Abstract—The Third Generation Partnership Project (3GPP) Fifth Generation (5G) networks employ network slicing in order to implement multiple service types corresponding to different business needs. Network slicing results in the creation of multiple end-to-end logical networks over shared physical infrastructure. 3GPP mandates that slices are to be deployed homogeneously within a Registration Area (RA) for ensuring User Equipment (UE) slice mobility over the area. However, studies have shown that the deployment cost of the network increases with the increased number of slices. This makes it a costly proposition to deploy all the supported slices homogeneously over the RA. In this paper, we propose that the slice deployment requirements can be more flexible while ensuring UE slice mobility through the use of an enhanced multi-connectivity protocol. The proposed protocol can also be used for ensuring slice mobility across multi-Radio Access Technology (RAT) networks, including 5G New Radio (NR) and Wireless Local Area Networks (WLANs). The performance improvements obtained due to the proposal are demonstrated through simulations in network simulator-3 (ns-3).

I. INTRODUCTION

Fifth Generation (5G) networks are expected to support a multitude of services with diverse and sometimes, conflicting latency and throughput requirements. For example, it should be feasible to deploy services requiring high data rates, e.g., high definition streaming, very low latencies such as tactile internet as well as applications such as Internet of Things (IoT) requiring low data rates over the network [1]. This diversity of services is made realizable within the Third Generation Partnership Project (3GPP) 5G network through the use of network slicing [2].

A network slice has been defined by 3GPP as a logical network that provides specific network capabilities and network characteristics [3]. Every slice can be configured to support a different service and is identified using an identity known as the Single Network Slice Selection Assistance Information (S-NSSAI). The S-NSSAI specifies the slice behavior in terms of services and features as well as comprises information which helps in differentiating between multiple slices. A User Equipment (UE) can be connected simultaneously to multiple slices for accessing various services. Separate Protocol Data Unit (PDU) sessions are created for the UE for each of the connected slices. Deployment of a network slice requires two types of resources viz., link resources with constraints such as bandwidth, latency, packet loss, etc., and node resources include compute and storage resources over each of the network nodes [4]. Each of these resource types is to be chained in a specific manner based on the slice-types supported.

The 3GPP specifications allow for the deployment of specific slice types within a certain geographic area [5]. These slices may not be available everywhere in the network. For example, a real-time slice supporting factory automation may be required only within the premises of the factory and would not be required elsewhere. However, the specifications also prescribe that slices should be deployed in a homogeneous manner within a given Registration Area (RA) for ensuring UE slice mobility [5]. As a result, network slices may need to be deployed on every node within the UE RA even though they may be not used beyond certain geographic limits (as within the factory automation example). This would lead to an increase in the network Capital Expenditure (CAPEX) and Operating Expense (OPEX) as it has been shown through studies that network costs increase sharply with the increase in the number of slice instances over a given node in the network [6]. Moreover, the 3GPP slice deployment specifications place a lot of additional demands on the 5G base station (gNB) capabilities when an increased number of slices are to be deployed, due to the higher resource requirements.

Also, 3GPP 5G networks support multiple Radio Access Technologies (RATs) such as Wireless Local Area Networks (WLANs) and 5G New Radio (NR) connected to a common core network. Due to the difference in the nature of these RATs, the latency and throughput guarantees that each of these RATs can provide are different from one another. As a result, not all standardized slice types may be supported by all the RATs. Due to all the above reasons, it is important to allow network slices to be deployed in a more flexible manner.

In this paper, we propose a new protocol based on enhanced multi-connectivity for ensuring UE slice mobility and user association, allowing for flexible deployment of slices over the network. Multi-connectivity [7] is a protocol defined by 3GPP for reducing handover failures in a heterogeneous network.

Within this protocol, the UE is connected to two 3GPP Radio Access Network (RAN) nodes which have different coverage areas (e.g., macro gNBs and femto gNBs) and the data corresponding to a given PDU session is delivered to the UE by splitting it across the gNBs/direct transfer from the core network. However, at present, UEs cannot be multi-connected to nodes with similar coverage areas (e.g., two femto cells). Also, the protocol is undefined when the RAN nodes belong
to different RATs such as a 3GPP gNB and a WLAN AP. We propose a few enhancements in the multi-connectivity protocol that not only remedy the above-mentioned issues but also ensure UE slice mobility. The proposed protocol allows UE multi-connectivity to two or more similar node types (e.g., two femto/macro cells) as well as RAN nodes belonging to diverse RATs in addition to dual connectivity to HetNets supported by 3GPP. Unlike the 3GPP multi-connectivity protocol, wherein the data for a single PDU session within a slice is split across the RAN nodes, the proposed protocol supports individual PDU sessions corresponding to different slices on the individual data connections. This results in reduced UE slice handover failures when a particular slice is not supported on a target cell. All of the proposed ideas are evaluated through ns-3 simulations.

The rest of the paper is organized as follows. The details of existing works in this area are presented in Section II. The succeeding section presents an overview of UE handover considering slice-support as defined by the 3GPP 5G standard, followed by details of the proposed protocol in Section IV. Experimental evaluations of the proposed protocol is presented in Section V, followed by conclusions in Section VI.

II. RELATED WORK

Both, UE slice mobility management and multi-connectivity have been recently studied in the literature. Studies on multi-connectivity [8]–[10] mainly focus on reducing handovers in Heterogeneous Network (HetNet) scenarios while improving robustness by maintaining connections to multiple base stations during handover. Authors in [8] propose the usage of dual connectivity in 5G millimeter wave networks for improving handovers. The proposal tracks UE channel quality on multiple links and uses a local coordinator in close proximity to the cell to manage the UE traffic. It also illustrates that path switching using dual-connectivity is faster than a handover. In another work [9], authors aim to reduce handover signaling cost in dense networks using multi-connectivity. This is done by using a split control plane and data plane architecture and maintaining connections to multiple APs at a time so that the service remains uninterrupted. Authors in [10] focus on inter-macro base station handovers for 5G base networks in a high-speed railway scenario. The authors define a scheme to reduce handover failures through coordination between two macro base stations by proposing changes on the protocol stack and by replicating Radio Resource Control (RRC) signaling on both the base stations.

In [11], the authors exploit the property of localization of slices to certain areas within the network given that for certain types of slices, UEs may remain stationary. They propose protocols for slice mobility with the aim to reduce control signaling by eliminating the location tracking functionality for stationary UEs and reduce location update frequency and paging frequency. Some of the research on slice handovers [12]–[16] focuses on defining SDN based network architectures and illustrate the protocols for slice-based handovers within the defined architectural framework. Authors in [12] propose an architecture for a 5G system with a two-tiered RAN made up of macro and femto base stations, which share the spectrum. In this work, the authors also illustrate slice-based handover procedure within the architecture and propose a scheme for allocation of power and sub-channels for the network slice. UEs. In [13], the authors propose an architecture for slice-based mobility management for HetNets. They propose a scheme for offloading slices within the same slice across different RATs. [14] proposes an architecture based on SDN and Network Function Virtualization (NFV) for managing network slices and their associated resources dynamically. The authors propose to use different elements for handling intra-slice and inter-slice mobility. Authors in [15] propose a hierarchical control model for HetNets comprising a global controller and multiple local controllers. The global controller manages the slice life-cycle and also co-ordinates the handover for users. It queries the local controller for the availability of the target slice and creates it, if unavailable to ensure a seamless handover. In [17], authors propose the design of a flexible SDN/NFV based 5G architecture with a focus on network slicing. They analyze and differentiate scenarios where UE mobility management is required to be performed within and across slices and illustrate the same through call flows. [16] proposes a framework capable of offloading data traffic from the 3GPP network to dynamically created non-3GPP network slices. Although these works explore and provide solutions for slice mobility and multi-connectivity separately, they do not look at slice mobility and slice specific service delivery, cohesively. To the best of our knowledge, our work is the first work to propose the use of multi-connectivity as a protocol to facilitate UE slice mobility. In the next section, we provide a brief overview of the 3GPP-defined handovers considering slice-support for better illustrating the differences with the proposed protocol. Although in this paper, we have mostly discussed multi-connectivity in the context of UE slice mobility, it can also be used for slice specific user association and service delivery through multiple base stations for stationary users.

III. UE SLICE MOBILITY IN THE 3GPP 5G NETWORK

In this section, we summarize the UE slice handover procedure as defined by 3GPP. Figure 1 illustrates the call flow for UE slice mobility considering slice handover support as defined by 3GPP [18]. As illustrated in the figure, the UE supports ‘n’ slices with ‘m’ active PDU sessions. The active PDU sessions correspond to the number of slices that the UE is connected to at present, with one PDU session per slice. When the received signal strength of the serving cell at the UE falls below a given threshold, the source gNB tries to initiate a handover to a possible target cell. Note that different algorithms for target cell selection and different criteria (e.g., signal strength, support for slices) may be used by the network provider at the source gNB to select the target cell/gNB. However, after the target selection, the proposed 3GPP protocol ensures that the UE is handed over to a single target cell/gNB. As illustrated in Figure 1, handover is initiated
towards the chosen target cell (gNB) by sending a **HANDOVER REQUEST** message. The target gNB can only admit the PDU sessions mapped to the supported slices (S-NSSAIs), and the remaining sessions are rejected. The information of the admitted sessions is then sent to the source gNB using the **HANDOVER REQUEST ACKNOWLEDGE** message. Only the admitted sessions are then handed over to the target. The handover is followed by path switch request/response signaling to indicate the new endpoints for the General Packet Radio Service (GPRS) Tunnelling Protocol (GTP) tunnels as the connected gNB is now changed. Note that if the target cell does not support a particular slice, then the PDU sessions belonging to that slice cannot be handed over and leads to session discontinuity. This is especially true of practical deployments, as some of the slices are available only within certain areas of the network [18]. Also, even if all the slices (S-NSSAIs) are supported on the target gNB, handovers may also fail due to unavailability of resources for some (or all) of the slices. In order to mitigate this issue, we propose a protocol using multi-connectivity for UE handover and UE association. The details of this protocol are provided in the next section.

IV. PROTOCOLS FOR UE SLICE MOBILITY

In this section, we describe the multi-connectivity feature as defined by the 5G standard, followed by details of the proposed enhancements which would help in enabling UE handovers with slice-support.

A. Multi-connectivity in the 3GPP 5G Network

Multi-connectivity provides a protocol for multi-mode UEs to be connected to two heterogeneous RAN nodes such as 5G NR femto gNB and a 5G NR macro gNB at the same time. Traffic corresponding to a PDU session is sent/received over these multiple radio connections. Although, 3GPP defines four variants of connectivity based on the core network node and RAN node type, e.g., Evolved Universal Terrestrial Radio Access (E-UTRA)-NR dual connectivity for 4G-5G node dual-connectivity, NR-NR Dual Connectivity (NR-DC) for 5G-5G node dual-connectivity, for the sake of this paper, we limit our discussion to NR-DC.

Figure 2 illustrates the connections for a particular UE, which is dual connected to two 5G gNBs. As illustrated in the figure, the dual-connected system consists of two gNBs, one of which known as the Master Node (MN) connected to the core network through the NG interface. The MN connects to the UE through the air interface (Uu). The signaling to the core network may be transferred through the MN. The other gNB known as the Secondary Node (SN) is connected to the MN over the Xn interface. The MN behaves as the master for the UE connection and is responsible for setting up PDU sessions in the RAN. The SN is also connected to the UE through Uu interface, and it can send radio control related configurations to the UE using the signaling Radio Bearer 3 (SRB3) message. Also, as illustrated, each gNB consists of units known as the gNB-Centralized Unit (gNB-CU) and one or more Distributed Units (gNB-DUs) under the control of a single gNB-CU. Typically, multi-connectivity is used in HetNet deployments where the MN is a larger coverage macro-gNB, and the SN is a lower power gNB with a smaller coverage area that overlaps with the MN's coverage area. Also, although protocols such as Long Term Evolution (LTE) WLAN Aggregation (LWA) [19] and LTE WLAN integration over IPSec (LWIP) were defined for multi-connectivity to other RATs in previous releases, at present, multi-connectivity to 3GPP 5G NR and non-3GPP WLAN has not been defined.

B. Proposed multi-connectivity protocol

In this section, we provide details of the proposed multi-connectivity protocol. The multi-connectivity protocol proposed in this article also allows UEs to connect to two (or more) base stations through multi-connectivity. The base stations may have similar cell sizes such as two femto gNBs
or have differing cell sizes, such as a macro gNB and a femto gNB. They may even belong to multiple RATs such as a macro gNB and a WLAN AP. The procedure for connecting 3GPP NR nodes such as two femto gNBs is the same as that used for 3GPP NR multi-connectivity. Also, 3GPP has not defined multi-connectivity between different RAT types such as NR gNBs and WLAN APs. The proposed system diagram for enabling such a configuration is illustrated in Figure 3. The protocol used is based on LWA [19] defined for release 13 LTE. The 3GPP defines as Wireless Termination (WT) point for 3GPP connections towards the non-3GPP network, the placement of which is implementation-dependent. We suggest such that WT can be placed together with the Non 3GPP InterWorking Function (N3IWF), a function defined by 3GPP to connect WLAN APs to the 5G core network. This entity referred to as the ‘modified N3IWF’ connects to WLAN APs within the network through the Y2 interface. The gNB-CU and the N3IWF are now connected over the Xw interface through the WT. The procedures for multi-connectivity using this system are presented in Section IV-C.

C. Slice handover using multi-connectivity

As described in Section III, 3GPP slice handovers may result in loss of PDU sessions due to the absence of support for all active S-NSSAIs on a single target cell. This situation would occur if slices are flexibly deployed or when the target gNB is in a different RA from the source gNB. By employing multi-connectivity, we can enable multiple connections from a single UE to multiple gNBs, especially when the gNBs support different slices (S-NSSAIs). The signaling procedure for the same is illustrated for two scenarios - firstly, when the target RAN nodes are gNBs and second when the target nodes are WLAN APs.

1) Scenario 1: Slice handover within 5G NR: The handover procedure within 5G NR RAN is illustrated in Figure 4. When the UE is mobile, it may be handed over to a target cell in accordance with the network provider’s chosen criteria. If the target cell only supports a few of the slices (S-NSSAIs), we then initiate an additional connection towards another gNB. The gNB now behaves as an MN and performs the ‘SN Addition Request’ to add another gNB which supports the remaining S-NSSAIs as an SN. Once the SN is added and configured, data is forwarded to enable seamless handover. PDU sessions are appropriately handed over to the supporting MN/SN, and the forwarding paths are updated. The UE context is released at the end of this procedure. As it is not necessary that a single target gNB must be able to support all the active S-NSSAIs on the UE, there is a reduced risk of handover failures through multi-connectivity. Within the proposed scheme, as an increased number of cells can be chosen as target cells for handovers, it is easier to distribute the traffic load across the network, leading to better network utilization. The proposed protocol also provides a framework for the deployment of more sophisticated approaches for handover and user association considering factors such as load, interference, etc.

2) Scenario 2: Slice handover between 5G NR and non-3GPP RAN: A similar procedure can be followed for handover within nodes comprising multiple RATs, when one of the target nodes may be WLAN AP.

The call flow for UE handover with multi-connectivity for non-3GPP networks is illustrated in Figure 5. As illustrated, a modified version of the N3IWF is added as an SN over the WT interface using ‘WT Addition Request’ message. Once the addition of SN is complete, the rest of the procedure is identical to that of the handover within 5G NR. The proposed procedure enables handovers with slice support in multi-RAT networks comprising 3GPP NR and non-3GPP WLANs. The proposed protocol allows for multi-RAT handovers within the RAN. This functionality is presently undefined by the 3GPP 5G standard, which only defines handovers between multi-RAT nodes through the 5G core. We demonstrate the performance of the proposed protocols through simulations in the succeeding section.
V. EXPERIMENTAL EVALUATION OF THE PROPOSED PROTOCOL

In order to evaluate the performance of the proposed handover protocol, we simulate a 7 cell cluster in ns-3. We use LTE as the reference RAT in the simulations, as to the best of our knowledge, an open-source simulator for 5G NR RAT is unavailable at present. Although an independent 5G millimeter wave simulator is available, it does not yet support the entire 5G NR protocol stack. The result for slice handover in LTE will be similar to that in 5G NR as the handover behavior of both these RATs are similar. This cluster, as illustrated in Figure 6 consists of LTE base stations (eNBs) placed in the center of the cell. As within practical deployments, different slices are supported within different areas of the network, and the supported slices are indicated in varying colors. We consider 3 User Datagram Protocol (UDP) applications corresponding to different slices, each with a data rate of 10Mbps. We carry out simulations for 10 different slice distribution configurations where slices are flexibly distributed within the network. The simulation parameters are provided in Table 1 and have been obtained from [20].

We assume that users are randomly distributed within the system and use the random wave point model to characterize user mobility. The users move with a maximum velocity of 40km/hr. We also assume that the user arrivals follow a Poisson distribution with an arrival rate of \( \lambda \) arrivals per second having a service rate of \( \mu \) within the system. The users experience log-normal path loss. For the multi-connectivity scenario, whenever a UE is handed over to a target cell without support for a given slice, we create another connection to a suitable secondary eNB. This secondary connection is torn down if the UE is handed over to another cell, where support for the slice that was present on the cell supported by secondary WLAN AP/gNB becomes available. The secondary eNB is chosen such that it has a suitable Reference Signal Received Power (RSRP) value. It should also support the slice that is required by the UE but unsupported by the target Master eNB. We evaluate the following system metrics for both 3GPP-defined and proposed handover protocols.

1) Rate of handover failure:: We measure the rate of handover failure for various user arrival rates (\( \lambda \)) for both the protocols. The service rate (\( \mu \)) is maintained at 1 user per second. We measure the handover failure rates for single connected users and also for multi-connected users for various slice configurations. The results are illustrated in Figure 7.

As shown in the figure, the rate of handover failure is reduced with multi-connectivity with an average improvement of 13.53% for the configurations considered in our simulations.

2) System throughput: We also evaluate the system throughput for the various slice configurations by varying the user arrival rate. The median system throughput for the 10 different configurations with single connected users and multi-connected users have been depicted in Figure 8. The confidence intervals across the 10 configurations have been plotted on the same graph with a line. As shown by the results, the system throughput in a multi-connected system is higher than that of the single-connected system. This is because PDU session continuity is maintained on handover by connecting to a secondary eNB if the target eNB does not support one or more slices present on the source eNB. Other than maintaining service continuity, we can observe that the resources from multiple eNBs are utilized for the sessions, thus improving resource utilization.
VI. CONCLUSIONS

In this paper, we have proposed that network slice deployment needs to be more flexible to minimize network costs and have provided an overview of the existing UE slice mobility protocols. We have also proposed a protocol based on UE multi-connectivity to enable handovers for slices that are unsupported on the target cell. The proposed protocol extends the usage of multi-connectivity and allows for connection to RAN nodes with similar coverage. It also addresses the gap within the 3GPP standard where multi-connectivity to nodes of 3GPP and non-3GPP RATs are presently undefined. The proposed protocol also reduces handover and user association failures due to lack of slice support on the target cell in comparison to the 3GPP standard protocols. The performance improvements obtained through the proposed protocol have been demonstrated through ns-3 simulations. This proposal can be easily integrated with the 3GPP 5G network due to its compatibility with the defined network framework. In future, we intend to study optimal network-slice deployment algorithms for better network utilization.

VII. ACKNOWLEDGEMENT

This work has been supported by the Department of Telecommunications, Ministry of Communications, India as part of the indigenous 5G Test Bed project.

REFERENCES

[18] 3GPP TS 38.300 v15.5.0, “NR; Overall description; Stage-2,” 2019.