $\ell'$, List $L$
    Precondition: Disjoint sets $\mathcal{U}$ and $\mathcal{M}$ provide optimal RB utilization s.t. $\mathcal{N} = \mathcal{U} \cup \mathcal{M}$
2:  Output: Optimal allocation $\tilde{\mathcal{U}}$, $\tilde{\mathcal{M}}$ without UE $\ell'$
3:  $\mu' \leftarrow index(\ell', L)$
4:  Update $L \leftarrow L \backslash \{L[\mu']\}$
5:  if $\ell' \in \mathcal{U}$ then
6:      $\tilde{\mathcal{U}} \leftarrow \mathcal{U} \backslash \{L[\mu']\}$ and $\tilde{\mathcal{M}} \leftarrow \mathcal{M}$
7:  else
8:      $\mathcal{M} \leftarrow \mathcal{M} \backslash \{L[\mu']\}$
9:      Set $\ell^\star \leftarrow \arg\max_{\ell \in \mathcal{M}} W^{\mathrm{m}}_{\ell}$
10:     Set $\mu^\star \leftarrow index(\ell^\star, L)$, $W^{\mathrm{m}} = W^{\mathrm{m}}_{L[\mu^\star]}$
11:     for $\mu = (\mu^\star - 1), \dots, 1$ do
12:         if $W^{\mathrm{m}} - W^{\mathrm{m}}_{L[\mu]} > \sum_{\mu=\mu+1}^{\mu^\star} W^{\mathrm{u}}_{L[\mu]}$ then
13:             $\mathcal{U} \leftarrow \mathcal{U} \cup \{L[\mu+1], \dots, L[\mu^\star]\}$,
                $\mathcal{M} \leftarrow \mathcal{M} \backslash \{L[\mu+1], \dots, L[\mu^\star]\}$
14:             Update $W^{\mathrm{m}} = W^{\mathrm{m}}_{L[\mu]}$
15:             $\mu^\star \leftarrow \mu$
16:         end if
17:         $\mu \leftarrow \mu - 1$
18:     end for
19:     $\tilde{\mathcal{U}} \leftarrow \mathcal{U}$ and $\tilde{\mathcal{M}} \leftarrow \mathcal{M}$
20: end if
```

The first step is to remove the departing UE ℓ' from the list L (UE ℓ' stored at index μ' in list L) mentioned in lines 3 and 4 of the Algorithm. If UE $\ell' \in \mathcal{U}$ leaves the system (i.e., $W^{\mathrm{m}}_{\ell'} > W^{\mathrm{m}}$), optimal allocation is obtained by removing the UE ℓ' from $\mathcal{U}$ i.e., $\tilde{\mathcal{U}} = \mathcal{U} \backslash \{L[\mu']\}$, while $\tilde{\mathcal{M}} = \mathcal{M}$ remains unchanged (line 6).

If UE $\ell' \in \mathcal{M}$ leaves the system then update multicast set $\mathcal{M}$ by removing UE ℓ'. Then, set index $\mu^\star$ to UE in list L with maximum RB requirement in $\mathcal{M}$ and hence $W^{\mathrm{m}} = W^{\mathrm{m}}_{L[\mu^\star]}$. Next, the difference between the required RBs in multicast transmission for the last UE (UE with maximum RB requirement) in $\mathcal{M}$ (UE index $\mu^\star$) and the second last UE (UE index $\mu^\star - 1$) in $\mathcal{M}$ is calculated, i.e., $W^{\mathrm{m}}_{L[\mu^\star]}$(or W^{m}) $- W^{\mathrm{m}}_{L[\mu^\star - 1]}$. Intuitively, algorithm evaluates whether UE shifting is required or not for optimal utilization of RBs. The calculated difference is compared with the required unicast radio resource of $L[\mu^\star]$. If the difference is less than or equal to $W^{\mathrm{u}}_{L[\mu^\star]}$, i.e., the unicast resource requirements of the last UE in $\mathcal{M}$ is more than $W^{\mathrm{m}} - W^{\mathrm{m}}_{L[\mu^\star - 1]}$, no change in $\mathcal{U}$ and $\mathcal{M}$ are required and the existing $\mathcal{U}$ and $\mathcal{M}$ sets remain optimal from the perspective of resource requirement. But if $W^{\mathrm{m}} - W^{\mathrm{m}}_{L[\mu^\star - 1]}$ is greater than $W^{\mathrm{u}}_{L[\mu^\star]}$ then optimal allocation is obtained by moving UE $\mu^\star$ from $\mathcal{M}$ to $\mathcal{U}$. This process is repeated in reverse order for all UEs in $\mathcal{M}$ by decreasing the loop index μ iteratively (line 11).

In order to illustrate the departure algorithm, let us take the reverse case of UE with ID 7 departing from the example given earlier. As shown in Table III, the following distribution of UEs across the two sets, $\mathcal{U} = \{4\}$ and $\mathcal{M} = \{2, 1, 5, 3, 7, 6\}$ achieves optimal allocation of RBs in the system. Now, when UE with ID 7 departs from $\mathcal{M}$ (and the system), the distribution of the remaining UEs with IDs $\{1, \dots, 6\}$ across $\mathcal{U}$ and $\mathcal{M}$ changes again and it goes back to the allocation given in Table II, i.e., the one before the arrival of UE 7 in the system. Thus, we get $\tilde{\mathcal{U}} = \{3, 6, 4\}$ and $\tilde{\mathcal{M}} = \{2, 1, 5\}$, which means that after the departure of the UE with ID 7, some UEs from set $\mathcal{M}$ are moved to set $\mathcal{U}$ to achieve the optimal allocation of resources (RBs) in the system. Now, we prove the optimality of Algorithm 2.

Theorem 2. *Algorithm 2 provides optimal solution on UE departure from the network.*

Proof. To prove $(\tilde{\mathcal{U}}, \tilde{\mathcal{M}})$ is optimal allocation after UE departure, we consider two scenarios as mentioned in Algorithm 2:
Scenario 1: UE $\ell' \in \mathcal{U}$ leaves the network
When unicast UE ℓ' leaves the network then the allocation obtained from Algorithm 2 is $\tilde{\mathcal{U}} = \mathcal{U} \backslash \{L[\mu']\}$(or $\tilde{\mathcal{U}} = \mathcal{U} \backslash \{\ell'\}$) and $\tilde{\mathcal{M}} = \mathcal{M}$. Hence, the RBs required to serve UEs in $\tilde{\mathcal{U}}$ reduces to $\tilde{W}^{\mathrm{u}} = W^{\mathrm{u}} - W^{\mathrm{u}}_{\ell'}$. However, the RBs required to serve UEs in $\tilde{\mathcal{M}}$ remains unchanged i.e., $\tilde{W}^{\mathrm{m}} = W^{\mathrm{m}}$. Thus, the total RBs required to serve UEs in $\tilde{\mathcal{U}}$ and $\tilde{\mathcal{M}}$ are

$$\tilde{W}^{\mathrm{s}} = \tilde{W}^{\mathrm{u}} + \tilde{W}^{\mathrm{m}} = W^{\mathrm{u}} - W^{\mathrm{u}}_{\ell'} + W^{\mathrm{m}}.$$

Suppose $\tilde{W}^{s}$ is not minimum, this implies that the allocation $(\tilde{\mathcal{U}}, \tilde{\mathcal{M}})$ is not optimal. The possible options to obtain the optimal allocation are as follows:

- Shift UEs from $\tilde{\mathcal{U}}$ to $\tilde{\mathcal{M}}$ to achieve the optimal allocation: Suppose a set of UEs $\mathcal{Z} \subset \tilde{\mathcal{U}}$ are shifted from $\tilde{\mathcal{U}}$ to $\tilde{\mathcal{M}}$. Let $\mathcal{T} = \{L[\tilde{\mu}], \dots, L[\hat{\mu}]\}$, where $\tilde{\mu}, \dots, \hat{\mu}$ are indices in L corresponding to set of UEs $\mathcal{Z}$, arranged in increasing order of $W^{\mathrm{m}}_{L[\mu]}$, $\forall\ L[\mu] \in \mathcal{T}$. Since, allocation $(\mathcal{U}, \mathcal{M})$ is optimal, therefore $W^{\mathrm{m}}_{L[\hat{\mu}]} - W^{\mathrm{m}} > \sum_{\mu=\tilde{\mu}}^{\hat{\mu}} W^{\mathrm{u}}_{L[\mu]}$. However, to further decrease the RB requirement $W^{\mathrm{m}}_{L[\hat{\mu}]} - W^{\mathrm{m}} \leq \sum_{\mu=\tilde{\mu}}^{\hat{\mu}} W^{\mathrm{u}}_{L[\mu]}$ must satisfy. This is a contradiction. Hence, $(\tilde{\mathcal{U}}, \tilde{\mathcal{M}})$ remains optimal allocation.
- Shift UEs from $\tilde{\mathcal{M}}$ to $\tilde{\mathcal{U}}$ to achieve the optimal allocation: When a set of UEs are shifted, optimality of solution can be proved using contradiction, similar to the case of UE shifting from $\tilde{\mathcal{U}}$ to $\tilde{\mathcal{M}}$.

Scenario 2: UE $\ell' \in \mathcal{M}$ leaves the network
The proof is similar to Case (ii) of UE arrival algorithm. ☐

C. Optimal Resource Allocation Algorithm

Algorithm 3 Resource Allocation Algorithm

```
Input: Event = {arrival, departure}
Output: Optimal UE allocation $\tilde{\mathcal{U}}$, $\tilde{\mathcal{M}}$ with the given action for UE $\ell'$.
if Event = arrival then
    Go to Algorithm 1
else
    Go to Algorithm 2
end if
```

In this section, we present a dynamic resource allocation algorithm for converged architecture described in Section II.

When there is either UE arrival or UE departure in the system, Algorithm 3 is executed. In Algorithm 3, we propose OPTImal resource allocation in Converged Unicast muLticast networks (OPTICUL). As per the given input, OPTICUL provides an optimal UE allocation to unicast and MBMS cells such that the total RBs required to serve all UEs (in set $\mathcal{N}$) are minimized.

Corollary 1. *The computational complexity of the OPTICUL algorithm is* $\mathcal{O}(|\mathcal{U}|)$ *in case of the arrival of a new UE and* $\mathcal{O}(|\mathcal{M}|)$ *in case of the departure of a UE, where* $|.|$ *denotes the cardinality of a set.*

When a new UE arrives in the system, the OPTICUL algorithm achieves the optimal solution by shifting UEs from set $\mathcal{U}$ to $\mathcal{M}$. Even if the optimal solution requires shifting of all UEs from set $\mathcal{U}$ to $\mathcal{M}$, the maximum number of operations required is equal to $|\mathcal{U}|$. When a UE departs from the system, the OPTICUL algorithm achieves the optimal solution in $\mathcal{O}(|\mathcal{M}|)$ operations. The optimal solution may require shifting of UEs from set $\mathcal{M}$ to $\mathcal{U}$, which results in shifting of up to $|\mathcal{M}|$ UEs, i.e., at most $|\mathcal{M}|$ operations is required. Thus, the OPTICUL algorithm determines the optimal solution in polynomial time.

V. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed algorithm (OPTICUL) through simulations. The simulations are performed in MATLAB [23]. Next, we describe the simulation settings.

A. Simulation Settings

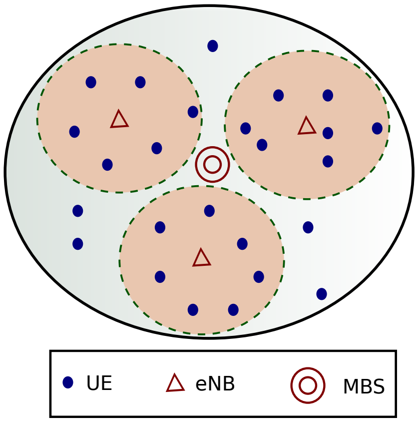

Fig. 5: Illustration of simulation settings.

We consider a system with a single MBMS cell and 3 unicast cells. The simulation parameters are considered as per 3GPP standard [24] and are listed in Table IV. UEs are distributed uniformly across the region in MBMS cell overlapping with unicast cells. Each UE is connected with both unicast and multicast cells simultaneously. However, UE is served either from a unicast or multicast cell at any given time. Based on the CQI and the data rate requirement of the individual UE, the required number of RBs (for both the unicast and the multicast transmission) to serve the UE is computed as per 3GPP standard [22]. First, we map CQI to the modulation coding scheme (MCS). Then, for the given rate requirement of each UE, MCS to transport block size (TBS) mapping is performed as given in Table 7.1.7.1-1 in [22]. The required number of RBs is obtained based on the TBS index using Table 7.1.7.2.1-1 in [22]. We consider dynamic arrival and departure of UEs in the system. Moreover, the same content is requested by all the UEs in the system. All the results are averaged over 500 iterations.

TABLE IV : Simulation parameters

Parameters	Values
Carrier Frequency	2 GHz
Number of MBMS cell	1
Number of LTE cells (unicast)	3
Channel Model	3D-UMa
Noise Figure-UE	9dB
Mobility	3km/h
UE Distribution	Uniform
Multicast BS (MBMS) Model	
Coverage radius	250 m
Transmit Power	43 dBm
Antenna Height	25 m
Antenna	Omni-directional
eNB BS (Unicast) Model	
Coverage radius	100 m
Transmit Power	37 dBm
Antenna Height	15 m
Antenna	Omni-directional

B. Performance Comparison

Now, we evaluate the performance of the algorithm proposed in Section IV. The proposed algorithm achieves efficient utilization, considering that sufficient resources are available to serve all UEs in the system. The performance of the proposed algorithm is compared against the "multicast only transmission scheme" wherein all UEs are served via multicast transmission mode through the MBMS cell. In the multicast scheme, all UEs are served via the MBMS cell. Therefore, the number of RBs required to serve all UEs is equal to the RBs required by the UE with the worst CQI associated with MBMS cell.

- *UE Arrival Only*

We consider a scenario where a content is streaming for a duration of 45 minutes. UE arrival process follows poisson distribution with average arrival rate $\lambda_a = 3$ UEs per minute. Each UE has a data rate requirement of $R = 3.5$ Mbps. We compare the performance of Algorithm 1 with the multicast scheme, assuming that once a UE arrives in the system, it stays for the remaining duration of the content streaming. In Figure 6, we observe that Algorithm 1 outperforms the multicast scheme in terms of RBs. The reason behind poor performance in multicast scheme is due to the fact that in multicast transmission, the required number of RBs depends on the channel condition experienced by the worst UE in the system. However, Algorithm 1 provides optimal resource utilization by splitting the UEs across unicast and multicast

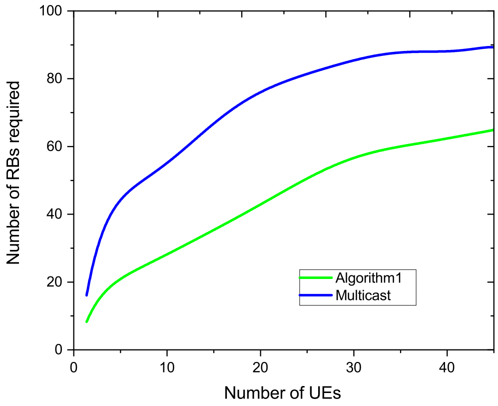

Fig. 6: Resource Utilization (RBs required) vs. number of UEs [$R = 3.5$ Mbps].

cells based on the channel conditions of each UE in the system. Thus, UEs with poor channel conditions for multicast transmission are served via unicast cell.

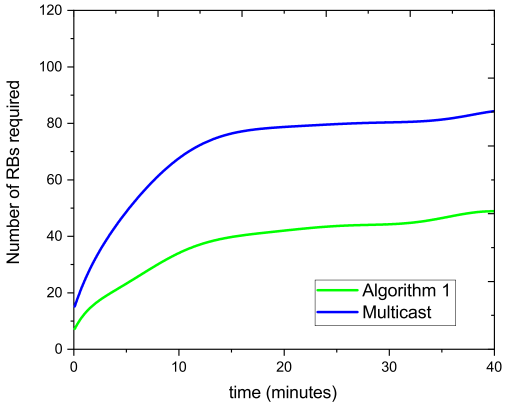

Fig. 7: Resource Utilization (RBs required) vs. numer of UEs [For $t = [0\ 10] : \lambda_b = 5$; and $t = [10\ 40] : \lambda_a = 1$, $R = 3.5$ Mbps].

Next, we analyze the performance when UE arrival process does not have uniform uniform distribution. Typically, we observe a large number of UEs arrive at the beginning (say for initial 10 minutes, bursty traffic) and after that UE arrival rate decreases considerably. We simulate the scenario by considering that for initial 10 minutes, the UE arrival follows poisson process with average arrival rate $\lambda_b = 5$ UEs per minute and thereafter the UE arrival rate goes down to $\lambda_a = 1$ UEs per minute. The data rate requirement of each UE is 3.5 Mbps. In Figure 7, the trend observed is similar to that of the previous case, except the fact that the number of RBs required to serve UEs increases sharply due to bursty traffic.

We also analyze the effect of required data rate (of UEs) on the performance of the proposed mechanism against the multicast scheme. Again, we consider UE arrival as a poisson process with average arrival rate $\lambda_a = 3$ UEs per minute for the duration of content streaming. We compare the number of RBs saved (difference between RBs required in multicast and Algorithm 1) in Figure 8, while varying the data rate requirement of each UE as 2.0 Mbps, 2.5 Mbps and 3.0 Mbps. We observe that the number of RBs saved by the proposed algorithm over the multicast scheme increases as the number of UEs increases in the system. To satisfy the higher data rate requirements for a given channel condition, UE requires more number of RBs. Therefore, as the data rate requirement of UE requesting the same content increases a significant improvement (in terms of the number of RBs saved) is observed in the performance of Algorithm 1.

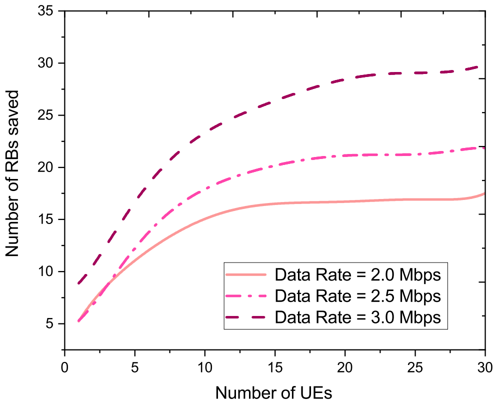

Fig. 8: Comparison of the number of RBs saved in Algorithm 1 against multicast scheme at different UE data rates [Data rates $R = \{2.0, 2.5, 3.0\}$ Mbps].

- *UE Arrival and Departure*

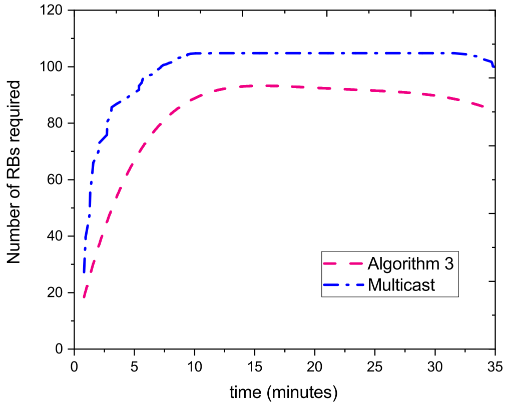

Fig. 9: Resource Utilization (RBs required) vs. time [For $t = [0, 10]$:$\lambda_a = 4$ and $t = [10, 45]$: $\mu_d = 1/10$, $R = 3.5$ Mbps].

Next, we consider a scenario where UEs arrive as well as depart from the system. The total duration of content streaming is 40 minutes. In this scenario, UEs arrive at an average arrival rate $\lambda_a = 4$ UEs per minute for initial 10 minutes. After that we consider only UE departures in the system. The departure time between UEs in the system is exponentially distributed with parameter μ_d. We choose $\mu_d = 1/10$. From the simulations, it is observed that the resource requirement to satisfy all the UEs in the system increases in the beginning, then remains constant and later decreases as the UEs depart using Algorithm 3 (OPTICUL) (Figure 9). However, in the multicast transmission scheme, RB requirement in the system is effected only if UE experiencing the worst channel condition departs. As Algorithm 3 allows switching of UEs from unicast to multicast transmission, we achieve optimal RB utilization.

VI. CONCLUSIONS

In this article, we have proposed a novel SDN-based architecture which enables the convergence of unicast and multicast services in the next-generation mobile network. The proposed architecture supports integration of dual connectivity with MBMS services and is able to utilize dual connectivity for improved radio resource utilization and network performance through flexible switching between unicast and multicast transmission modes, hitherto not possible in the existing 3GPP architecture. We have also developed an efficient radio resource allocation algorithm aligned with the proposed architecture. Our algorithm considers dynamic arrival/departure of UEs. We prove the optimality of the proposed algorithm. The proposed algorithm also points towards the usage of "flexible bandwidth in MBMS cells", currently under discussion in 3GPP [25]. Using extensive simulations, we observe that the proposed algorithm outperforms the multicast only scheme in various considered scenarios.

While this work focuses on the optimization of radio resources, the proposed architectural framework can easily be extended to optimize other network parameters. For example, the load on different network nodes (UPF/eNB/gNB) can be distributed evenly by flexibly using multicast or unicast transmission modes for different UEs in the system.

REFERENCES

[1] KPMG and EROS Now, "Unravelling the Digital Video Consumer," 2019, [Online] Available: https://home.kpmg/in/en/home/insights/2019/09/ott-digital-video-market-consumer-india.html.

[2] 3GPP TS 23.246 v16.1.0, "Multimedia Broadcast/Multicast Service; Architecture and Functional description," 2019, [Online] Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx.specificationId=829.

[3] 3GPP TS 26.346 v16.6.0, "Multimedia Broadcast/Multicast Service improvements; Protocols and codecs," 2020, [Online] Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1452.

[4] 3GPP TR 26.849 v12.1.0, "Multimedia Broadcast/Multicast Service (MBMS) improvements; MBMS operation on demand," 2015, [Online] Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1476.

[5] D. Gomez-Barquero, D. Navratil, S. Appleby, and M. Stagg, "Point-to-Multipoint Communication Enablers for the Fifth Generation of Wireless Systems," *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 53–59, 2018.

[6] W. Guo, M. Fuentes, L. Christodoulou, and B. Mouhouche, "Roads to Multimedia Broadcast Multicast Services in 5G New Radio," in *IEEE BMSB*, 2018, pp. 1–5.

[7] T. Tran, D. Navrátil, P. Sanders, J. Hart, R. Odarchenko, C. Barjau, B. Altman, C. Burdinat, and D. Gomez-Barquero, "Enabling Multicast and Broadcast in the 5G Core for Converged Fixed and Mobile Networks," *IEEE Transactions on Broadcasting*, 2020.

[8] J. Liu, W. Chen, Z. Cao, and K. B. Letaief, "Dynamic Power and Subcarrier Allocation for OFDMA-based Wireless Multicast Systems," in *IEEE ICC*, 2008, pp. 2607–2611.

[9] R. O. Afolabi, A. Dadlani, and K. Kim, "Multicast Scheduling and Resource Allocation Algorithms for OFDMA-based Systems: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 240–254, 2012.

[10] S. Pizzi, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and G.-M. Muntean, "A Unified Approach for Efficient Delivery of Unicast and Multicast Wireless Video Services," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8063–8076, 2016.

[11] P. Scopelliti, G. Araniti, G.-M. Muntean, A. Molinaro, and A. Iera, "A Hybrid Unicast-Multicast Utility-based Network Selection Algorithm," in *IEEE BMSB*, 2017, pp. 1–6.

[12] R. Kaliski, C.-C. Chou, H.-Y. Meng, and H.-Y. Wei, "Dynamic Resource Allocation Framework for MooD (MBMS Operation on-demand)," *IEEE Transactions on Broadcasting*, vol. 62, no. 4, pp. 903–917, 2016.

[13] R. Kaliski, C.-C. Chou, and H.-Y. Wei, "Further Enhanced Multimedia Broadcast/Multicast Service in LTE-Advanced Pro," *IEEE Communications Standards Magazine*, vol. 3, no. 3, pp. 44–51, 2019.

[14] L. Militano, M. Condoluci, G. Araniti, and A. Iera, "Multicast Service Delivery Solutions in LTE-Advanced Systems," in *IEEE ICC*, 2013, pp. 5954–5958.

[15] J. F. Monserrat, J. Calabuig, A. Fernandez-Aguilella, and D. Gomez-Barquero, "Joint Delivery of Unicast and E-MBMS Services in LTE Networks," *IEEE Transactions on Broadcasting*, vol. 58, no. 2, pp. 157–167, 2012.

[16] D. Lee, J. So, and S. R. Lee, "Power Allocation and Subcarrier Assignment for Joint Delivery of Unicast and Broadcast Transmissions in OFDM Systems," *IEEE Journal of Communications and Networks*, vol. 18, no. 3, pp. 375–386, 2016.

[17] G. Araniti, V. Scordamaglia, A. Molinaro, A. Iera, G. Interdonato, and F. Spano, "Optimizing Point-to-Multipoint Transmissions in High Speed Packet Access networks," in *IEEE BMSB*, 2011, pp. 1–5.

[18] G. Araniti, M. Condoluci, L. Militano, and A. Iera, "Adaptive Resource Allocation to Multicast Services in LTE Systems," *IEEE Transactions on Broadcasting*, vol. 59, no. 4, pp. 658–664, 2013.

[19] S. ul Zuhra, P. Chaporkar, and A. Karandikar, "Toward Optimal Grouping and Resource Allocation for Multicast Streaming in LTE," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 239–12 255, 2019.

[20] J. Chen, M. Chiang, J. Erman, G. Li, K. Ramakrishnan, and R. K. Sinha, "Fair and Optimal Resource Allocation for LTE Multicast (eMBMS): Group Partitioning and Dynamics," in *IEEE INFOCOM*, 2015, pp. 1266–1274.

[21] E. Haleplidis, K. Pentikousis, S. Denazis, J. H. Salim, D. Meyer, and O. Koufopavlou, "Software-defined Networking (SDN): Layers and Architecture terminology," in *RFC 7426*. IRTF, 2015.

[22] 3GPP TS 36.213 v15.2.0, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," 2018, [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2427.

[23] "MATLAB Toolbox," https://in.mathworks.com.

[24] 3GPP TR 36.931 v.15.0.0, " Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) requirements for LTE Pico Node B,," 2020.

[25] 3GPP RP-210739 v2, "WID on New Bands and Bandwidth Allocation for LTE based 5G Terrestrial Broadcast," 2021, [Online]. Available: https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_91e/Inbox/multiple_RP-210739_EBU/v2.