IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 18, 2020, accepted September 3, 2020, date of publication October 13, 2020, date of current version November 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3030762

A Comprehensive Survey on Mobility Management in 5G Heterogeneous Networks: Architectures, Challenges and Solutions

EMRE GURES[1], IBRAHEEM SHAYEA[1], (Member, IEEE), ABDULRAQEB ALHAMMADI[2], (Graduate Student Member, IEEE), MUSTAFA ERGEN[1], AND HAFIZAL MOHAMAD[3], (Senior Member, IEEE)

[1]Electronics and Communication Engineering Department, Faculty of Electrical and Electronics Engineering, Istanbul Technical University (ITU), 34467 Istanbul, Turkey
[2]Center for Wireless Technology, Faculty of Engineering, Multimedia University, Cyberjaya 63100, Malaysia
[3]Faculty of Engineering and Built Environment, Universiti Sains Islam Malaysia, Bandar Baru Nilai 71800, Malaysia

Corresponding author: Emre Gures (gurese18@itu.edu.tr)

This work was supported in part by the 2232 International Fellowship for Outstanding Researchers Program of TÜBİTAK, conducted at Istanbul Technical University (ITU), under Project 118C276, and in part by the Universiti Sains Islam Malaysia (USIM), Malaysia.

ABSTRACT With the rapid increase in the number of mobile users, wireless access technologies are evolving to provide mobile users with high data rates and support new applications that include both human and machine-type communications. Heterogeneous networks (HetNets), created by the joint installation of macro cells and a large number of densely deployed small cells, are considered an important solution to deal with the increasing network capacity demands and provide high coverage to wireless users in future fifth generation (5G) wireless networks. Due to the increasing complexity of network topology in 5G HetNets with the integration of many different base station types, in 5G architecture mobility management has many challenges. Intense deployment of small cells, along with many advantages it provides, brings important mobility management problems such as frequent handover (HO), HO failure, HO delays, ping-pong HO and high energy consumption which will result in lower user experience and heavy signal loads. In this paper, we provide a comprehensive study on the mobility management in 5G HetNet in terms of radio resource control, the initial access and registration procedure of the user equipment (UE) to the network, the paging procedure that provides the location of the UE within the network, connected mode mobility management schemes, beam level mobility and beam management. Besides, this paper addresses the challenges and suggest possible solutions for the 5G mobility management.

INDEX TERMS Mobility management, handover, heterogeneous networks, 5G network.

I. INTRODUCTION

The Ericsson mobility report predicts that the number of cellular broadband subscribers, which was 4.4 billion in 2016, will increase to 8.3 billion in 2022 [1]. To cope with the ever-increasing demand and service requirement, small cells such as pico and femto base stations (BSs) are deployed to the network of macrocells in the same geographic region, forming HetNet. HetNets have the potential to scale the system capacity to users significantly and to provide uninterrupted high-rate communication services to users with more reliably. Therefore, if users switch from one cell to another, the network must transfer a radio link to the new cell to ensure the continuity of the communication service. One of the major challenges with 5G HetNet is mobility management.

Deploying a large number of small cells into a HetNet leads to increase cell edges and greater inter-cellular interaction. This causes frequent HO events and more radio link failures (RLF) [2]. However, if a user connects to a neighboring cell for a short time and then bounces back the former source cell, this event is called ping-pong HO (PPHO). PPHO is undesirable because it causes unnecessary power consumption. Additional signal load occurs due to mutual signaling

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

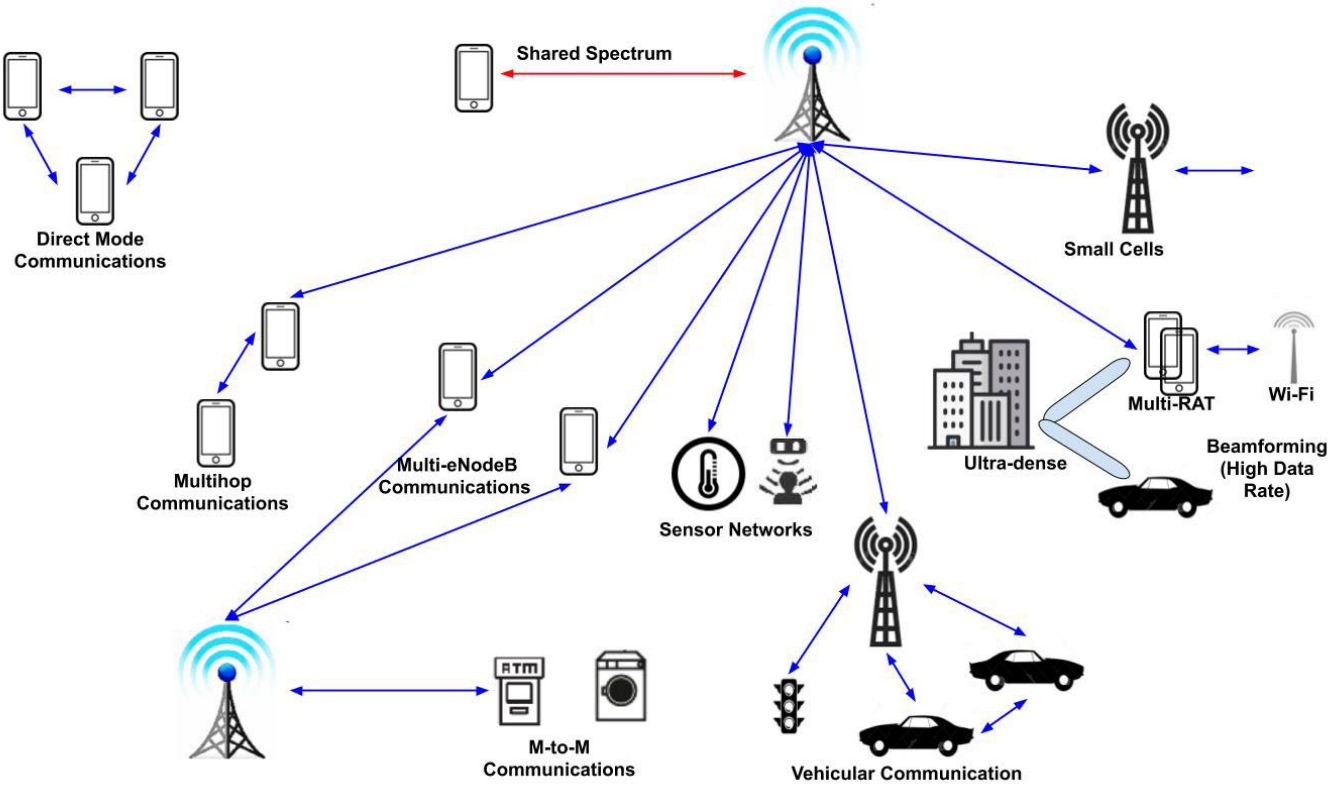

FIGURE 1. General 5G cellular network architecture and the different services supported.

during HO failure (HOF) and PPHO, which inefficiently uses the resources reserved for data transfer while shortening the battery life of the user equipment.

We believe that mobility management in 5G systems will support additional features that enhance the user experience and improve new use cases that will be available in the future. Figure 1 shows a general 5G cellular network architecture and the different services supported. The basic requirements of mobility management in 5G systems are that HO operation can be performed without interruption, it has configurable HO security and link monitoring mechanism. Besides, the mobility management has to create a service-specific mobility configuration by providing mobility support for multi-slicing networks. Network slicing is a form of virtual network architecture and allows for flexible and efficient use of limited resources by creating multiple virtual networks on top of a physical infrastructure. Virtual networks are then customized to fulfill the specific needs of applications, services, devices, customers or operators. Thus, rather than using complex and expensive hardware resources that are not used at full capacity, they directly address devices that require special functionality and provide savings and efficiency. For example, different network slices are allocated to internet of things (IoT) and machine-to-machine (M2M) services, which prioritize low-cost connectivity rather than reliability, because they can tolerate temporary interruption of networks due to their asynchronous communication with high-reliability IP telephony and internet access services. However, the paging process used to determine the location of the UE, the registration process which enables the UE to connect to the network to benefit from the services and applications and mutual signaling in the network during HO processing, cause latency and power consumption in the network. For this reason, it is essential to have a configurable balance between power saving and latency in 5G mobility management. In addition, radio resource control (RRC) is also an important role that controls the communication between UEs and BS in the highest layer of the control plane. The idle state and connected state of RRC are introduced in long-term evaluation (LTE), whereas inactive state presented recently in 5G. The RRC Inactive state is a new state between RRC Idle and RRC Connected states, and its purpose is to improve both latency and power consumption performance.

Extensive studies have been conducted in the literature to overcome management problems in HetNets. The main reason for performing HO from one cell to another is users mobility. In [3], proposed a method to predict the user velocity from the number of HOs in HetNets by using a minimum variance-neutral technique. In [4], the authors proposed a

frequent HO mitigation algorithm to reduce overall HOs and increase network efficiency for ultra-dense HetNets by grouping users with frequent HO experiences as fast-moving users and ping pong users. In this algorithm, fast-moving users are associated with macro BSs, while ping pong users try to optimize their HO parameters so that the effects of such unnecessary transfers are reduced. If ping pong effects cannot be avoided by adjusting the HO parameters, these users are associated with macro BSs. In [5], the authors are tried to optimize cross-tier HO in terms of delay characteristics for HetNet. In [6], proposed a network framework to seamlessly switch between two neighboring macro evolved NodeBs (eNB) in the 5G control/user plane split. The proposed framework is enable uninterrupted transfer by integrating of the HO assisted micro eNB in the overlapping region with a DC communication. In [7], a method is proposed to achieve the weight of the HO metrics using the analytical hierarchy processing (AHP) technique and then sorts the cells to select the best HO target with the grey rational analysis (GRA) method. Thus, the HO rate is minimized and RLF is reduced. In [8], the authors examined the effects of channel fading on mobility management in HetNet. The results show that increasing the sampling period of the HO decision, decreases the fading effect while increasing the ping-pong effect. The work in [9] presented mobility sensitive user association rule. The rule tries to overcome congestion by directing UEs to small BS while monitoring dynamic changes in channel conditions that caused by user mobility in the network topology. Thus, it avoids frequent HO and PPHOs between small BS. However, it takes into account the specific aspects of millimeter-wave (mmWave) communication, such as its specific directional, susceptibility to clogging, and non-line-of-sight propagation propagation effects, and the UE is distributed within the network accordingly. The mmWaves tend to be blocked by objects due to their high-frequency structure. To overcome this problem, in [10], a framework is proposed that predicts obstacle-caused data rate degradation before the degradation occurs by expanding the status area thanks to the successive camera images over time. Using a deep reinforcement learning to decide HO timings, challenges in addressing large dimensional were overcome.

A comparison table of the literature survey is depicted in Table 1. Descriptive information about the surveys in the literature is given in the prepared table. Besides, the features of the presented survey were specified and compared with other studies. In this paper, the RRC states in the highest layer of the control plane are introduced. Then, various registration procedures were mentioned and their usage and working principles were explained. Here, the balance between power savings and control signal load associated with the recording procedure is mentioned. The paging procedure used in 5G to find the location of the UE on the network is described by comparison with its version in LTE. Two trade-offs encountered during the paging procedure are mentioned; these are the trade-off between power and delay caused by the size of the discontinuous reception (DRX) cycle and the number of cells to be paged and power saving. Then, in order to provide an uninterrupted service to UEs, its association with the cell that will provide the best performance was examined for different HO scenarios. The advantages of using mmWave systems in 5G HetNets were mentioned and how these systems will be integrated, the management of beam level mobility and the measurement parameters used for this purpose are explained in detail. Afterwards, various difficulties such as HO problems, signal overhead, power consumption, security and delay are described in detail, along with the solution suggestions presented in the literature. Finally, effective solution methods such as software-defined networks (SDN), conditional HO (CHO), and dual connectivity (DC) to meet the fundamental requirements of 5G mobility management and their application in the literature are mentioned.

This paper is organized as follows. Section II presents the types of RCC states and their capabilities in LTE/5G networks. Registration and paging of the UE are discussed in Section III and Section IV, respectively. Section V explains the connected mode mobility and HO management. The beam level mobility and beam management are discussed in Section VI. Challenges in mobility management are addressed in Section VII. Section VIII provides some suggested solutions for the mobility management issues. Finally, the paper concludes in Section IX.

II. RRC STATES

Design standardization of the control plane is still in the developing process in 5G. The state machine design for the 5G radio access network (RAN) contains numerous divergent and partially contradictory 5G use cases. For this reason, the mobility framework consisting of the state machine design is one of the important issues of the control plane [22]. The RRC layer processes the control plane signals between the UE and RAN on the radio interface and controls the UE's transition mobility to the cells. The RRC is the highest layer in the control plane of the access stratum (AS). Its main tasks are broadcasting system information, establishing or releasing RRC connections, paging, transferring non-access stratum (NAS) messages used to control communication between the UE and core network (CN) [23].

The state machine of universal mobile telecommunication system (UMTS) technology consists of a total of 5 states, one idle state, and four connected states (CELL_DCH, CELL_FACH, CELL_PCH and URA_PCH). The idle state is optimized for low power and network resource consumption, whereas connected states are optimized for high UE efficiency and fast connection re-establishment. The main distinction between the 4 connected states is that the power consumption levels are different, and so many situations complicate the structure.

TABLE 1. Literature survey comparison.

Survey	Year	Description	Survey Platform
[11]	2011	• Mobility management solutions for vehicle networks are classified and revised based on vehicle to vehicle and vehicle to infrastructure communications. • Conformity of traditional mobility management techniques to vehicle networks is discussed. • Several open research topics in mobility management for vehicle networks are outlined.	Vehicular ad hoc networks (VANET)
[12]	2012	• In this study, a comprehensive literature review about the mobility management architectures required for seamless HO of UEs in HetNet is presented. • An analysis was made by examining the main objectives, assumptions and requirements of the architectures selected from the literature. Thus, instructions are provided containing requirements and features needed for future architectures. • A new architecture called Context Aware Mobility Management System was presented and basic functional assets that should be a part of future architectures were determined.	HetNet
[13]	2012	• Mobility management of fourth generation (4G) HetNets are introduced and the challenging issues are discussed.	4G/HetNets
[14]	2013	• A comprehensive discussion is presented on the key aspects of mobility management support, research challenges and solutions for the two-layer macrocell-femtocell LTE-A system. • Based on the discussion for the HO decision phase, existing HO decision algorithms for the macrocell-femtocell network were investigated based on the primary HO decision criteria used in the literature. • By classifying these algorithms, the main advantages and disadvantages of the most representative algorithms per class are discussed.	LTE-Advanced
[15]	2015	• A comprehensive summary of mobility models commonly used in WSNs. • Sink mobility management plans were reviewed from an evolutionary perspective. • By analyzing the fundamentals and flaws of the existing solutions, several open problems that have not been discovered so far have been identified.	WSN
[16]	2016	• Mobility management techniques in VANET were examined for vehicle to infrastructure, vehicle to vehicle and hybrid vehicle communication modes. • Some technical challenges in mobility management in VANET were discussed.	VANET
[17]	2016	• Techniques and challenges for integrating IP over WSNs were described. • It also provided an overview of the IPv6 protocol and thoroughly examined the algorithms developed to address the features of mobility management within IPv4 and IPv6 • A comparison between IPv4 and IPv6 has been made.	IP-enabled wireless sensor network (IP-WSN)
[18]	2019	• State-of-the-art research addressing user mobility in a MEC environment for different types of services has been studied. • Also, open research problems related to mobility management are discussed in these systems.	MEC-enabled Systems
[19]	2019	• LTE HO management and 5G were examined in a general framework. • General concepts and possible difficulties related to radio access mobility in cellular networks were explained.	LTE/5G
[20]	2020	• In order to reduce the mobility management overhead, tracking area update (TAU) and paging procedure were examined in terms of complexity, latency, and computational cost and various solution schemes were proposed.	LTE
[21]	2020	• Considering the new architectural changes introduced by SDN, network functions virtualization (NFV), multi-access edge computing (MEC), specific approaches to vertical migration in 5G are explained. •Adressing the requirements of different vertical use cases, the evolutionary steps of mobility management based on new architectural elements are discussed in terms of efficiency, latency and scalability.	5G/HetNet
This survey		• We introduced general principles of 5G mobility management, such as RRC states, initial access and registration procedures, and reachability, and highlighted current research focuses and major challenges in these areas. • We have defined and categorized inter-RAN HO procedures for connected state, and simplified the description of the procedures to provide a general understanding of how each procedure works. We explained the beam management categories and architectures for mmWave cellular systems, and highlighted the key points of beam level mobility that should be considered. • We explained in detail the challenges encountered in 5G mobility management and the solutions that can be effective in this regard.	5G/HetNet

The fact that there are four connected states in UMTS and the majority of their features overlap has complicated the implementation of RRC states. In the LTE system, the number of RRC states to two, RRC Idle and RRC Connected, to eliminate this complexity. If there is no RRC connection between the UE and the eNB, the UE is in the RRC Idle state, else if there is an RRC connection between them, the UE is in the RRC Connected state. The RRC Idle state implies to

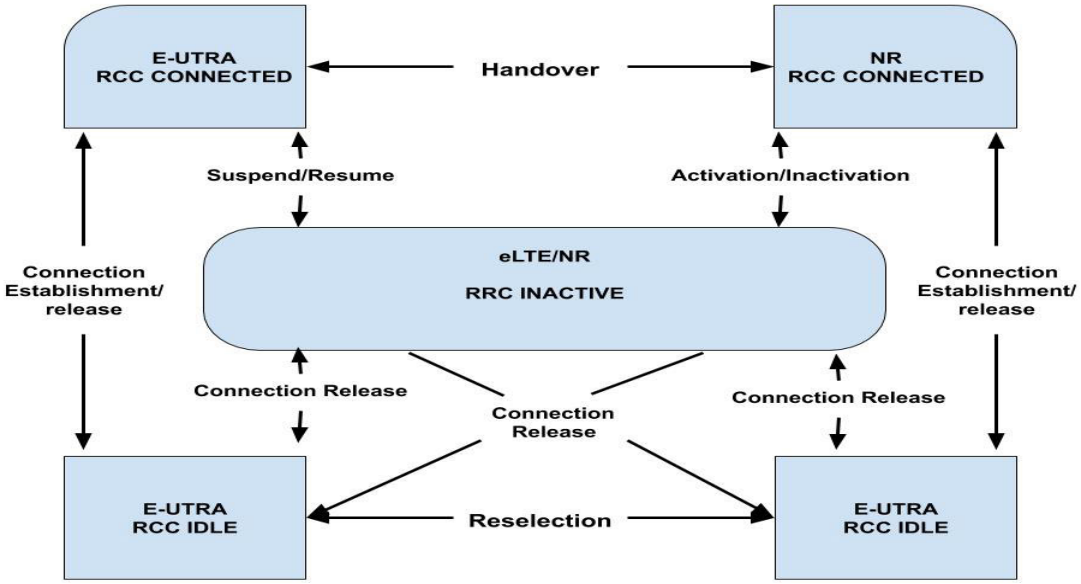

FIGURE 2. RRC state diagram in 5G [24].

optimize power consumption and network consumption. The UE's AS context is not stored in the RRC Idle state, so when the UE switches to the RRC Idle state, this context stored in both the network and the UE is deleted. This is due to the transitions between RRC Idle and RRC Connected states require a high signal procedure, where they are insufficient for the sparse transmission of small packets. When the UE is in the RRC Idle state, it performs cell selection and cell re-selection based on the parameters provided by the RRC. On the other hand, the UE is in the RRC Connected state, it maintains its connection to the eNB and can use resources on the network with its wireless interface and also perform mobility under the network control.

In the 5G state machine design, there are two inferences are made from the problems encountered in UMTS and LTE. The first inference is the storage of the UE AS context in the UE and the network during low activity, reducing the delay in signal transitions and the signaling load. Thus, low activity status should be added to the new 5G technology. The second inference is interface re-establishment between RAN and CN that increases the control plane delay. Therefore, the new low activity status to be added must maintain this connection.

The state machine of 5G consists of the newly introduced RRC Inactive state, the RRC Idle, and RRC Connected states in LTE. Figure 2 shows the RRC state diagram in 5G.

For a UE to establish RRC connection, it must be in either RRC Inactive or RRC Connected, otherwise the UE is in the RRC Idle state. The features and capabilities of the newly added RRC Inactive state will be discussed in detail below.

A. RRC IDLE STATE

In the case of RRC Idle, the UE AS context stores neither in the UE nor in the network. The use of this state should be limited to some cases, such as boot state and error recovery state only when the power is on. Since the UE has no RRC connection in the RRC Idle state, the BSs have no content about the UE, which in most cases leads to the location of the UE being known by the network at the level of the tracking areas (TAs). The main purpose of the RC Idle state is to save power. In the absence of data to be received or transmitted, the UE switches to the RRC Idle state and reduces battery consumption by switching off its receiver and transmitter, except for periodic paging reception. In the case of RRC Idle, the UE periodically monitors the call channel, checks incoming cases ad selects the cell to camp based on the mobility measurements. In case of RRC Idle, the UE's camping in a cell has purposes such as receiving system information for the camped cell, initiating an RRC connection setup on the camped cell when needed, receiving call messages for mobile termination calls in the camped cell, and receiving public warning system notifications [25].

B. RRC CONNECTED STATE

In the case of RRC Connected, the UE AS content stores in both the network and the UE, which optimizes for high UE activity. In the case of RRC Connected, the UE can transmit and receive user plane data and control plane signaling. Moreover, this is the case that consumes the most battery since the UE needs to constantly monitor the link quality of the serving cell and its neighbors and periodically give feedback to the information of the radio link. Therefore, DRX is used to save power when there is no continuous data transmission. In DRX, the UE is RRC Connected and goes into a state that the timer will use less power when it expires.

C. RRC INACTIVE STATE

In LTE technology, if the UE remains permanently in the RRC Connected state, it avoids connection delays, but

unnecessarily has to monitor the control channel and report measurement, resulting in increased power consumed by the UE. To ensure power efficiency, the UE must switch from RRC Connected to RRC Idle. The UE must release the CN/RAN connection and delete the UE AS contexts with this transition. The UE will need to get the UE AS context from CN when it switches back to the RRC Connected state. In the case of the newly introduced RRC Inactive with 5G, the UE AS context is stored in both the UE and the network. RAN/CN connection is kept active to minimize control plane delay and UE power consumption. The latency performance resulting from RRC state transitions improves with the reduction of the CN signal count with the integration of the RRC RRC Inactive state compared to RRC Idle. Moreover, the network with AS content knows the location of the UE in a preset area and the UE can move within that network without needing to notify the network. This results in lower UE power and network resource consumption compared to network controlled mobility. Since the CN link is kept active and the UE AS content is stored in RAN, the number of CN signals required to paging a UE is expected to decrease.

RRC Inactive state configurability promises to serve UEs with different usage needs with a single flexible low activity state instead of having multiple low activity states optimized for each use situation, thus enabling new services to enter the market quickly. The behavior of a UE with RC Inactive status can be configured over the cycle length of a standardized mechanism such as DRX, and low activity states of various 5G devices and UEs can be managed without increasing the number of RRC states.

In performance analysis for small packet transmission in [22], it has been observed that the RRC inactive state has a superiority over 71%, 88%, and 79%, respectively, in terms of overhead, delay, and power consumption signals compared to LTE Idle state. In the study on 5G networks containing machine-type communication in [26], it appears that adopting an optimal configuration of the RC Inactive state saves more than 200% latency reduction compared to the RRC Idle state and approximately 40% UE power savings compared to RRC Connected.

III. INITIAL ACCESS AND REGISTRATION

When the UE becomes active to take advantage of network services and capabilities, it needs to be registered on the network. This ensures that the UE is monitored on the network and becomes reachable. Several conditions can initiate the registration procedure of the UE, and these can be listed as initial registration, periodic registration, mobility registration, and emergency registration [27]. Initial registration is performed by the UE to connect to a network after the device is turned on. Periodic registration is the process by which the network checks the UE periodically to perform a new registration so that the UE in the registration area (RA) can be sure whether its registration is deleted without notifying the network. Mobility registration is the registration by the UE when the user changes the location and the TA of the cell to which it is linked is not in the RA list. Emergency registration is used by the UE only when it wishes to register for emergency services. Here, it would be useful to explain what the expressions of the registration area and the tracking area. When the UE location is changed, the location update procedure is triggered, and the network sends a paging message to the possible BSs where the UE can be found and according to the responses, it determines the location of the UE and tries to connect with it [28]. Cells that do not intersect with each other are grouped to form the structure called TA and each TA has an identity. TAs also group and form RAs. For instance, the UE is connected to cell 1 in TA1 as shown in Figure 3.

When the UE switches to a different TA region within the same RA, for example, TA2, it decides that it does not leave the RA region by comparing it with the identity broadcasted periodically from the BSs. For example, if the UE made contact with cell 7 in the TA 4 region, which is on the different RA list, the location update procedure would be triggered after comparing their identities. As a result, it notifies the network that the UE has changed its location, and the network appoints the RA of the cell it is into the UE, and after that the UE uses this RA. The network will periodically update the registry within certain periods to find out whether the UEs have left the RA zones or reaches information whether it is out of use without informing. There is a trade-off between paging message density and registration update concerning the size of the TA list. If the size of the TA list expands, it is imperative to send a paging message to more cells and the signal load in the system and the amount of power consumed accordingly will also increase. If the size of the TA list is small, it is necessary to update it more frequently.

The fundamental of the registration procedure is based on the transmission of control messages between the UE, next-generation NodeB (gNB) and access and mobility management function (AMF). After the UE is turned on, the UE needs to select the cell and establish the RRC connection with the gNB, where there are input access functions that enable synchronization with the target cell and should determine the network/public land mobile network (PLMN) to which it will connect. In NAS signal exchange that occurs between the UE and AMF during the registration procedure, the signal sent by the UE is encapsulated and sent to the gNB thanks to the RRC protocol. Following this process, the signal is transmitted to AMF thanks to the next generation-application protocol (NG-AP) in gNB [29]. As a result of all these transactions, UE's are registered to the 5G services.

IV. REACHABILITY

Paging is a system access function used by the network to locate the UE and is triggered when there is a downlink packet for the UE. In LTE, the location of the UE is determined by the call message sent by the mobility management entity (MME) to the BSs in the tracking area list (TAL) when the UE is in the Idle state. The BSs to which the UE is connected responds to the call message and reports the location of the

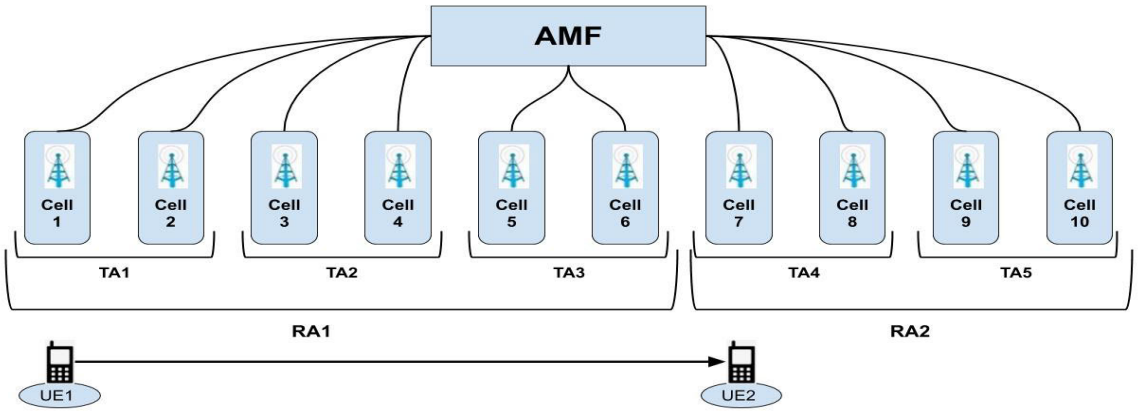

FIGURE 3. 5G mobility management architecture.

UE to the network. The paging procedure with 5G technology has become more severe than LTE with increasing BS. As a result, an increase in both signaling overhead and battery consumption is predicted. To reduce the effects of these facts, paging and location tracking schemes should be planned well. However, the increase in signaling load in the paging procedure does not match with the latency target of 5G. In 5G, the UE can be paged by both CN and RAN. When the UE is in the RRC Idle state, the paging procedure is performed by AMF, and in the worst-case scenario, the AMF will need to paging the entire network if the UE is not known in which RA. In case the UE is RRC Inactive, it can be called by both CN and RAN. The UE monitors the call status of the paging channel. The UE follows the paging channel (PCH) to find out if a paging message has arrived. A single call procedure can be used to simplify the system, as two different call procedures will cause complexity [30]. To limit power consumption, the UE periodically wakes up at certain times, called DRX cycles, instead of constantly monitoring the PCH. Depending on the size of the DRX cycle, a trade-off occurs between power consumption and latency. If the DRX cycle is short, the power consumption will be low. Otherwise, the latency will be small.

V. CONNECTED MODE MOBILITY

A UE aims to be associated with the BS that will provide it with the best performance in the network. For this purpose, it generates measurement reports periodically or upon the request of the network via reference signals received from neighboring BSs and service BS. Since each BS does not operate on the same carrier frequency, measurement gaps are needed where the UE adjusts the carrier frequency with respect to neighboring BSs. Measurement gaps refer to the time interval in which no uplink or downlink transmission occurs. The HO procedure is initiated when the measured value is above a predetermined threshold for a specified time to trigger (TTT). In future studies, it is predicted that more efficient HO structures will be achieved with the reduction of HOF, along with algorithms where the TTT value can be dynamically adjusted according to the speed of the users. The general 5G architecture is shown in Figure 4. Interfaces between gNBs are called XN nodes, while the connection between gNBs and 5G core (5GC) is called the N2 node. The HO procedure is divided into an XN-based HO procedure and an N2-based HO procedure, depending on which node the transaction is managed through. Figure 5 shows the 5G mobility architecture formed according to the relevant interfaces and HO use cases.

A. XN-BASED INTER-gNB HO

In this HO procedure, operation is managed over the X2 node with different target and service gNB in different cells. This procedure can only be performed when the target and service gNBs are connected via the same AMF (intra-AMF) via N2 interface. The flow chart is shown in Figure 6.

1. It is the preparation phase of the HO procedure. Serving according to the reports sent by the UE, the gNB decides to start the HO procedure and sends a HO request message to the target gNB. This message contains some information necessary for the target gNB to prepare itself. If the target has sufficient resources in the gNB to service the UE, it initiates the admission control procedure and sends the "HO request acknowledgment" message to the gNB that it has accepted the HO procedure. Source gNB transmits the HO command message to the UE. Source gNB sends SN status message to target gNB and routes data it receives from CN to target gNB. With the SN status message, it is tracked whether the received data is sent in the correct order, whether the data is duplicated, and contains information about how small blocks of data can be combined and restored to their original state.

2. The target gNB sends the message "N2 Path Switch Request" to the AMF and transmits the information that the UE has changed its cell. Along with this message, the target gNB also provides a list of protocol data unit (PDU) sessions

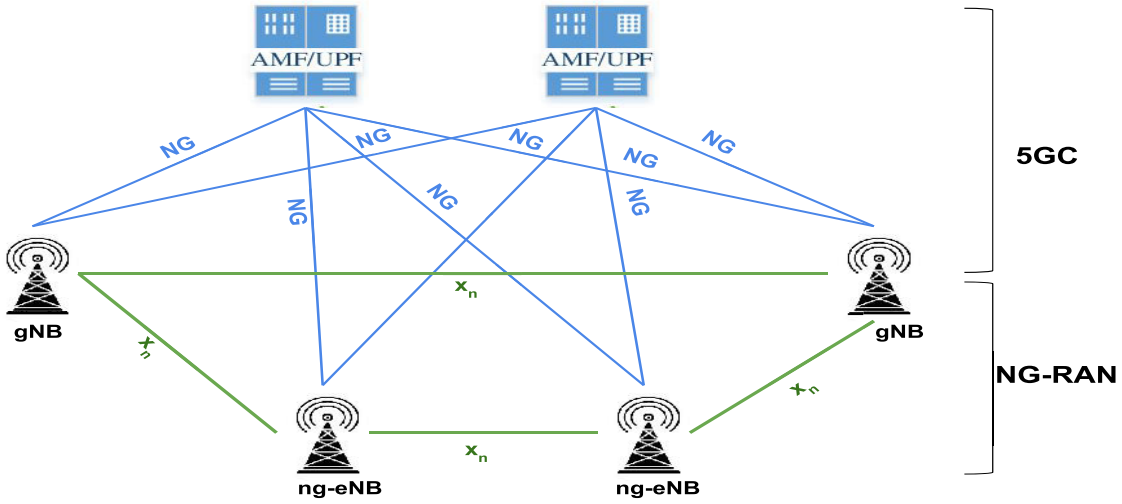

FIGURE 4. Overall architecture of 5G [31].

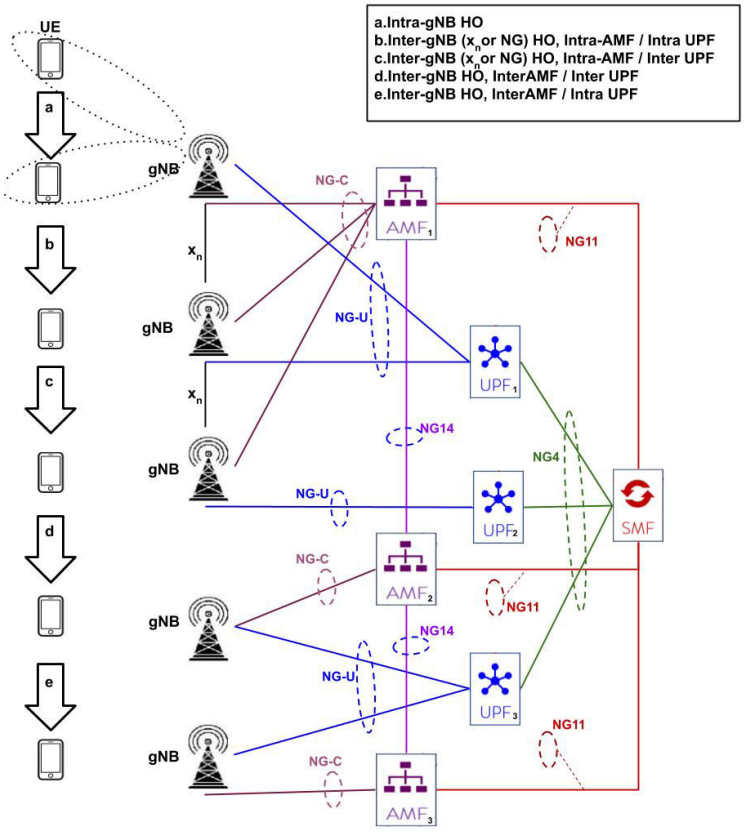

FIGURE 5. 5G mobility architecture based on relevant interfaces and HO use cases [19].

that need to be changed to the AMF. However, if there are not enough user plane resources in the target gNB to meet the requirements of the user plane security application, the corresponding PDU sessions are rejected.

3. The Nsmf_PDUsession service runs on PDU Sessions, and the service operations exposed by this network function allow other network functions to create, modify and broadcast PDU Sessions. The Nsmf_PDUSession_UpdateSMContext Request message is forwarded to the session management function (SMF) by the AMF. AMF sends N2 SM information by paging this request for each PDU session in the list of PDU sessions received in the previous step. With the

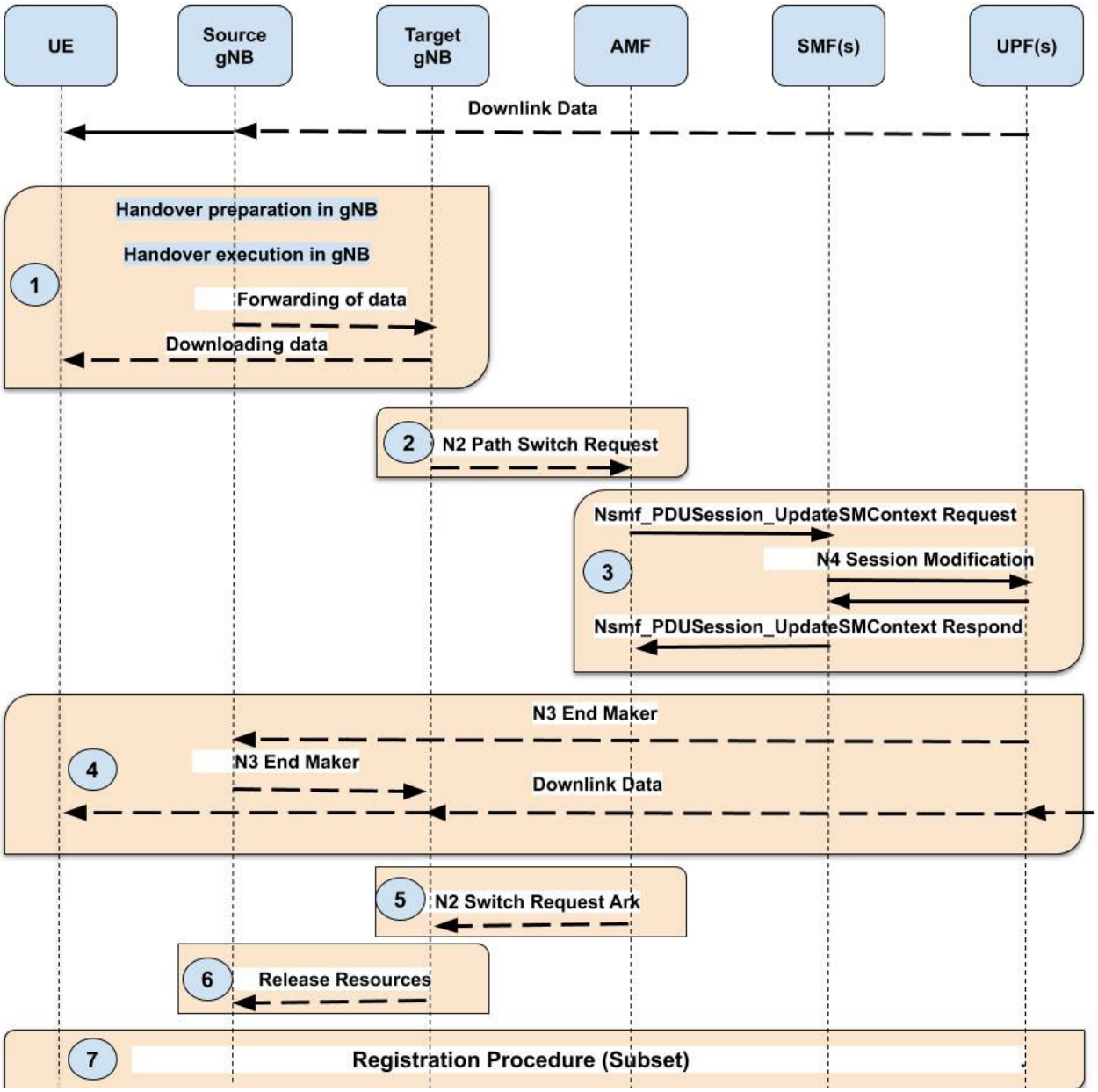

FIGURE 6. X2-based inter-gNB HO [32].

Nsmf_PDUSession_UpdateSMContext Request, either the PDU sessions are changed or the PDU sessions are rejected and then released. Thus, it is provided that downlink data is sent to the target gNB. For PDU sessions modified according to the target gNB, the user plane function (UPF) sends an N4 SessionModification Request by the SMF. UPF sends SessionModification Response message to SMF. As a result of these downlink data paths are updated. SMF then sends the message ''Nsmf_PDUSession_UpdateSMContext'' to the AMF that the PDU sessions have been successfully changed.

4. Thus, future downlink packages will be sent to the target gNB. One or more ''end markers'' will be sent instantly to the source gNB by the UPF after the change of path. Thus, it is understood that the N3 tunnels between the UPF and the source gNB have been removed.

5. AMF sends the ''N2 Path Switch Request Acknowledgment'' message to target gNB after acquiring CN tunnel information.

6. At this stage, the target gNB contacts with the source gNB to report that the HO procedure is completed. The source gNB releases the resource it allocated to the UE.

7. In some cases, the UE may need to initiate the registration procedure after the HO has completed. An example of this is when the UE goes out of RA.

B. N2-BASED INTER-gNB HO

The N2-based HO procedure is applied when there is no X2 connection between the source gNB and the target gNB due to the spread of the network. This HO procedure is carried out via CN. In the preparation phase, which is the first phase of this two-step HO procedure, it is aimed to prepare gNB and SMF/UPF in order to ensure the implementation phase of HO in a fast and uninterrupted manner.

1) PREPARATION PHASE

The flow chart of the preparation phase is shown in Figure 7. Several steps involved in the preparation phase. These steps are as follows:

1. Due to the communication conditions or load balancing, source gNB decides to HO and sends a message to the AMF containing the list of destination ID, PDU session ID SM N2 information.

2. This is a conditional situation. If the UE resource is at a location that the AMF cannot manage, it connects to a new AMF that can serve that location and sends the Namf_Communication_CreateUEContext message.

3. Nsmf_PDUSession_UpdateSMContext request which is related to each PDU session is sent to SMF by the target AMF.

4. The SMF checks whether the HO can be performed for the PDU sessions specified in the previous step by looking at the target ID. If the UE is out of the range of the UPF that can connect to the gNB, the SMF allocates a new intermediate UPF. In this case, if the CN tunnel information is required, the message ''N4 Session Modification Request'' is sent to the UPF by the SMF. UPF responds to this request with the ''N4 Session Establishment Response''. If the SMF chooses a new intermediate UPF and the target UPF allocates CN tunnel information, the N4 Session Establishment Request message is sent to the target UPF. Thus, the target will have UPF package detection, application, and reporting capabilities. Intermediate UPF sends the N4 Session Establishment Response message to AMF. It should be noted that the situation mentioned in step 4 is a process.

5. AMF requests HO to the target gNB with responses from the SMF. Besides, it requests information about PDU sessions. The target gNB's response to the target AMF also includes information about the N3 tunnels it provides for each PDU session can perform HO and the rejected PDU sessions.

6. The target AMF routes the PDU session information from the target gNB to the SMF. If data from UPF to gNB will be routed, SMF is tasked with creating the routing tunnels needed.

7. With this phase, the target AMF notifies the source AMF that the necessary preparation has been completed.

2) EXECUTION PHASE

After the preparations for the HO are completed, there are several steps involved in the execution phase as shown in Figure 8. These steps are as follows:

1. A HO command, including information in the preparation phase, is sent to the source gNB by the source AMF. The source gNB informs the UE after the message it receives, thereby initiating the procedure that allows the UE to switch to the target gNB.

2. The source gNB sends the upstream ''RAN Status Transfer'' message to the source AMF. This information is then transmitted to the target gNB via the target AMF.

3. With this step, UPF can provide packet transmission directly or indirectly to the target gNB.

4. The UE sends a message to the target gNB that it confirms the HO and moves to this cell synchronously with the target cell. Then it passes the information that the target AMF has been moved to the target cell. The target AMF resource reports the release of resources provided by the AMF to UE. The target gNB sends the downlink user plane data and uplink user plane data directed by the UE to the UPF.

5. AMF inform all SMFs about PDU sessions. This information also includes N3 tunnel information. UPF sends one or more ''end marker'' packages to N3 tunnels on the old path.

6. The registration procedure is initiated by the UE.

7. In the last step, the units that are no longer associated with the UE will release the relevant resources.

C. HO FROM 5G TO LTE

5G and old technology devices should work in coordination with each other. Therefore, the procedure for HO of the UE from gNB to eNB is illustrated in Figure 9, and the simplified flowchart of this procedure is as in Figure 10.

1. According to the measurement results from the UE, the HO procedure is triggered from gNB to eNB, and then gNB sends a message to AMF that HO is required. By looking at the ID of the target BS, it realizes that the HO will happen to the LTE BS and contacts packet data network gateway control plane function (PGW-C) and SMF to create the context to help send the SM context to MME.

2. AMF selects an MME that can communicate with the target eNB based on the information provided by gNB. Thus, the AMF can send a reallocation request to MME. MME will then operate S1-based transfer procedures and request serving gateway (SGW) to set up Sessions and ask eNB to initiate HO processes.

3. Where indirect routing is valid, the SMF is triggered by AMF to set up routing tunnels.

4. The UE is handed over to the target cell.

5. For indirect routing, downlink data transmission can be performed and uplink data transfer is provided via the target eNB.

6. eNB informs the MME that the HO is complete, and MME communicates with the AMF and reports that

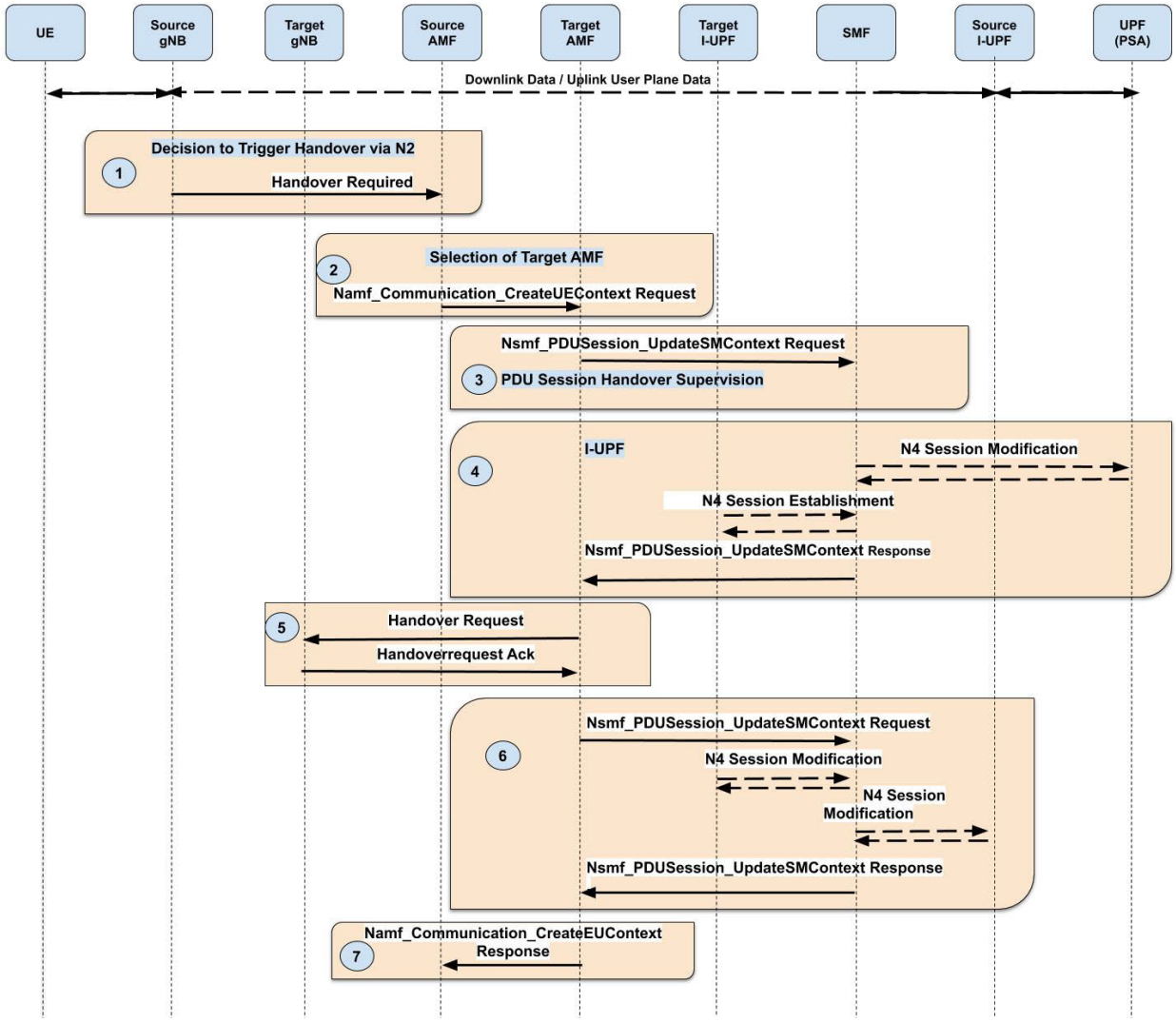

FIGURE 7. Preparation phase of N2-based HO [32].

the HO is complete. MME transmits tunneling information to SGW, PGW-C, and SMF to direct the downlink path to eNB.

7. In the last step, TA is updated and the installation of dedicated bearer is performed.

D. HANDOVER OF A PDU SESSION PROCEDURE BETWEEN THIRD GENERATION PARTNERSHIP PROJECT (3GPP) AND UNTRUSTED NON-3GPP ACCESS

5GC architecture supports devices connected via both non-3GPP access technologies and 3GPP access technologies. Connecting a device to the network over non-3GPP access technologies in most cases means that the device is connected via a Wi-Fi access network rather than a 3GPP radio network. Besides, ''untrusted'' access means that the 3GPP-defined mobile network operator doubts the reliability of the non-3GPP access network. This is because Wi-Fi network typically use password-based access authorization methods and sometimes lack payload encryption. The security issues make Wi-Fi networks unreliable to allow access to mobile network structure and services. The 5GC architecture includes non-3GPP Interworking Function (N3IWF), which acts as a gateway to the mobile network, and a port for devices that communicate over a non-3GPP access network [27]. Thus, traffic to and from devices can be routed between the untrusted access network and the mobile network over the public Internet.

1) HANDOVER OF A PDU SESSION PROCEDURE FROM UNTRUSTED NON-3GPP TO 3GPP ACCESS

This section describes how to move a PDU Session from untrusted non-3GPP access to 3GPP access. Such HOs are

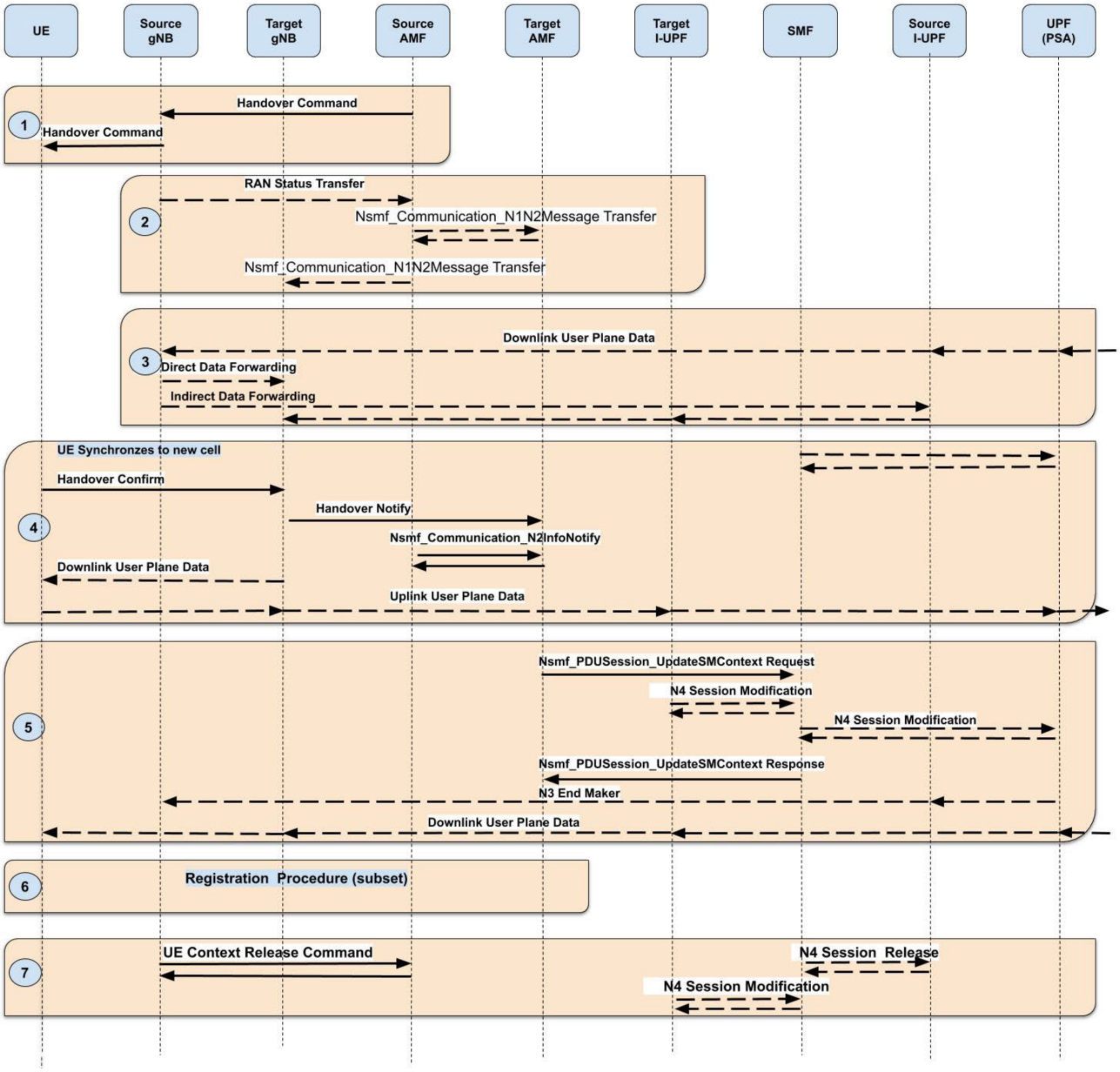

FIGURE 8. Execution phase of N2-based HO [32].

performed using the PDU Session Establishment procedure at the target 3GPP access, following the steps shown in Figure 11.

1. If the UE is not already registered for 3GPP access, the registration procedure is initiated. Then, to move a PDU Session from non-3GPP access to 3GPP access, the UE performs a PDU Session Establishment procedure using the PDU Session ID of the PDU Session to be moved.

2. After the PDU Session is moved, the SMF executes the release of resources over non-3GPP access. Thereby, user plane resources and PDU Session context in the N3IWF are released.

2) HANDOVER OF A PDU SESSION PROCEDURE FROM 3GPP TO UNTRUSTED NON-3GPP ACCESS

The HO procedure from 3GPP access to untrusted non-3GPP access follows the steps shown in Figure 12.

1. The UE starts the registration procedure if it is not registered to untrusted non-3GPP access yet. Next, the PDU Session Establishment procedure is executed over non-3GPP access to move a PDU Session from 3GPP access to non-3GPP access.

2. After the PDU Session is moved, the SMF executes resource release in 3GPP. Thereby, user plane resources and PDU Session context in AMF and NG-RAN are released.

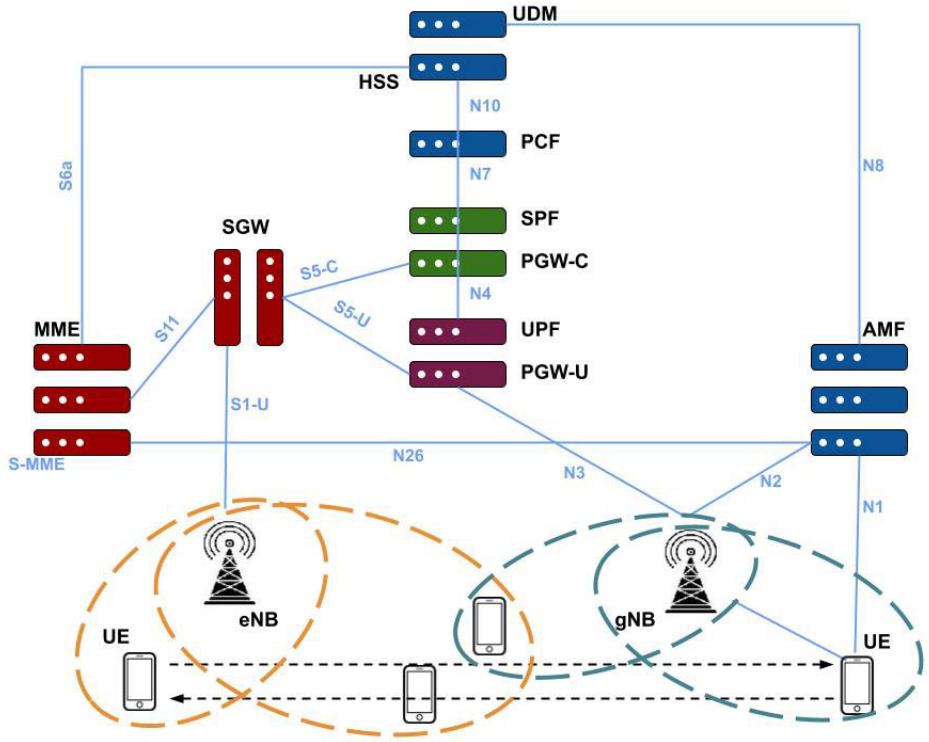

FIGURE 9. HO process from 5G to LTE.

VI. BEAM LEVEL MOBILITY AND BEAM MANAGEMENT

Growth in the number of mobile devices, applications, and services causes an increase in data traffic on the network. It is predicted that the system capacity needed in 5G will be 1000 times the system capacity in 4G and the data rate will increase 100 times compared to the data rate in 4G [34]. With the new radio access technology (RAT) deployment formed with 5G, system bandwidth reaching 1GHz can be achieved with mmWave ranging from 10GHz to 100GHz integrated in bands sub-6GHz. However, mmWaves are subject to high path and high penetration loss by their nature. Besides, blocking caused by various objects also affects the robustness of the communication. To cope with these issues, large-scale and high-dimensional antenna arrays are used. Thus, with beamforming technology, by using narrow beam, by concentrating the electric field in the desired direction, the gain increases and satisfactory communication quality can be achieved by compensating the loss effects. The gain mentioned here is achieved by using multiple antennas can be multiplexing gain, diversity gain or antenna gain. Thus, while the data rate increases, error, and signal-to-interference-plus-noise ratio (SINR) performance of the system improves [35].

There are different beamforming architectures used in UE and gNBs, these architectures are analog beamforming, digital beamforming, and hybrid beamforming, a mixture of these two architectures. In analog beamforming structures, while signal processing is carried out in the analog domain, the transmission or reception of the beam takes place over a single RF chain for all antenna elements, resulting in low flexibility of the structure. In analog beamforming, a pair of analog-to-digital converters are needed, thus saving power in the system [36]. In analog beamforming, the phase of each antenna is controlled by low-cost but constant amplitude phase shifters [37]. The constant amplitude of phase shifters is a factor that degrades system performance. In digital beamforming independent RF chain and data converters are required for each antenna element. Thus, the signals received in the frequency domain are processed, allowing the transmitters to send the beam in an infinite number of directions. In digital beamforming, the transceiver creates a strong structure by taking samples from the received signal and weighing them differently, increasing the flexibility of the system. However, establishing an RF chain for each antenna that creates large-scale antenna arrays increases the total cost, while creating such a structure increases the complexity of the system and the amount of energy it consumes. Due to the disadvantages described above, both beamforming structures cannot be used in mmWave. In this case, hybrid beamforming, which is a mixture of these two beamforming structures, comes to the fore. The main idea in the hybrid architecture is to divide the large-size digital signal processing into a large-size analog signal processing and the reduced-size digital signal processing [38]. Since there is usually very few effective

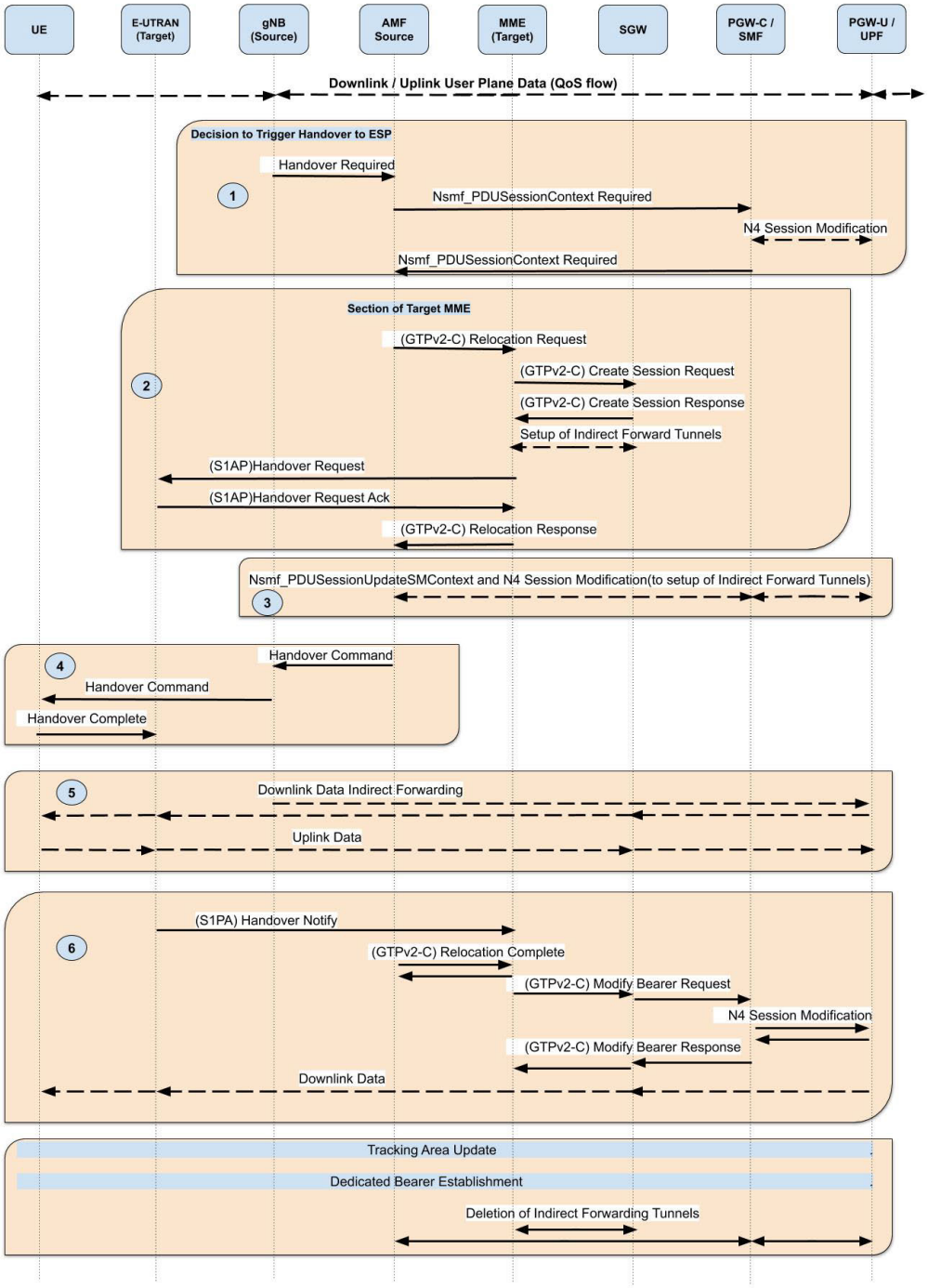

FIGURE 10. Flow chart of HO process from 5G to LTE [32].

scattering due to high loss in mmWave frequencies, the optimum number of data streams is generally much lower than the number of antennas. The number of RF chains needed is subtracted from the number of streams, and the hybrid architecture can meet that number. Thus, both the cost required to build the structure decreases and the energy consumed decreases. In [39], when compared to hybrid architecture and completely digital architecture, it is seen that it provides almost the same product performance with lower transceiver-receiver complexity and cost.

A. NR EVALUATIONS FOR BEAM MANAGEMENT

Synchronization signal (SS) blocks and bursts are used to ensure signal synchronization in 5G. Each block consists

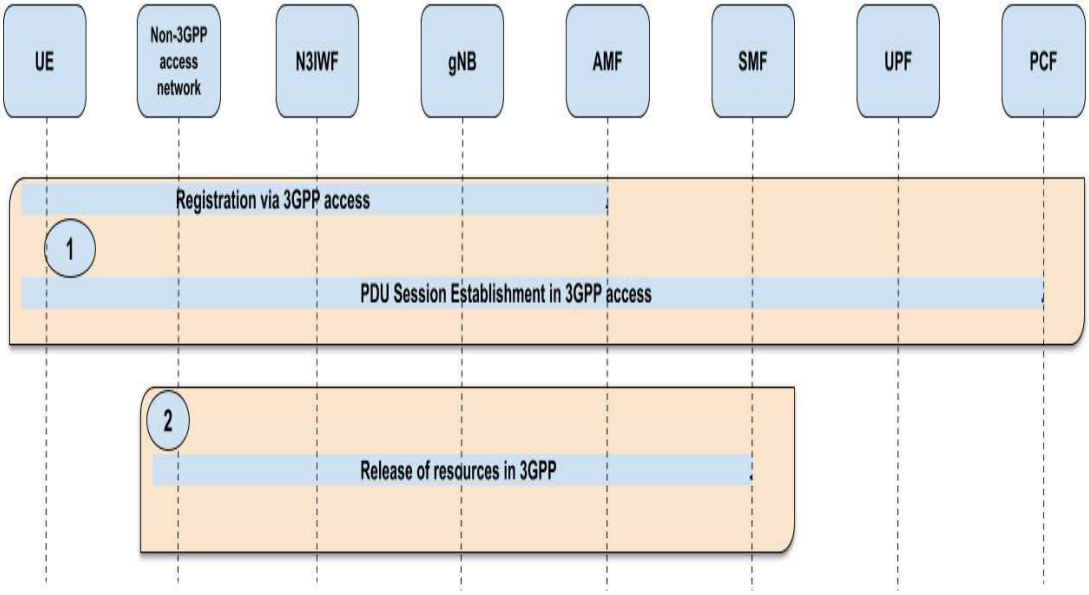

FIGURE 11. Handover of a PDU Session procedure from untrusted non-3GPP to 3GPP access [33].

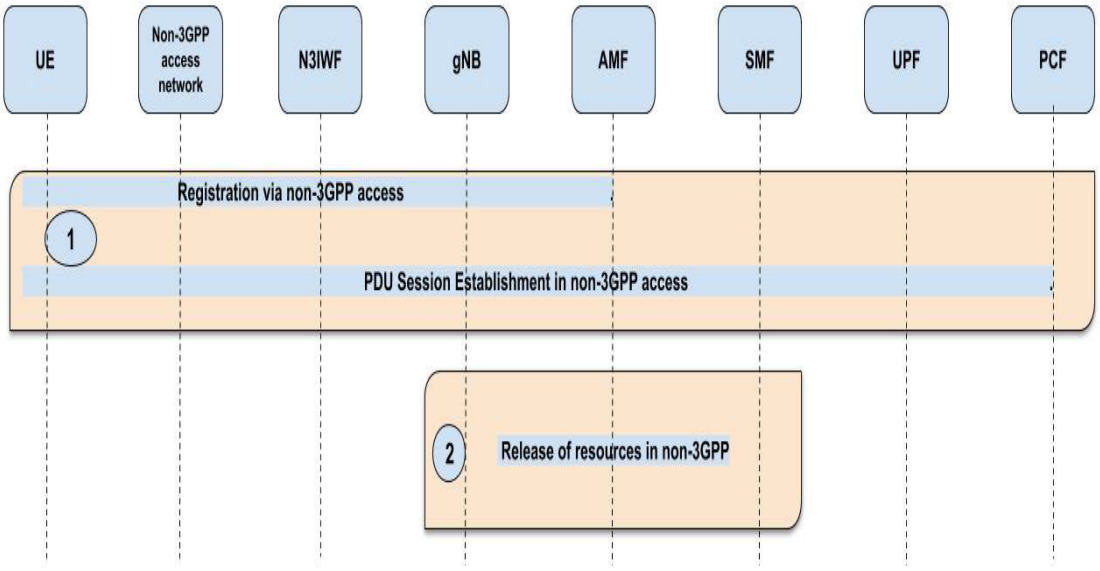

FIGURE 12. Handover of a PDU Session procedure from 3GPP to untrusted non-3GPP access [33].

of four orthogonal frequency-division multiplexing (OFDM) symbols in the time domain and 240 subcarriers in the frequency domain. The block structure includes primary SS (PSS), secondary synchronization signal (SSS), and physical broadcast channel (PBCH). The synchronization signal block (SSB) is measured using the demodulation reference signal and the associated PBCH to measure the reference signals received power of the SSB and select the appropriate beam for communication according to this value. Transmission of SSB is provided in the first 5 seconds of SS bursts [40]. SSB has different periods such as $T_{SS} = \{5, 10, 20, 40, 80, 160\}$ ms. While the UE is accessing the network for the first time, it determines its periodicity as 20 ms. Each gNB inside the cell transmits the SSB in different angular directions and scans the cell. Measurement reports containing signal strength and signal quality information are transmitted to the network. According to the reports, narrow beams are selected and thus, a better data transmission is provided.

Channel state information reference signal (CSI-RS) is a downlink signal and used in radio resource management (RRM) calculations. These RRM calculations are used to manage mobility when UE is in RRC Connected state

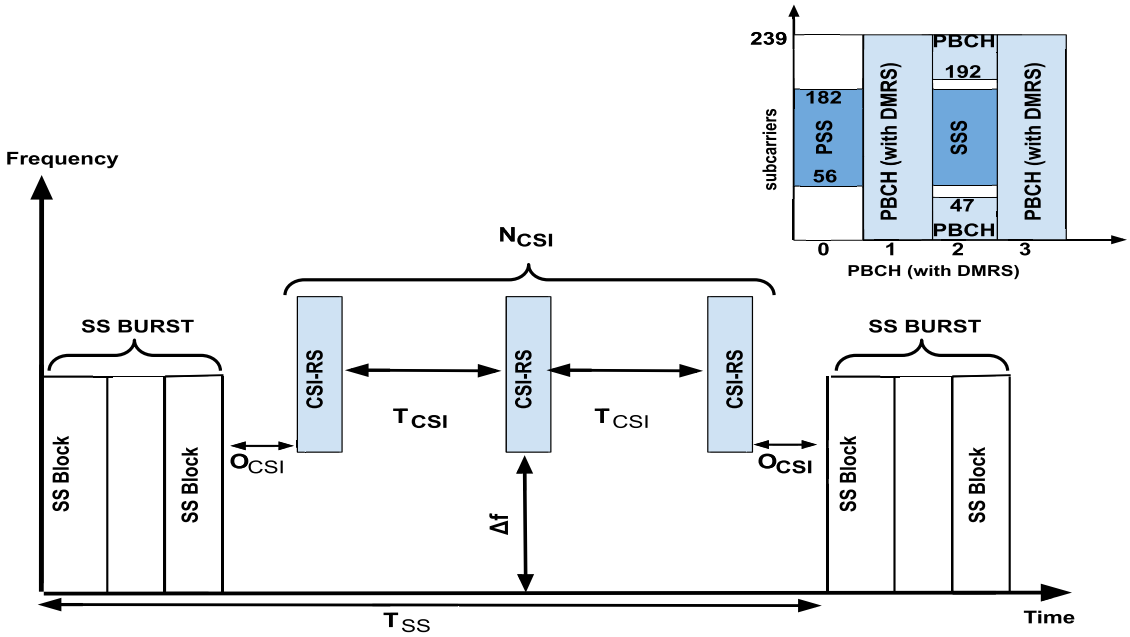

FIGURE 13. CSI-RS measurement window [38].

[31]. Because each CSI-RS has its unique identifier, UEs know which sources of frequency and time domain to send CRI-RS signals to. Thus, they provide cell synchronization using the same SS bursts and then use them as reference when searching for CSI-RS resources. CSI-RS measurement window as shown in Figure 13.

There should be a certain offset in the frequency and time domain between the CSI-RS packets and associated SS bursts, and these offsets should be periodic. There are two cases according to the values that this offset will take. The first case is that the time offset between consecutive CSI-RS packets is the same as the time offset between the CSI-RS and associated SS burst. In the second case, the period between CSI-RS packets is the case where the offset value between CSI-RS packets and the associated SS bursts is different [38]. There is a periodicity between CSI-RS packets and these periodic times can be $T_{CSI} = \{5, 10, 20, 40, 80, 160, 320, 640\}$ ms. The network can broadcast CSI-RS signals to the entire frequency band or broadcast within a certain frequency range.

The sounding reference signal (SRS) is an uplink signal sent by the UE. By receiving this signal, gNB obtains information about channel quality. Thus, the network knows which frequency range can provide better signal quality within the frequency band, and thus, it provides better quality service by allocating the frequency band according to each UE. It is divided into three groups according to the way of sending SRS signals. The first is periodic SRS signals are sent periodically. In the second case, it is sent periodically again, but the gNB can intervene and activate the signal transmitting or inactivate the signal transmitting. In the third case, the sending of the reference signal is triggered by the gNB.

B. BEAM MANAGEMENT

The purpose of beam management is to provide the most efficient connection between gNB and UE without interruption. In 5G, in the beam communication with mmWave, control signals are transmitted in the physical and medium access control (MAC) layers for beam adjustment and directionality. Thanks to the control element in the MAC layer, these control signals are transmitted between the UE and gNB. The beam management procedure in 5G consists of 4 parts, which are beam sweeping, beam measurement, beam reporting and beam determination.

- Beam sweeping is the process of grinding a spatial space with beams transmitted and received at predetermined intervals and directions for a predetermined period of time. Each Tx beam is transmitted through an reference signal source to sweep over multiple Tx beams for beam management. To sweep over multiple Rx beams, each Tx beam is transmitted over the same number of reference signal sources by the number of Rx beams. For example, assuming that transmit-receive point (TRP) has N number of Tx beams and UE has M number of Rx beams, then the number of CSI-RS transmission instants required is NxM [41]. The number of Tx/Rx beams should be large to make sufficient spatial field scanning. Effective beam measurement and reporting procedures are important at this stage to ensure low overhead and UE complexity [42].
- Beam measurement is the acquisition of some properties of beamforming signals received by gNB or UE to evaluate the quality of the signals using different measurements. In the previous section, it mentioned that

beam scanning based CSI-RS for downlink and SRS for uplink were measured by measuring beam.

- Beam reporting is the procedure for sending beam quality information to the RAN based on beam measurements.
- Beam detection is the technique of choosing the Tx/Rx beam(s) that best meet the predetermined criteria from the beam pool. Here, TRP provides the information of the beams used in data transmission to the UE so that the UE can choose the most suitable beam.

C. STANDALONE (SA) VS NON-STANDALONE (NSA) BEAM MANAGEMENT

The ability to work in two different stages, such as SA and NSA, is one of the hallmarks of the new generation RAN. The architecture of SA and NSA is introduced in detail in 3GPP Version 15 (phase-1) ([24], [43]). SA and NSA architectures on existing 4G networks are shown in Figure 14. In the SA phase, the UE is connected to the new wireless evolved NodeB (NReNB) and routed to the 5G CN through the NG interface. In the NSA phase, the UE connects to NReNB and is then forwarded to 4G eNB via the X2 interface or directly via the evolved package core (EPC) developed by the S1 interface. In the early stage of 5G deployment, wherein 5G CN is not implemented, the eNB is directly connected to the EPC via S1-interface and to the NReNB via the X2-interface. The NReNB may also be connected to the EPC via the user-plane S1-U interface and to other NReNB via the user-plane X2-U interface.

In SA beam management, the mobile terminal needs to wait for the gNB to serve it in the best direction it chooses during the beam determination phase to evaluate random access channel (RACH) opportunities to provide random access. In non-standalone beam management, the UE uses the support of the LTE layer to get information about RACH related opportunity, so full beamforming RACH can be planned without having to track the versatile SSB and associated RACH. In [44], it has been found that an NSA configuration utilizing multi-connectivity in mmWave networks offers improved end-to-end performance, and when the radio link fails, it provides more flexible and improved reactivity, reduces load effect in beam reporting, and provides more reactive reporting service.

VII. MOBILITY MANAGEMENT CHALLENGES

In this section, the challenges encountered in mobility management in 5G HetNets are discussed in detail and the solution suggestions presented in the literature are described to deal with these challenges. Challenges in 5G mobility management, the effects of these challenges and proposals are summarized in Table 2.

A. USE OF mm-WAWE BANDS

With the explosion of mobile traffic demand, mmWave provides an important opportunity to overcome the conflict between capacity requirements and spectrum shortage. The mmWave is to significantly increase the communication capacity by making use of the enormous size spectrum. Unlike crowded bands sub-6GHz, mmWave bands between 30 and 300 GHz have raw bandwidth and provide high data rates thanks to low order modulation. In addition to its enormous benefits, mmWave brings a lot of challenges.

Precipitation causes absorption, scattering, and diffraction of radio waves, resulting in increased transmission losses and decreased signal levels. This can seriously affect the propagation of the mmWave signals and lead to high signal attenuation along the propagation path. Attenuation caused by rain increases significantly with increasing parameters such as frequency, rain intensity, and effective length. This affects the reliability of the communication link and may cause existing connections to become unusable. It is essential to use real measurement data to make more accurate estimates and better performance in the design of 5G system channels. Accordingly, rain rate and rain attenuation at 38 GHz were analyzed based on measurements performed at University of Technology Malaysia (UTM)'s Skudai Campus for one month in a tropical region with heavy rains such as Malaysia in [45]. The results showed that the rain attenuation at 38 GHz was critical and that specific rain attenuation at 0.01% of the time could result in 18.4 dB/km.

The mmWave connections are very sensitive to fast channel variations and suffer from serious free space loss and atmospheric absorption. To cope with this problem and increase system capacity, both UE and BS use progressive array beamforming and large MIMO techniques. Thanks to the small wavelength of the mmWave signals and advances in low-power complementary metal-oxide-semiconductor (CMOS) RF circuits, it is possible to use large antenna arrays even in a small structure such as a telephone. Directive communication with increased system capacity can be achieved with beamforming techniques to compensate for the loss of propagation. In addition, beamforming techniques mitigate performance requirements in each antenna and RF circuit and greatly suppress interband interference. On the other hand, by using MIMO structures and beamforming techniques in both UE and BS, high energy and spectral efficiency are also provided as well as increasing the robustness of communication.

The mmWave signals suffer from difficulties such as high path loss, severe channel interruption and blocking by building materials such as bricks and mortar, and even the human body [46]. Therefore, the quality of communication between the UE and the serving cell is highly variable due to factors that can cause rapid drops in signal strength, such as the movement of obstacles or the change of the body's position relative to mobile devices. Thanks to multiple connectivity, a UE can be connected to different cells simultaneously and have multiple signal paths. Thus, if the connection quality on a data path decreases, it can change the data path and ensure the continuity of the communication by preventing the connection quality from decreasing. In mmWave cellular

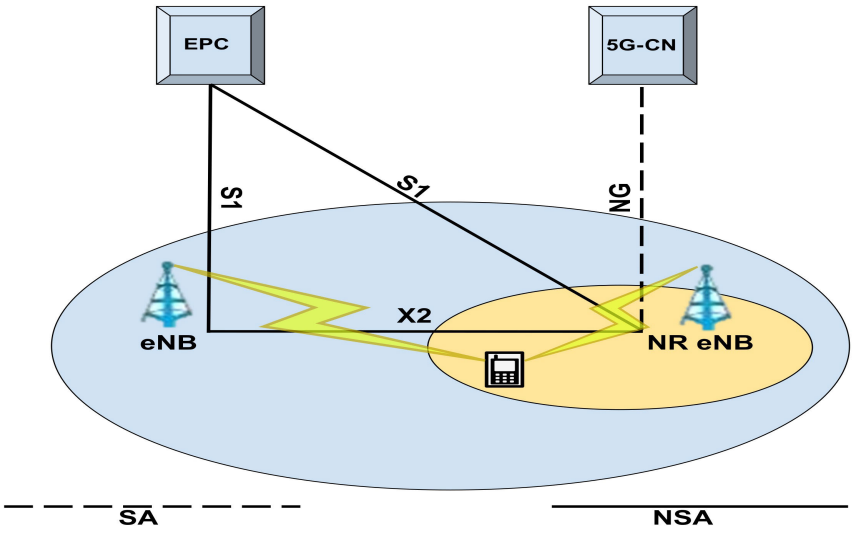

FIGURE 14. Architecture phases of SA and NSA.

networks, multiple connectivity can be established between 5G mmWave cells or between 5G mmWave cells and 4G cells. Multiple connectivity between 5G mmWave cells offer higher bandwidth, while multiple connectivity between 5G mmWave cells and 4G cells provide more robust communication. A new multicellular measurement reporting system is proposed in [47]. In this system, each UE broadcasts SRS directionally in varying directions over time to scan the angular space. Each potential service cell in the system scans in all angular directions in order to better analyze the dynamics of the channel and monitor signal strength with the variance of SRS it receives from UEs. A central controller then acquires full directional information from potential cells, making the service cell selection and timing decision. According to the simulation results, the proposed system has shown that the delay in the creation of digital beams in BS can be significantly reduced and that a UE can see multiple mmWave cells in a reasonable cell density state. In [48], a DC protocol is provided that allows UEs to physically connect with both 4G and 5G cells simultaneously. Thus, thanks to a local coordinator, an uplink control signaling system provides fast patch switching in case of any malfunction. The local coordinator, which manages the intercellular traffic, performs the tasks of the control plane as the path switching and tasks of the data plan as traffic anchor in the packet data convergence protocol (PDCP) layer. Using the DC framework, faster mobility management is provided compared to the hard HO diagram. Moreover, dynamic TTT adaptation has been proposed to improve key decision timing in very uncertain situations. Along with the proposed framework, according to the simulation results, it was observed that higher efficiency stability was achieved and improvement was provided in parameters such as delay, packet loss, control signal load.

However, in mmWave, multiple connectivity is much more complex because transmissions in multiple connectivity are likely to be omnidirectional. For this reason, every potential connection between the network and the UE needs to be examined continuously, but following the changing directions slows down the adaptation rate of the network and the quality of the connection makes it difficult to provide a robust service. In addition, because the UE and BS can only listen to one direction at a time, the necessary control signaling to change routes is difficult to occur. To cope with these difficulties, different methods should be used instead of constantly monitoring each connection. Synchronization signals are transmitted by beam scanning along the angular coverage zone during cell discovery. At this stage, the best performer is chosen from the combinations of beam pairs between UE and gNB. However, sequential search causes large amounts of access latency and low initial access efficiency. Therefore, using the repetitive neural networks (RNN) called gated recurrent unit in [49], a beam sweeping model based on the dynamic distribution of user traffic is presented. Spatial distributions of users are provided from the data in the cellular network call detail records (CDR). Using RNN to predict CDRs with high accuracy, mmWave uses historical CDR data to quickly determine the scan direction in the cellular system.

RF receiver and transmitter should be as compact as possible to reduce energy and cost-effectively in 5G mmWave systems. As described in the previous section, hybrid analog/digital architectures are used to reduce the number of RF chains. It is easier to build scalable array systems with these architectures. In the transmitter part, data flows are first processed by the digital pre-encoder and then converted to RF frequency and mapped to all antennas for transmission over a phase shifter network. The biggest challenge to maximizing capacity in hybrid analog/digital architectures is the creation of analog and digital beamforming matrices [50].

When a UE is within the coverage of the BS for the first time and the beams between the UE and BS need to be

TABLE 2. Summary table of challenges.

Challenges	Description	Consequences/Effects	Proposal	Ref.
Use of mmWave bands	• The mmWaves suffer from challenges such as high path loss, severe channel interruption and blocking by building materials such as bricks and mortar, even by the human body. • The mmWave connections are very sensitive to rapid channel variations, that can cause serious free space loss and atmospheric absorption. • Precipitation can seriously affect the propagation of mmWave signals and lead to high signal attenuation and low signal levels along the path of propagation.	• The quality of communication between the UE and the serving cell is quite variable due to factors that can cause rapid decreases in signal strength, such as the movement of obstacles or the change of the body's position relative to mobile devices.	• Using progressive array beamforming and large multiple-input-multiple-output (MIMO) techniques is an important solution to deal with severe free space and atmospheric absorption and increase system capacity. • Since multiple connectivity allow a UE to connect with more than one cell at the same time, it can ensure the continuity of communication by preventing the connection quality from decreasing. • In the design of 5G mmWave system channels, it is very important to use real measurement data to make more accurate predictions and provide better performance.	[45], [46], [47], [48], [49], [50], [51], [52], [53]
Load balancing	• In cellular networks, UEs are rarely distributed evenly across the network due to the random positioning of cells and the mobility of UEs.	• Although there are resources in the network that can be used in neighboring cells, some cells may become congested due to the intense UE association. This situation creates a load imbalance within the network, increases the rate of HOF and causes decrease in the quality of service (QoS) provided to the UEs.	• In order to provide UEs with more available resources and better network performance, load balancing schemes need to be created to transfer the network load from congested cells to lightly loaded small BSs.	[54], [55], [56], [57], [58], [59]
HO problems	• Intensive placement of small cells in HetNets created to meet the rapid increase in the number of mobile devices causes many challenges that reduce QoS such as in the interferance, frequently and unnecessarily HO, HOF and PPHO.	• Frequent and unnecessary HOs, HOFs and PPHOs cause unnecessary or erroneous procedures to be executed within the network and thus unnecessary signaling load. This results in unnecessary use of resources allocated to UEs, applications and services, and consuming energy for faulty procedures.	• Models that will minimize the unnecessary HO number and optimize the HO decision-making procedure should be used.	[7], [60], [61], [62], [63]
Signaling overhead	• With the increasing number of UEs, HO procedure will occur more frequently. The signal packet transmissions that HO must be able to perform cause additional signal overhead in the network. • Extensive deployment of BSs to serve UEs results in more BSs being subjected to paging procedures to find the location of a UE within the network. In this case, it causes additional signal loads on the network.	• The additional signal load on the network causes more interruptions in data transfer and latency in communication.	• Reducing the number of HO by optimizing HO parameters. • Design models where fewer BSs will be paged to find a user on the network. • To design models in which notification area concepts are eliminated.	[64], [65], [66], [67]
Power consumption	• HetNet is a structure in which many BSs operating at different frequencies work in an integrated manner. The UE makes measurements both inter-frequency and intra-frequency according to the carrier frequencies of itself and the BSs. • Deploying large numbers of small cells in a HetNet leads to increased cell edges and greater intercellular interaction, leading to frequent HO procedures. • The mmWave signals suffer from isotropic path loss. To reduce this effect, mmWave systems transmit with narrow and electrically steerable beams.	• Increasing the number of measurements made depending on the density of the network increases the battery power consumption of the UE. • Numerous antennas are used in mmWave systems to perform beamforming. Mixer and phase shifters integrated in antennas increase power consumption. • Power consumption increases as the number of BSs to be paged in the paging procedure performed to find the location of the UE.	• Power consumption can be reduced with hybrid beamforming structures that use fewer RF chains than digital beamforming. • The narrowing of the measuring gaps can have an effect that reduces power consumption. • The UE checks if the HO criterion is met by making measurements within the DRX cycle. Models with a smaller DRX cycle have lower power consumption.	[22], [26], [68],

TABLE 2. *(Continued)* **Summary table of challenges.**

Challenges	Description	Consequences/Effects	Proposal	Ref.
Energy efficiency	• BSs have an important share in total energy consumption in the network. Small BSs, an important solution to meet large mobile traffic demand and increasing capacity needs, lead to increased energy consumption. These BSs have energy consumption costs, even when the number of associated users is limited. However, a significant amount of energy is consumed during the HO procedure.	• Network operators' OPEXs increase due to increased energy consumption within the network. • Growth in carbon footprint resulting from the mobile communications industry that has increased exponentially over the years. • Due to the faulty procedures such as frequent and unnecessary HO, HOF and PPHO, the energy consumed by the network increases.	• To reduce energy consumption, it may be the solution to transfer the load from highly load cells to these idle cells or to close these cells if the system capacity will not have a serious effect. • It is important to make energy efficiency calculations at the decision stages of HO procedures.	[69], [70], [71], [71], [72]
Security	• 5G HetNets host a large number of small cells from different technologies. Frequent HO procedures cause mobility security issues within the network.	• With the authentication procedure between UEs and BSs, the network takes security precautions against malicious users. However, the transmission of signal packets and BS interfaces required for the authentication procedure causes HO latency.	• To reduce HO latency, proactive models in which the number of messaging in authentication procedures is reduced by means of various protocols instead of repeated authentication procedures for each HO procedure can be a solution.	[73], [74], [75]
URLLC	• URLLC is one of the core services of 5G wireless communication systems and this service requires strict requirements in terms of latency and reliability.	• To ensure high reliability, it is necessary to use an extended code word with an excess that will cause an increase in latency. To reduce latency in URLLC, the channel block length is finite in packet transmission. This leads to poor transmission rate and better probability of decoding.	• Multiple connectivity is an important solution to increase reliability in URLLC. • Large antenna systems have extraordinary features that can be useful with the ability to make multiple spatial degrees of freedom that determine the URLLC. Large antenna systems alleviate the need for powerful coding schemes, thereby maintaining high reliability for short packets and reducing the need for retransmission, providing high capacity for spatial division multiplexing.	[76], [77], [78]

aligned. In mmWave systems using hybrid beamforming, this procedure creates a significant load on the system. Reducing the discovery time of new users is essential. In addition, the reciprocal positions of the user and the BS are not known, and therefore the search procedure is a phenomenon that increases overheads. There are various approaches in the literature to find the location of users. Arrival statistics depend on the entry ways of the users along with the roads and buildings that make up the cell. With the help of these statistics, the first access procedure can be accelerated by scanning more frequently. For this purpose, an online method is proposed in [51] to obtain these statistics. In this way, by analyzing the typical behaviors of the users, the acquisition time can be accelerated by scanning more frequently on the places that the users use more. Moreover, optimization of a particular aspect that failed in the past can be achieved by these statistics. Thus, performance evaluation is made using real data. It has been argued that the presented method significantly reduces discovery time compared to methods that are not based on user statistics. A gradient descent algorithm is proposed in [52] to solve the beam alignment problem at 5G mmWave, enabling the characterization of pure Nash equilibrium and enabling users to learn optimum beam widths. The game is characterized by parameters such as beam width, the distance between the transceiver, the possibility of alignment, and the transmit power to align the beamforming direction to maximum efficiency. Along with simulation results, the effects of transmit power, mmWave frequency, and learning rate on beam alignment and convergence time are shown. In [53], integrated machine learning and coordinated beamforming solution are presented for highly mobile mmWave applications. The solution offered for coordinated beamforming also serves a number of BS users simultaneously. With the deep learning model, it is learned how the beam generator vectors of BS can predict omni or semi-omni beam models from signals received from distributed BSs. Signals received from distributed BS provide the information necessary to adapt to the environment in which the user is located. The results showed that the user with a deep learning model with sufficient learning time and a robust beamforming system was successful in adapting to changing environments.

B. LOAD BALANCING

HetNet structures, in which small BSs work integrated with macro BSs, expand network coverage at a lower cost, while also improving network performance parameters such as spectrum efficiency and energy efficiency. In highly dense HetNets, there is a load imbalance between the cells due to

the random positioning of the cells and the mobility of the UEs. Load imbalance within the network increases the rate of HOF and reduces the efficiency of network performance. For example, in traditional user association architectures that determine the target cell for which HO will be realized based on the maximum signal-to-noise ratio (max-SINR) criterion, most users associate with macro BSs because the power supplied by the macro BSs is greater than the small BSs. Thus, while small BS resources are used inefficiently, macro BSs become overloaded. As a result, load imbalance occurs between the BSs in the network, and the QoS provided to users is reduced. Theoretically, if a UE is handed over to an already overloaded cell, this results in a shortage of resources in the loaded cell and a decrease in the QoS provided in both the existing UEs in the cell and the new UE, and the resulting HO is considered unsuccessful. For this reason, load balancing, which will be achieved by transferring the load from overloaded cells to idle cells, has an important place in recent research. In these studies, the parameters to be used in the HO of UEs are adjusted according to the load conditions of the small cells in the environment. The point to be noted here is the correct adjustment of the parameters, otherwise this will result in inefficient use of network resources and inadequate QoS provided to UEs.

A cluster-based load balancing algorithm is proposed in [54] instead of a mobility load balancing across the entire network to eliminate load unbalance for better network performance. The proposed algorithm models the network to create clusters from overloaded cells and their n-layered neighbors. Thus, dynamically adjusting cell individual offset (CIO) parameters for each cluster, load balancing is done locally, thereby avoiding unnecessary load balancing actions. Simulation results show that in the low-speed UE scenario, the proposed algorithm provides an increased network performance throughput of 6.42% compared to a network without load balancing. Current load balancing solutions generally formulate this problem according to the full CSI between the user and each BS, and this formulation at the global level complicates structures while generating additional signal load. In [55], the load balancing problem is solved by a low complexity successive offloading scheme created using a limited CSI feedback structure where each user measured and reported sync signal strengths of nearby BSs instead of a global CSI feedback structure. Thus, it has been demonstrated with simulation results that it can achieve results close to optimal load balancing performance with lower complexity. In [56], a utility-based mobility load balancing (UMLB) algorithm that determines the edge UEs of an overloaded cell required for load transfer to occur from an overloaded cell to lightly loaded neighboring cells and provides the best neighboring cell HO by calculating the total utility of these candidate UEs with each neighboring cell, and load balancing efficiency factor (LBEF), a newly introduced term that will indicate the order of overloaded cells for the UMLB algorithm process, are presented. According to simulation results, UMLB minimizes standard deviation with a higher average UE data rate than current load balancing algorithms.

It is thought that using machine learning algorithms in load balancing management will provide a better service to users and bring great advantages. In [57], a deep learning-based mobility strength optimization solution is presented, which learns the appropriate values of the parameters required for the mobility model of each cell in the network. Optimum mobility setting has been made for transmission parameters depending on the distribution of users and their speed in the network. Mobility sensitive load balancing approach has been applied to provide approximately the same QoS to each user. Simulation results show that the proposed system optimizes HO performance and can learn parameter settings that will provide better performance compared to the benchmarked reference. [58] provides a framework that combines real-world data urban incident detection with proactive load balancing. First, urban incident detection was proposed to predict changes in cellular hot spots based on Twitter data. Then, with the a proactive load balancing strategy, simulation is done by taking into consideration the distorted hot spots in the urban area. At the last stage, proactive load balancing strategy is optimized to estimate the best activation time. BSs supported by unmanned aerial vehicles (UAV) are a promising solution to deal with the problem of balanced network load with features such as flexible distribution, wireless coverage everywhere, and high data rate advantages. However, how UAVs can be deployed autonomously and dynamically within the network is an important challenge. In [59], UAV BSs intelligent distribution plan based on machine learning is proposed and performance measurements are evaluated in the real data set. Data preprocessing is performed to process data received over the network, clear, and convert. Next, predictions are made with the hybrid approach, which includes the Autoregressive Integrated Moving Average (ARIMA) model and the eXtreme Gradient Boosting (XGBoost) model. ARIMA is a linear prediction model and XGBoost is a nonlinear prediction model. The proposed hybrid model is used to estimate the number of future users with XGBoost nonlinear prediction based on ARIMA's linear prediction. According to the prediction results obtained in the last stage, UAVs are deployed to dynamically meet the demands in hot spot areas. Simulation results have proven that this approach is a successful scheme in load balancing.

C. HO PROBLEMS

The rapid growth of the number of mobile devices associated with the cellular network in the new generation 5G networks has led to the demand for high data traffic. Since macrocells cannot meet this need, HetNets are created by integrating small cells with lower transmission power and coverage compared to macrocells into existing networks [79], [80]. In addition to the extraordinary benefits of small cells, it also brings with it many difficulties that reduce QoS such as interference, frequent and unnecessary HO, HOF, and

PPHO due to the intense deployment of small cells in the network. Thus, an increase in signaling load occurs. This increase causes the resources provided by the network to be used unnecessarily and the network consumes energy for a faulty procedure. In the literature, there are studies modeling HO decision problems to ensure the continuity of the service provided to mobile users by minimizing the number of unnecessary HOs and to reduce the signaling load in HetNets.

In [7], a GRA-HO method is presented to reduce both HOF and HO rate in HetNet where small cells are densely deployed. Using the AHP technique, the HO metrics are weighted and a ranking is made between the existing cells using the GRA method and the cell with the highest order is selected for the HO. In order to prevent abnormalities and avoid unnecessary HO, the max-min normalization technique is used, which takes into consideration the benefit and cost-efficiency. Simulation results showed that the proposed method improved energy efficiency while decreasing HO rate and HOF compared to conventional multiple attribute decision-making (MADM) methods including Vise Kriterijumska Optimizacija IKompromisno Resenje (VIKOR) and simple additive weight. VIKOR is an algorithm that helps the decision-maker reach a final decision by enabling the determination of the compromise solution of a problem with contradictory criteria and listing the chosen set of alternatives [60]. In [61], a MADM technique called Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) for sorting HO candidate cells by characteristics and weight of each feature to model the HO decision. MADM is concerned with the selection of the best alternatives that are characterized by multiple features. Two methods are presented in HetNet, such as the entropy weighting technique (named PE-TOPSIS) and a standard deviation weighting technique (named PSD-TOPSIS) to weight HO metrics. According to the simulation results, it is seen that both methods suggested reduce the number of HOs and HOF, thereby increasing user efficiency. But PSD-TOPSIS performs better with a more complex structure. Thus, one of the two methods can be used depending on the size and capacity of small cells. In [81] and [82] proposed a weighted scheme to adjust values of TTT and HO margin (HOM) based on three parameters: SINR, cell load and user speed. This scheme aims to address the conflict problem between mobility robustness optimization and load balancing optimization.

In [62], an SDN based mobility strategy was proposed using the Markov chain formulation to predict the possibilities of the neighboring eNB transition and the available resource to solve the HO latency problem. With this strategy, the most appropriate eNB selection is made and these selections are assigned to mobile nodes according to OpenFlow tables. Thus, both HOF and HO latency is reduced compared to conventional LTE HO operations. In [63], a velocity-based self-optimization algorithm is proposed to set HO control parameters in 4G and 5G networks. This algorithm uses the power received by the user and the speed information of the users to set the HO margin and the trigger time during the user's movement on the network. Simulation results show that there is a significant reduction in PPHO and RLF rates compared to existing algorithms.

D. SIGNALING OVERHEAD

Inter-cell HO occurs more frequently due to the intense deployment of huge numbers of small cells in HetNet with 5G. During HO operation, reciprocal signal packets must be sent between the source cell, the target cell, and the UE, so that users can be registered with the target cell by performing HO. More frequent occurrence of HO will create an additional signal load on the network and lead to more interruptions during data transfer. This causes a trade-off to occur between the additional signal load from the frequent HO and the network coverage. For this reason, it is essential to achieve this balance with smart mobility management solutions. One solution to reduce the signal load would be to reduce the number of HO. An optimization algorithm is proposed in [64] to adjust the HO margin and TTT adaptively according to the user speed and received signal reference strength. Thus, it is aimed to decrease the HO number and HOF rate. Simulation results show that the average PPHO and HOF parameters are significantly reduced compared to the studies in the literature with the proposed algorithm. Besides, it reduces latency and interruption during communication. A wireless test-bed based on monitoring the location of the user equipment is proposed in [65], designed to evaluate the performance of ultra-dense networks and mobility schemes related to these networks. A mobility management scheme is provided to proactively track the location of the UE and provide uninterrupted HO service to users. Thanks to the SRSs transmitted by the UE, its position in the network is tracked. Thus, it eliminates the necessary signaling expenses for HO, but also brings with it the computational complexity. In [66], the CN resource control signaling scheme for the active state HO is proposed to evaluate the signal load as a function of network density, user mobility, and session features. Moreover, a smart HO scheme has been created so that the HO signal delay is tried to be minimized. This HO procedure, configured in control/data separation architecture networks, reduces mobility signals as it connects to a BS that offers UE wide coverage. Simulation results also show a reduction in signal load and HO latency.

Besides, with the increase in the number of BSs, it causes an increase in the number of paging procedures the system will perform in order to find a user in the network, and therefore in the total signaling load. In [67], a new solution gNB-based UE mobility tracking is proposed that eliminates the concepts of TAL and the corresponding IoT/UE based monitoring TAU/RAN-based notification area. This will save power because IoT/UEs do not report their location to the TAU. The gNBs, give the IoT/UEs always known positions, and take over the task of monitoring and placing them. Since the paging procedure is eliminated, these devices can be quickly accessed significantly. Moreover, this solution creates a slight signaling load with very low paging delay.

Simulation results showed that 92% reduction in signaling load is achieved compared to conventional TAU and paging structures.

E. POWER CONSUMPTION

HetNet consists of an enormous number of BSs operating at different frequencies. The UE makes both inter-frequency and intra-frequency measurements from different BSs. If the source cell and the target cell to which HO will occur are running at the same carrier frequency, this is called intra-frequency HO, and the UE does not need to change the carrier frequency and measurements are made according to the order of the cells. If the source and target BSs operate on different carrier frequencies, this is called Inter-frequency HO and the UE has to adjust the carrier frequency within the measurement ranges according to the carrier frequencies of neighboring BSs. In this case, measurements are made according to the quality of the carrier frequencies. It performs both intra-frequency and inter-frequency measurements in the order of priority assigned by BS in a UE target cell selection. Due to these measurements, an increase in battery power consumption occurs depending on the density of the network.

Many factors that decrease the QoS such as interference, frequent and unnecessary HO, HOF, and PPHOs caused by the intense deployment of small cells in HetNet will also cause unnecessary use of resources and unnecessary power consumption. However, due to the more frequent HO with HetNet, the frequency of sending mutual signal packets between the source cell, the target cell, and the user increases the power consumption in the network.

Power consumption on mobile devices is an important challenge in mmWave systems. It suffers from isotropic path loss due to the nature of mmWave signals. To deal with this situation, mmWave systems transmit with narrow and electrically steerable beams. A large number of antennas are used to realize beam steering. Analog beamforming shapes the beams along a single RF chain for all antenna elements. This operation is performed in the analog domain and can only transmit/receive in one direction at any time. It also saves power because it uses a single analog-digital converter. It also uses a single analog-to-digital converter, which saves power. However, since it can only transmit/receive in one direction, the flexibility of the system is low. In digital beamforming, a separate RF chain and data converters are required for each antenna element. Thus, the received signals can be processed in the digital field and the beams of the receiver/transmitter can be directed to an infinite number of directions, thus offering a significantly faster search. However, power consumption is high as digital architectures require a separate RF chain and an analog-to-digital converter per antenna. In [68], the power consumption of multi-carrier receivers is analyzed for systems operating at both 28 GHz and 140 GHz. It has been observed that the power consumed by the mixer and phase shifter among the RF front-end components is dominant. Alternatively, the hybrid analog-digital architecture in the transceiver is considered a cost-effective solution. Alternatively, the hybrid analog-digital architecture in the receiver/transmitter is considered a cost-effective solution. Hybrid beamforming uses fewer RF chains than digital beamforming, allowing it to use more antenna array elements while reducing energy consumption and system design complexity

F. ENERGY EFFICIENCY

HetNets are a promising solution to meet massive mobile traffic demand and growing capacity needs in 5G networks. Small BSs, which are deployed intensely and uncoordinated into HetNets, can meet these needs and cause an increase in energy consumption. This is because even in scenarios where there are no users associated with small cells, these cells have energy costs. In this case, the clever solution would be either transferring the load from cells with highly load to these idle cells, or shutting them off if the system capacity will not have a serious impact. BSs have a huge share in total energy consumption across the entire network. Therefore, network operators are trying to create efficient power management schemes to reduce operational expenditures (OPEX). OPEX is a cost model that includes technical and commercial operations, administration costs [69]. In addition, a significant amount of energy is consumed during the transfer of users to different cells. In order to avoid high energy consumption due to poor connection conditions of the target cell, it is very important to make energy efficiency calculations correctly in the decision phase of the HO process. In this way, the battery life of the users can be extended with the energy efficiency transfer procedures to be created. However, due to the growth in the carbon footprint resulting from the mobile communication industry and increasing exponentially over the years, efforts to reduce energy consumption are becoming more and more important.

In [70], the authors present a fuzzy logic-based game theoretical framework to optimize the ideal transmission BS power levels to serve UEs to ensure energy efficiency in HetNets. The proposed fuzzy HO scheme consists of two modules such as HO decision and target BS selection. The desired energy efficiency, PP rate, and efficiency can be flexibly adjusted by the network operator with many parameters such as speed, SINR, efficiency, and BS load in the HO decision. In the target BS selection module, the most suitable BS selection is made according to the order of preference created by considering the parameters. According to simulation results, it has been observed that energy consumption can be improved in the management of ping-pong transfers for high-speed users. In [71], a HO algorithm has been proposed that provides self-optimization of the system to improve energy efficiency and PPHO ratio. The main concept of the proposed algorithm is to select the system energy reduction gain (ERG) and PPHO rate threshold parameters in each time period through the sample of power consumption and SINR performance, and the network feedbacks the ERG and PPHO rate conditions to compare with two threshold parameters. ERG

is a parameter that expresses the energy consumption gap between different systems as a percentage. As a result of the comparison, the system optimizes the TTT and HOM's values of the next time period to achieve better system performance.

Thanks to their high mobility, UAVs provide more flexible and faster distribution for communication systems and offer higher connection capacity since they have higher line of sight channels compared to ground-to-ground connections. Because of these features, UAVs are predicted to play an important role in future wireless systems. The paper presented in [83] aims to analyze spectrum efficiency maximization and energy efficiency maximization designs and reveal basic trade-off by optimizing communication time allocation and UAV trajectory for a UAV-enabled mobile transition system. Simulation results have shown that there is a new fundamental trade-off between spectrum efficiency maximization and energy efficiency maximization.

Ultra-reliable low latency communications (URLLC) is a service intended to be used in critical applications requiring low latency in 5G wireless communication systems. The work in [72] presents an HO technique for 5G URLLC, bypassing the role of the source gNB, aiming to send HO requests directly between the UE and the target gNB based on the measurements of the UE. Thus, energy efficiency can be improved and HO can be achieved faster by reducing the user plane delay.

G. SECURITY

In 5G HetNet, which consists of the deployment of a huge number of small cells from different technologies, mobility security due to frequent HO is an important problem. With mutual authentication between UEs and BS, malevolent users take precautions against network effects such as Manin-the-Middle attacks, Denial of Service attacks, impersonation attacks and repeat attacks. Secure transport authentication is required to take precautions against these attacks and to provide robust communication when transferring in different networks. However, HO delays may occur due to messaging in the authentication procedure and the interfacing of the BSs. Therefore, it is not efficient to initiate an authentication procedure for each HO. In [73], a more secure vertical HO authentication scheme is proposed using the certificate-based authentication procedure, which is symmetric with the key distribution extensible authentication protocol-transport layer security (EAP-TLS), to ensure that the user equipment receives certificates from a foreign network. EAP is an authentication protocol used in network and internet connections, and EAP-TLS is a standard supported between wireless vendors using the TLS protocol. In [74], anonymous mutual authentication by key agreement is provided for those which HO, taking advantage of the trapdoor collision feature of the chameleon hash functions and the tamper resistance of the blockchains. In [75], two fixed trajectory groups for HO node authentication scheme is presented for the mobile relay node (MRN). Thus, authentication is performed before MRNs in a train arrive at the next BS, thereby ensuring uninterrupted communication.

H. ULTRA-RELIABLE LOW-LATENCY COMMUNICATION

URLLC, along with enhanced mobile broadband (eMBB) and massive machine-type communications (mMTC), is considered among the core services of 5G wireless communication systems. Enabling URLLC is particularly difficult because it needs strict requirements (1 ms unidirectional delay with 99.999% reliability within the radio access network) in terms of latency and reliability. In wireless communication, reliability is commonly defined as the probability of success in delivering a packet within the utmost time required. Multiple connections are a strong approach to increase reliability. Also, simultaneously supporting various services like eMBB and URLLC is another problem of 5G. In 5G, two technologies that provide versatile coexistence are flexible physical layer structure and network slicing. Features such as BS-specific frequency band and transmission time interval length greatly affect the guarantee of traffic requirements [76]. To reduce latency, the channel block length is finite in packet transmission, leading to poor transmission rate and better probability of decoding error [77]. The law of large numbers does not apply to such a situation, and Shannon capacity cannot be used to characterize system capacity. However, URLLC is harder to implement than eMBB and mMTC, because the URLLC targets extremely high reliability and ultra-double QoS requirement. Specifically, to attain high reliability, it is necessary to use an extended code word with redundancy, which results in increased latency. Additionally, it is imperative to use a brief packet/code word to ensure low latency, which are factors that reduce reliability performance. Alternatively, IoT applications like industrial robots are particularly prone to security threats due to the broadcast nature of wireless communication. Traditionally, by cryptography, security in the upper layers of the communication system is increased. However, it requires additional channel usage to implement secret key exchange and management protocols, which complicates the structure. In URLLC, the channel block length is limited and therefore the encryption method might not be available in URLLC applications. On the other hand, URLLC is more convenient as there is no need for a complex key exchange procedure in physical layer security. Large antenna systems have outstanding features that may be beneficial with their ability to make an oversized number of spatial degrees of freedom that determine for URLLC [78]. Large antenna systems provide high SNR connections thanks to the array gain. Semi-deterministic connections specific to systems operating below 6GHz in diversity scattering environment are practically the result of channel hardening phenomenon and practically immune to fast fade. However, with the initial property, it alleviates the requirement for strong coding schemes and thus maintains high reliability for brief packets and therefore the need for retransmission may be significantly reduced. Thus, high capacity is provided for spatial division multiplexing.

In multi-user systems, since multiple users can exchange data at the identical time, this feature is accustomed improve latency caused by multiple access. However, this may result in additional computational delay of the multiple antenna processing used to separate users.

VIII. SOLUTIONS IN THE LITERATURE

The basic requirements of mobility management are that HO operations are reliably configurable, have a robust link tracking mechanism, and uninterrupted HO scheme. In order to meet these requirements, some solutions defined in the literature are mentioned in this section and various applications of these solutions are explained. The solutions described in this section are the use of DC in the HO procedure, SDN based mobility management schemes, and conditional HO.

A. DUAL CONNECTIVITY

In order to increase the robustness of the communication systems, DC offered in the 12th release of 3GPP specifications is an important tool. In DC, the UE connects to two RATs at the same time, so if a problem is encountered in the current connection, data flow is carried out over the other connection to ensure the continuity of the communication, thus increasing the mobility performance. In this way, 0 ms HO interruption time, one of the needs of 5G, can be met by DC. Especially for mmWave channels where the connections can deteriorate rapidly, providing DC with cells operating sub-6GHz bands can be an important solution for the continuity of the communication. The mobility robustness provided in this way in cellular systems is called macro-diversity. In addition, it can increase QoS by providing flexible and dynamic solutions for resource allocation in traffic management. For example, it can offer solutions that can program network traffic flexibly by routing small volumes of delay sensitive data traffic to macro BS and large volumes of delay tolerant traffic to small BS.

In [84], a DC-assisted prevenient HO model has been proposed to manage HO operation in high-speed rail scenarios in C/U plane split networks in long-term evaluation-advanced (LTE-A) technology among macro BSs. The presented model consists of a two-level trigger mechanism, a micro-macro HO procedure and a macro-micro HO procedure. In micro-macro HO, the target macro BS is premade secondary eNB, making mobility robust by supporting RRC signal replication in the case of DC for two macro BSs. In macro-micro HO, mobility coordination of C-plane transition is performed by mutual signaling in the X2 interface between macro BSs. Thus, the period of discontinuation of the service is reduced. However, as the model causes additional signal overload, the two-level HO model requires additional BS resources.

In [2], a local anchor-based user-centric network (UCN) architecture, in which the DC technique is applied, is presented. The proposed architecture aims to examine the performance improvement potential of small cell use in UCN. Accordingly, among the small cells, master eNBs and slave eNBs are selected and the local anchor acts as the MeNB while neighboring small cells play the role of SeNB. Among the femtocells, those with strong abilities are chosen as anchor femtocells, and their corresponding clusters called virtual cells (VC) are created according to the geographic distribution of surrounding femtocells. In the proposed architecture, to implement the DC technique, the anchor femtocells act as MeNB, which is responsible for both control and data transmission for UEs, while other femtocells within VC only function as SeNBs that support user plane transmission for UEs. According to the simulation results, the proposed architecture has shown a significant improvement in HOF performance compared to the current LTE system.

An efficient HO algorithm is proposed in [85] that can operate between 4G and 5G RATs. The proposed HO algorithm determines the target cell where HO will occur, not just considering RSS. In addition, it tries to provide better performance to the users by considering the parameters such as the density and coverage of the target cell, the number of users in the target cell, and the status of the radio channel. For multi-RAT DC mobility management with a split data mechanism, a system using a Markov decision process is proposed, where an award-based controller drives HO. Here, the reward mechanism derives a reward function from the parameters listed above. The split bearer mechanism incorporated into the architecture to reduce the frequent HO effects caused by ultra-dense networks and user speeds, provides alternative routes for data packets. When the proposed mechanism is compared with frequent HO mitigation, the HO count decreased.

In [86], a new perspective is presented to ensure energy efficiency and guarantee QoS in 5G HetNets. Accordingly, a dynamic femtocell gNB on/off strategy, together with Markov-based load prediction scheme and cell association, provides load balancing by optimizing the traffic load, thus aiming to maximize network energy efficiency. In addition, a DC-based HO procedure has been proposed to reduce latency during transmission and ensure uninterrupted communication. The simulation results show that the proposed algorithm makes a significant contribution to energy efficiency.

A smart, fast and light-weight HO procedure should provide minimal signaling overhead and HO latency. Accordingly, in [87], a DC radio access network architecture with logical separation between control and data planes is proposed. The proposed scheme attempts to predict future HO events and expected HO time by combining physical proximity information provided by the UE's location information with user context information such as speed, direction, and HO history, and RF performance derived from signal strength and signal quality metrics. The simulation results show that the proposed scheme significantly reduces HO signaling latency compared to the traditional LTE X2 HO procedure.

In [88], the advantages of allowing disaggregated associations in DC scenarios where users can simultaneously consume radio resources of two different service cells are explored. A comprehensive mathematical analysis of the

probability of association as well as the capacity performance metrics for aggregated transmissions is modeled by a meticulous stochastic geometry-based model of HetNet, a two-tier co-channel with flexible associations. It is concluded that the best spectrum aggregation for users in separated regions is where UL and DL are allowed to be split.

B. SDN BASED SOLUTIONS

SDN architecture is a solution to reduce the complexity of today's traditional network structures. SDN provides tools and mechanisms to make applications and services more precise, more flexible, programmable, and more manageable. In addition, SDN promises to mobile operators cost reduction and new applications and services to be released in less time. It separates the SDN control plane from the data plane, providing a global overview and a central control. As the control plane is executed by the SDN controller, the data plane for applications and network services is simplified and abstracted through the SDN controller. When preparing recipes for the SDN controller, general control cell, or HetNet, the SDN switch deals with data transfer and behavior changes on the network by following controller commands. SDN uses the OpenFlow protocol for device level abstraction. Devices within the network become simple routing devices that can be programmed through an open interface implemented by the Openflow protocol. Thus, instead of configuring the network in every device, it is done only on the controller. This enables network resources to be used more efficiently and can provide dynamic optimization of resources and flow management. SDN-based solutions, which are presented to meet the needs that occur with 5G, have an important place in the literature.

In [89], a seamless mobility management scheme based on the principle of tracking users' movements is proposed using the distributed hash table (DHT). In this architecture, there is a master and slave relationship mechanism between distributed SDN controllers. Accordingly, the slave SDN controller updates the DHT table periodically or dynamically depending on the network status, and the master SDN determines the decisions of the interconnection network by making calculations on the routing information. Simulation results show that the proposed framework provides better performance in terms of latency and signal overhead compared to proxy mobile IPv6 and hierarchical mobile IPv6.

In [90], an SDN based HO scheme is proposed to provide communication with minimum delay, which is one of the basic needs of 5G. The SDN controller collects BS status information such as the target cells' load density and the signal strength, as well as mobility information of the UE, such as the UE's direction and estimated sojourn time, and processes this information to perform cell selection. Linear programming is used in the designed scheme to reduce computational complexity and thus minimize overall HO delay time. The simulation results show that the proposed method is successful in finding cells with strong signal strength, long sojourn time and lightly loaded according to the direction of movement of the UE. In [91], it is aimed to select the most suitable cell for an uninterrupted HO operation by using fuzzy analytical hierarchy process (FAHP) and multipath transmission control protocol (MPTCP) in SDN based HetNets according to service needs and user preferences. While FAHP is used to select the most suitable cell with the data it receives from SDN, MPTCP provides the assurance of uninterrupted connection. Simulation results showed that the proposed method reduces HO times and provides uninterrupted HO. In [92], an SDN/NFV based network architecture that can be used in 5G ultra dense networks is proposed. The proposed network architecture consists of 3 planes: data plane consisting of several small cells that provide traffic to users and transmit measurement reports to controllers, control plane with SDN controller that controls and manages the mobile mobility of RATs and mobile nodes, and application plane consisting of programs keeping the network abstract to determine future behavior according to network requirements. In the proposed architecture, software defined HO management engine (SDHME) is designed to be responsible for HO management in 5G networks. In SDHME, the target network selection for each mobility node is determined according to the application's QoS requirements, the state of the mobility node, and the network conditions provided by the control plane. The simulation results show that the proposed scheme significantly reduces the HOF rate and HO latency compared to the traditional LTE HO strategy.

In [93], the HO management framework that allows the use of HO algorithms that can use multiple metrics in decision-making processes, centralized, proactive in order to have appropriate mobility management in IEEE 802.11 Wi-Fi networks is presented. The HO management framework is an SDN-based management framework that runs in a real-life test environment and can be created, tested and evaluated by large-scale HO algorithms. Based on the HO management framework 's capabilities, a machine learning algorithm called ABRAHAM has been created to predict the future load status of access points and future client station location, and estimate the future received signal strength indicator value. Simulation results showed that the proposed algorithm provides an average efficiency increase of up to 139% compared to the IEEE 802.11 standard HO algorithm.

The frequent occurrence of HO transactions with 5G leads to security issues and delays that conflict with 5G targets if authentication procedures are performed inefficiently. For this purpose, a new authentication approach using blockchain and SDN techniques is proposed in [94] to remove re-authentication on repeated HO. With the proposed approach, users' privacy is protected and a fast and secure connection is achieved by eliminating the duplicate authentication procedure. The results show that the proposed model reduces latency compared to similar models. In [95], a capability-based privacy protection HO authentication mechanism has been proposed to address the common HO authentication issue in SDN-based 5G HetNets. The authentication HO module is integrated into the SDN controller to monitor users and predict their future location and pre-prepare cells

or select suitable cells in advance. In the proposed scheme, mutual authentication is performed between the UE and the BSs, without any other BSs or protocols being activated, so that authentication is performed at less cost. Besides, the proposed scheme has been shown to provide security protection through different safety tests. The results showed that the proposed scheme reduced both communication cost and computation cost compared to the standard HO scheme.

In [96], a software-defined IoT system called UbiFlow is introduced for unified ubiquitous flow control and mobility management in urban HetNets. UbiFlow uses multiple controllers to divide urban-scale SDN into different geographic segments and provide distributed control of IoT flows. Distributed controllers perform tasks such as managing mobility in coordination with each other, optimizing HO, determining the access point, and flow scheduling. As a result of comparing the UbiFlow system with the existing systems using SDN in the literature, it was seen that efficiency increased by 67.21 %, delay decreased by 72.99 % and vibration increased by 69.59 %.

C. CONDITIONAL HO

CHO is a method introduced in NR 3GPP version 16 and used to ensure robust communication. RLF can be caused by poor conditions in the communication links, errors from synchronization, and malfunctions that can occur while running random access procedures [97]. Also, malfunction may occur due to the timing of the HO command and measurement reports not being received by the network. To deal with such situations, CHO is a promising solution, but the excessive signaling load should be considered and HO should be well parameterized. The flow diagram of the CHO is as in Figure 15. In the CHO, the HO command is given a little early, that is, when the connection between the serving cell and the UE is strong, and access to the target cell is done when the connection between the UE is at a sufficient level. Thus, HO success is increased.

In [97], even with the simplest CHO scheme that can be prepared, it was observed that user mobility increased and interruptions decreased compared to inter-frequency HO. As mmWave communication is vulnerable to blockages, a new prediction-based CHO scheme using deep learning technology has been proposed in [99] to tolerate sudden changes in signal reception power and thus increase the robustness of the CHO. The results show that the proposed scheme can increase the early preparation success rate and reduce the signaling overhead compared to existing CHO schemes. In [100], the mobility performance of CHO was analyzed for 5G mmWave systems with beamforming systems. In addition, a random access procedure has been proposed that increases the chances of contention-free random access during CHO and reduces signaling and downtime. The UE initiates random access by sending a RACH preamble to the target cell, but RACH collision occurs when multiple UEs use the same preamble during random access towards the same receive beam of a target cell, which is then further

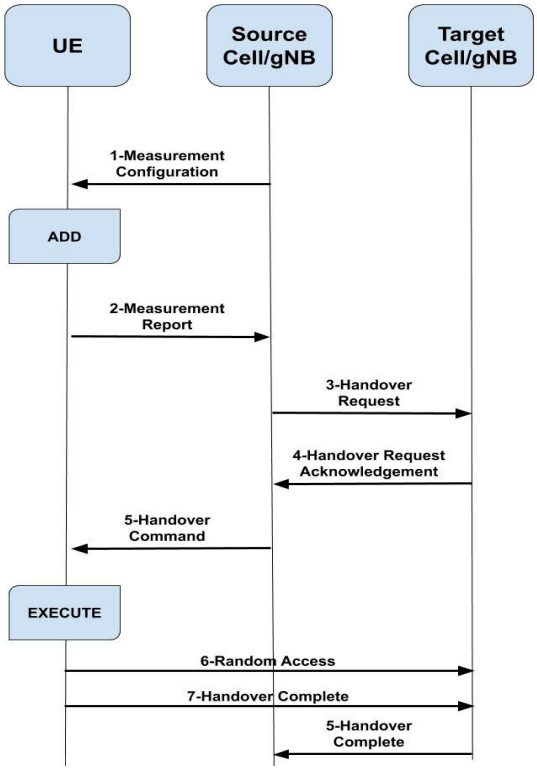

FIGURE 15. CHO procedure [98].

signaling and randomization resolved by the delay to complete access. In order to eliminate the risk of collisions in the HO, the radio access that defines the UE signal without further signal and delay if the network reaches the prepared beam using the allocated preamble of the UE is called contention-free radio access. The simulation results showed that with the proposed scheme, overall failure performance was improved and the number of reverts to contention-based random access was significantly reduced.

IX. CONCLUSION

In this paper, a comprehensive survey has been discussed the details of 5G mobility management. It has highlighted most important parts that provides a better understanding of the mobility management in recent generation such as requirements, architecture and challenges. This paper has been further discussed the newly introduced RRC inactive status, initial access and procedure of registration and paging. In addition, inter-RAN HO procedures in connected state and integrated mmWave cells with 5G network have been explained with a details in literature survey. Finally, several challenges have been addressed such as HO issues, signaling overhead, power consumption, security and latency. In this regard, some effective solutions have been demonstrated to meet the requirement of 5G mobility management. We believe that this survey will provide the researchers guidelines and a good platform to further their research in the same scope of this survey.

TABLE 3. Summary of abbreviations.

Abbreviation	Definition
3GPP	Third generation partnership project
4G	Fourth generation
5G	Fifth generation
5GC	Fifth generation core
AHP	Analytical hierarchy processing
AMF	Access and mobility management function
ARIMA	Autoregressive integrated moving average
AS	Access stratum
BS	Base station
CDR	Call detail records
CHO	Conditional handover
CIO	Cell individual offset
CMOS	Complementary metal-oxide-semiconductor
CN	Core network
CSI-RS	Channel status information reference signal
DC	Dual connectivity
DHT	Distributed hash table
DRX	Discontinuous cycle
EAP-TLS	Extensible authentication protocol transport layer security
eMBB	Enhanced mobile broadband
eNB	Evolved NodeB
EPC	Evolved packet core
ERG	Energy reduction gain
FAHP	Fuzzy analytical hierarchy process
gNB	Next-generation NodeB
GRA	Grey rational analysis
HetNet	Heterogeneous network
HO	Handover
HOF	Handover failure
HOM	Handover margin
IoT	Internet of things
LBEF	Load balancing efficiency factor
LTE	Long term evaluation
M2M	Machine to machine
MAC	Medium access network
MADM	Multiple attribute decision making
MEC	Multi-acccess edge computing
MIMO	Multiple-input-multiple-output
MME	Mobility management entity
mMTC	Massive machine type communication
mmWave	Milimeter wave
MPTCP	Multipath transmission control protocol
MRN	Mobile relay node
N3IWF	Non-3GPP interworking function
NAS	Non access stratum
NFV	Network functions virtualization
NG-AP	Next generation application protocol
NReNB	New radio evolved NodeB
OFDM	Orthogonal frequency division multiplexing
OPEX	Operational expenditure
PBCH	Physical broadcast channel
PCH	Paging channel
PDCP	Packet data convergence protocol
PDU	Protocol date unit
PGW-C	Packet data network control plane function
PLMN	Public land mobile network
PPHO	Ping-pong handover
PSS	Primary signal synchronization
QoS	Quality of service
RA	Registration area
RACH	Random access channel
RAN	Radio access network
RAT	Radio access technology
RLF	Radio link failure
RNN	Repetative neural network
RRC	Radio resource control
RRM	Radio resource management
SDHME	Software defined handover management engine
SDN	Software defined network
SGW	Serving gateway
SINR	Signal-to-interference-plus-noise ratio
SMF	Session management function
SS	Signal synchronization
SRS	Sounding reference signal
SSB	Synchronization signal block
SSS	Secondary signal synchronization
TA	Tracking area
TAL	Tracking area list
TAU	Tracking area update
TOPSIS	Technique for order preference by similarity to ideal solution
TRP	Transmit receive point
TTT	Time to trigger
UAV	Unmanned aerial vehicle
UCN	User-centric network
UE	User equipment
UMLB	Utility based mobility load balancing
UMTS	Universal mobile telecommunication system
UPF	User plane function
URLLC	Ultra reliable low latency communication
VANET	Vehicular ad hoc network
VC	Virtual cell
VIKOR	Vise Kriteri-jumska Optimizacija IKompromisno Resenje
WSN	Wireless sensor network
XGBoost	eXtreme Gradient Boosting

APPENDIX

The abbreviations and acronyms used are first introduced in the text and, for convenience, the list of abbreviations used in this article is summarized in Table 3.

REFERENCES

[1] P. Cerwall, P. Jonsson, R. Möller, S. Bävertoft, S. Carson, and I. Godor, "Ericsson mobility report. On the pulse of the networked society," Hg. v. Ericsson, White Paper, Jun. 2015. [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2015/ericsson-mobility-report-june-2015.pdf

[2] H. Zhang, N. Meng, Y. Liu, and X. Zhang, "Performance evaluation for local anchor-based dual connectivity in 5G user-centric network," *IEEE Access*, vol. 4, pp. 5721–5729, 2016.

[3] R. Tiwari and S. Deshmukh, "MVU estimate of user velocity via gamma distributed handover count in HetNets," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 482–485, Mar. 2019.

[4] M. M. Hasan, S. Kwon, and S. Oh, "Frequent-handover mitigation in ultra-dense heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 1035–1040, Jan. 2019.

[5] X. Xu, X. Tang, Z. Sun, X. Tao, and P. Zhang, "Delay-oriented cross-tier handover optimization in ultra-dense heterogeneous networks," *IEEE Access*, vol. 7, pp. 21769–21776, 2019.

[6] Z. Zhang, Z. Junhui, S. Ni, and Y. Gong, "A seamless handover scheme with assisted eNB for 5G C/U plane split heterogeneous network," *IEEE Access*, vol. 7, pp. 164256–164264, 2019.

[7] M. Alhabo, L. Zhang, and N. Nawaz, "GRA-based handover for dense small cells heterogeneous networks," *IET Commun.*, vol. 13, no. 13, pp. 1928–1935, Aug. 2019.

[8] K. Vasudeva, M. Simsek, D. Lopez-Perez, and I. Guvenc, "Impact of channel fading on mobility management in heterogeneous networks," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2206–2211.

[9] A. S. Cacciapuoti, "Mobility-aware user association for 5G mmWave networks," *IEEE Access*, vol. 5, pp. 21497–21507, 2017.

[10] Y. Koda, K. Nakashima, K. Yamamoto, T. Nishio, and M. Morikura, "Handover management for mmWave networks with proactive performance prediction using camera images and deep reinforcement learning," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 2, pp. 802–816, Jun. 2020.

[11] K. Zhu, D. Niyato, P. Wang, E. Hossain, and D. In Kim, "Mobility and handoff management in vehicular networks: A survey," *Wireless Commun. Mobile Comput.*, vol. 11, no. 4, pp. 459–476, Apr. 2011.

[12] S. Fernandes and A. Karmouch, "Vertical mobility management architectures in wireless networks: A comprehensive survey and future directions," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 1, pp. 45–63, 1st Quart., 2012.

[13] B. R. Chandavarkar and G. R. M. Reddy, "Survey paper: Mobility management in heterogeneous wireless networks," *Procedia Eng.*, vol. 30, pp. 113–123, Mar. 2012, doi: 10.1016/j.proeng.2012.01.841.

[14] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility management for femtocells in LTE-advanced: Key aspects and survey of handover decision algorithms," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 64–91, 1st Quart., 2014.

[15] Y. Gu, F. Ren, Y. Ji, and J. Li, "The evolution of sink mobility management in wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 507–524, 1st Quart., 2016.

[16] V. Kumar and P. K. Dahiya, "Mobility management in vehicular adhoc networks: A review," *IOSR J. Electron. Commun. Eng.*, vol. 11, no. 1, pp. 85–96, 2016.

[17] S. M. Ghaleb, S. Subramaniam, Z. A. Zukarnain, and A. Muhammed, "Mobility management for IoT: A survey," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, p. 165, Dec. 2016.

[18] M. Mehrabi, H. Salah, and F. H. P. Fitzek, "A survey on mobility management for MEC-enabled systems," in *Proc. IEEE 2nd 5G World Forum (5GWF)*, Sep. 2019, pp. 259–263.

[19] M. Tayyab, X. Gelabert, and R. Jantti, "A survey on handover management: From LTE to NR," *IEEE Access*, vol. 7, pp. 118907–118930, 2019.

[20] A. A. R. Alsaeedy and E. K. P. Chong, "A review of mobility management entity in LTE networks: Power consumption and signaling overhead," *Int. J. Netw. Manage.*, vol. 30, no. 1, p. e2088, Jan. 2020.

[21] N. Akkari and N. Dimitriou, "Mobility management solutions for 5G networks: Architecture and services," *Comput. Netw.*, vol. 169, Mar. 2020, Art. no. 107082.

[22] S. Hailu, M. Saily, and O. Tirkkonen, "RRC state handling for 5G," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 106–113, Jan. 2019.

[23] J.-H. Lee, K. S. Sung, S. Kang, and J. Shin, "A study of the radio resource control connection re-establishment procedure on the UE side in 3GPP," in *Proc. 17th Int. Conf. Adv. Commun. Technol. (ICACT)*, Jul. 2015, pp. 260–262.

[24] *Nr; Radio Resource Control (RRC); Protocol Specification*, document TS 38.331, 3GPP, 2018.

[25] S. Yi, S. Chun, Y. Lee, S. Park, and S. Jung, *Radio Protocols for LTE and LTE-Advanced*. Hoboken, NJ, USA: Wiley, 2012.

[26] A. Khlass, D. Laselva, and R. Jarvela, "On the flexible and performance-enhanced radio resource control for 5G NR networks," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, pp. 1–6.

[27] S. Rommer, P. Hedman, M. Olsson, L. Frid, S. Sultana, and A. Mulligan, *5G Core Networks: Powering Digitalization*. New York, NY, USA: Academic, 2020.

[28] Y.-B. Lin, R.-H. Liou, and C.-T. Chang, "A dynamic paging scheme for long-term evolution mobility management," *Wireless Commun. Mobile Comput.*, vol. 15, no. 4, pp. 629–638, Mar. 2015.

[29] N. M. Akshatha, P. Jha, and A. Karandikar, "A centralized SDN architecture for the 5G cellular network," in *Proc. IEEE 5G World Forum (5GWF)*, Jul. 2018, pp. 147–152.

[30] S. Hailu and M. Saily, "Hybrid paging and location tracking scheme for inactive 5G UEs," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–6.

[31] *NR and NG-Ran Overall Description; Stage 2 (Release 15)*, 3GPP, Sophia Antipolis, France, 2017.

[32] *Procedures for the 5G System (5GS)*, document TS 23.502, 3GPP, 2019.

[33] 3GPP, *5G; Procedures for the 5G System (5GS)*, document 3GPP TS 23.502 version 15.4.1 Release 15, 2019.

[34] J. S. Kim, W. J. Lee, and M. Y. Chung, "A multiple beam management scheme on 5G mobile communication systems for supporting high mobility," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2016, pp. 260–264.

[35] J. Mietzner, R. Schober, L. Lampe, W. Gerstacker, and P. Hoeher, "Multiple-antenna techniques for wireless communications–a comprehensive literature survey," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 2, pp. 87–105, 2nd Quart., 2009.

[36] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, C.-L. I, and A. Ghosh, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.

[37] T. Obara, S. Suyama, J. Shen, and Y. Okumura, "Joint processing of analog fixed beamforming and CSI-based precoding for super high bit rate massive MIMO transmission using higher frequency bands," *IEICE Trans. Commun.*, vol. 98, no. 8, pp. 1474–1481, 2015.

[38] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, 1st Quart., 2019.

[39] T. E. Bogale and L. B. Le, "Beamforming for multiuser massive MIMO systems: Digital versus hybrid analog-digital," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 4066–4071.

[40] Evolved Universal Terrestrial Radio Access. (2015). *Radio Resource Control (RRC) Protocol Specification*. [Online]. Available: http://www.3gpp.org

[41] Y.-N.-R. Li, B. Gao, X. Zhang, and K. Huang, "Beam management in millimeter-wave communications for 5G and beyond," *IEEE Access*, vol. 8, pp. 13282–13293, 2020.

[42] E. Onggosanusi, M. S. Rahman, L. Guo, Y. Kwak, H. Noh, Y. Kim, S. Faxer, M. Harrison, M. Frenne, S. Grant, R. Chen, R. Tamrakar, and A. Q. Gao, "Modular and high-resolution channel state information and beam management for 5G new radio," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 48–55, Mar. 2018.

[43] *Evolved Universal Terrestrial Radio Access. User Equipment (UE) Radio Transmission and Reception*, document TS 36:V10, 3GPP, 2011.

[44] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "Standalone and non-standalone beam management for 3GPP NR at mmWaves," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 123–129, Apr. 2019.

[45] I. Shayea, T. Abd. Rahman, M. Hadri Azmi, and A. Arsad, "Rain attenuation of millimetre wave above 10 GHz for terrestrial links in tropical regions," *Trans. Emerg. Telecommun. Technol.*, vol. 29, no. 8, p. e3450, Aug. 2018.

[46] J. Lu, D. Steinbach, P. Cabrol, and Pietraski, "Modeling the impact of human blockers in millimeter wave radio links," *ZTE Commun. Mag.*, vol. 10, no. 4, pp. 23–28, Dec. 2012.

[47] M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Multi-connectivity in 5G mmWave cellular networks," in *Proc. Medit. Ad Hoc Netw. Workshop (Med-Hoc-Net)*, Jun. 2016, pp. 1–7.

[48] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2069–2084, Sep. 2017.

[49] A. Mazin, M. Elkourdi, and R. D. Gitlin, "Accelerating beam sweeping in mmWave standalone 5G new radios using recurrent neural networks," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–4.

[50] L. Li, D. Wang, X. Niu, Y. Chai, L. Chen, L. He, X. Wu, F. Zheng, T. Cui, and X. You, "mmWave communications for 5G: Implementation challenges and advances," *Sci. China Inf. Sci.*, vol. 61, no. 2, Feb. 2018, Art. no. 021301.

[51] H. Soleimani, R. Parada, S. Tomasin, and M. Zorzi, "Fast initial access for mmWave 5G systems with hybrid beamforming using online statistics learning," *IEEE Commun. Mag.*, vol. 57, no. 9, pp. 132–137, Sep. 2019.

[52] W. Attaoui, K. Bouraqia, E. Sabir, M. Benjillali, and R. Elazouzi, "Beam alignment game for self-organized mmWave-empowered 5G initial access," in *Proc. 15th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jun. 2019, pp. 2050–2057.

[53] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37328–37348, 2018.

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

[55] P. Han, Z. Zhou, and Z. Wang, "User association for load balance in heterogeneous networks with limited CSI feedback," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1095–1099, May 2020.

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

[57] A. Mohajer, M. Bavaghar, and H. Farrokhi, "Mobility-aware load balancing for reliable self-organization networks: Multi-agent deep reinforcement learning," *Rel. Eng. Syst. Saf.*, vol. 202, Oct. 2020, Art. no. 107056.

[58] B. Ma, B. Yang, Y. Zhu, and J. Zhang, "Context-aware proactive 5G load balancing and optimization for urban areas," *IEEE Access*, vol. 8, pp. 8405–8417, 2020.

[59] J. Hu, H. Zhang, Y. Liu, X. Li, and H. Ji, "An intelligent UAV deployment scheme for load balance in small cell networks using machine learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–6.

[60] M. Karaatlı, N. Ömürbek, and G. Köse, "Analitik hiyerarşi süreci temelli topsis ve vikor yöntemleri ile futbolcu performanslarinin değerlendirilmesi," *Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, vol. 29, no. 1, pp. 25–61, 2014.
[61] M. Alhabo and L. Zhang, "Multi-criteria handover using modified weighted TOPSIS methods for heterogeneous networks," *IEEE Access*, vol. 6, pp. 40547–40558, 2018.
[62] T. Bilen, B. Canberk, and K. R. Chowdhury, "Handover management in software-defined ultra-dense 5G networks," *IEEE Netw.*, vol. 31, no. 4, pp. 49–55, Jul. 2017.
[63] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, and A. Alquhali, "Velocity-aware handover self-optimization management for next generation networks," *Appl. Sci.*, vol. 10, no. 4, p. 1354, Feb. 2020.
[64] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, S. Alraih, and K. S. Mohamed, "Auto tuning self-optimization algorithm for mobility management in LTE—A and 5G HetNets," *IEEE Access*, vol. 8, pp. 294–304, 2020.
[65] N. Malm, L. Zhou, E. Menta, K. Ruttik, R. Jantti, O. Tirkkonen, M. Costa, and K. Leppanen, "User localization enabled ultra-dense network testbed," in *Proc. IEEE 5G World Forum (5GWF)*, Jul. 2018, pp. 405–409.
[66] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, "Predictive and core-network efficient RRC signalling for active state handover in RANs with Control/Data separation," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1423–1436, Mar. 2017.
[67] A. A. R. Alsaeedy and E. K. P. Chong, "Mobility management for 5G IoT devices: Improving power consumption with lightweight signaling overhead," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8237–8247, Oct. 2019.
[68] P. Skrimponis, S. Dutta, M. Mezzavilla, S. Rangan, S. H. Mirfarshbafan, C. Studer, J. Buckwalter, and M. Rodwell, "Power consumption analysis for mobile mmWave and sub-THz receivers," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, Mar. 2020, pp. 1–5.
[69] S. Verbrugge, S. Pasqualini, F.-J. Westphal, M. Jager, A. Iselt, A. Kirstadter, R. Chahine, D. Colle, M. Pickavet, and P. Demeester, "Modeling operational expenditures for telecom operators," in *Proc. Conf. Opt. Netw. Design Modeling*, 2005, pp. 455–466.
[70] K. Vasudeva, S. Dikmese, I. Guven, A. Mehbodniya, W. Saad, and F. Adachi, "Fuzzy-based game theoretic mobility management for energy efficient operation in HetNets," *IEEE Access*, vol. 5, pp. 7542–7552, 2017.
[71] B. Zhang, W. Qi, and J. Zhang, "An energy efficiency and ping-pong handover ratio optimization in two-tier heterogeneous networks," in *Proc. IEEE 8th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2018, pp. 532–536.
[72] A. Mukherjee, "Energy efficiency and delay in 5G ultra-reliable low-latency communications system architectures," *IEEE Netw.*, vol. 32, no. 2, pp. 55–61, Mar. 2018.
[73] J. Huang and Y. Qian, "A secure and efficient handover authentication and key management protocol for 5G networks," *J. Commun. Inf. Netw.*, vol. 5, no. 1, pp. 40–49, 2020.
[74] Y. Zhang, R. Deng, E. Bertino, and D. Zheng, "Robust and universal seamless handover authentication in 5G HetNets," *IEEE Trans. Dependable Secure Comput.*, early access, Jul. 9, 2019, doi: 10.1109/TDSC.2019.2927664.
[75] R. Ma, J. Cao, D. Feng, H. Li, and S. He, "FTGPHA: Fixed-trajectory group pre-handover authentication mechanism for mobile relays in 5G high-speed rail networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 2126–2140, Feb. 2020.
[76] K. Zhang, X. Xu, J. Zhang, B. Zhang, X. Tao, and Y. Zhang, "Dynamic multiconnectivity based joint scheduling of eMBB and uRLLC in 5G networks," *IEEE Syst. J.*, early access, Apr. 6, 2020, doi: 10.1109/JSYST.2020.2977666.
[77] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Resource allocation for secure URLLC in mission-critical IoT scenarios," 2019, *arXiv:1911.13154*. [Online]. Available: [Online]. Available: http://arxiv.org/abs/1911.13154
[78] P. Popovski, C. Stefanovic, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.
[79] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, and S. Alraih, "Dynamic handover control parameters for LTE–A/5G mobile communications," in *Proc. Adv. Wireless Opt. Commun. (RTUWO)*, Nov. 2018, pp. 39–44.
[80] A. Abdulraqeb, R. Mardeni, A. M. Yusoff, S. Ibraheem, and A. Saddam, "Self-optimization of handover control parameters for mobility management in 4G/5G heterogeneous networks," *Autom. Control Comput. Sci.*, vol. 53, no. 5, pp. 441–451, Sep. 2019.
[81] I. Shayea, M. Ismail, R. Nordin, M. Ergen, N. Ahmad, N. F. Abdullah, A. Alhammadi, and H. Mohamad, "New weight function for adapting handover margin level over contiguous carrier aggregation deployment scenarios in LTE-advanced system," *Wireless Pers. Commun.*, vol. 108, no. 2, pp. 1179–1199, Sep. 2019.
[82] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, S. Alriah, and A. B. Abas, "Advanced handover self-optimization approach for 4G/5G HetNets using weighted fuzzy logic control," in *Proc. 15th Int. Conf. Telecommun. (ConTEL)*, Jul. 2019, pp. 1–6.
[83] J. Zhang, Y. Zeng, and R. Zhang, "Spectrum and energy efficiency maximization in UAV-enabled mobile relaying," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[84] P.-J. Hsieh, W.-S. Lin, K.-H. Lin, and H.-Y. Wei, "Dual-connectivity prevenient handover scheme in Control/User-plane split networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3545–3560, Apr. 2018.
[85] T. Mumtaz, S. Muhammad, M. I. Aslam, and N. Mohammad, "Dual connectivity-based mobility management and data split mechanism in 4G/5G cellular networks," *IEEE Access*, vol. 8, pp. 86495–86509, 2020.
[86] X. Huang, S. Tang, Q. Zheng, D. Zhang, and Q. Chen, "Dynamic femtocell gNB On/Off strategies and seamless dual connectivity in 5G heterogeneous cellular networks," *IEEE Access*, vol. 6, pp. 21359–21368, 2018.
[87] A. Mohamed, M. A. Imran, P. Xiao, and R. Tafazolli, "Memory-full context-aware predictive mobility management in dual connectivity 5G networks," *IEEE Access*, vol. 6, pp. 9655–9666, 2018.
[88] M. A. Lema, E. Pardo, O. Galinina, S. Andreev, and M. Dohler, "Flexible dual-connectivity spectrum aggregation for decoupled uplink and downlink access in 5G heterogeneous systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 11, pp. 2851–2865, Nov. 2016.
[89] A. S. D. Alfoudi, S. H. S. Newaz, R. Ramlie, G. M. Lee, and T. Baker, "Seamless mobility management in heterogeneous 5G networks: A coordination approach among distributed SDN controllers," in *Proc. IEEE 89th Veh. Technol. Conf. (VTC-Spring)*, Apr. 2019, pp. 1–6.
[90] J. Lee and Y. Yoo, "Handover cell selection using user mobility information in a 5G SDN-based network," in *Proc. 9th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2017, pp. 697–702.
[91] H. Tong, X. Liu, and C. Yin, "A FAHP and MPTCP based seamless handover method in heterogeneous SDN wireless networks," in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2019, pp. 1–6.
[92] A. Gharsallah, F. Zarai, and M. Neji, "SDN/NFV-based handover management approach for ultradense 5G mobile networks," *Int. J. Commun. Syst.*, vol. 32, no. 17, p. e3831, Nov. 2019.
[93] E. Zeljkovic, N. Slamnik-Krijestorac, S. Latre, and J. M. Marquez-Barja, "ABRAHAM: Machine learning backed proactive handover algorithm using SDN," *IEEE Trans. Netw. Service Manage.*, vol. 16, no. 4, pp. 1522–1536, Dec. 2019.
[94] A. Yazdinejad, R. M. Parizi, A. Dehghantanha, and K.-K.-R. Choo, "Blockchain-enabled authentication handover with efficient privacy protection in SDN-based 5G networks," *IEEE Trans. Netw. Sci. Eng.*, early access, Aug. 28, 2019, doi: 10.1109/TNSE.2019.2937481.
[95] J. Cao, M. Ma, Y. Fu, H. Li, and Y. Zhang, "CPPHA: Capability-based privacy-protection handover authentication mechanism for SDN-based 5G HetNets," *IEEE Trans. Dependable Secure Comput.*, early access, May 14, 2019, doi: 10.1109/TDSC.2019.2916593.
[96] D. Wu, X. Nie, E. Asmare, D. I. Arkhipov, Z. Qin, R. Li, J. A. McCann, and K. Li, "Towards distributed SDN: Mobility management and flow scheduling in software defined urban IoT," *IEEE Trans. Parallel Distrib. Syst.*, vol. 31, no. 6, pp. 1400–1418, Jun. 2020.
[97] H. Martikainen, I. Viering, A. Lobinger, and T. Jokela, "On the basics of conditional handover for 5G mobility," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1–7.
[98] D. Chandramouli *et al.*, "Access control and mobility management," in *5G for the Connected World*. Hoboken, NJ, USA: Wiley, 2019, pp. 225–282.

[99] C. Lee, H. Cho, S. Song, and J.-M. Chung, "Prediction-based conditional handover for 5G mm-wave networks: A deep-learning approach," *IEEE Veh. Technol. Mag.*, vol. 15, no. 1, pp. 54–62, Mar. 2020.
[100] U. Karabulut, A. Awada, I. Viering, A. Noll Barreto, and G. P. Fettweis, "Analysis and performance evaluation of conditional handover in 5G beamformed systems," 2019, *arXiv:1910.11890*. [Online]. Available: http://arxiv.org/abs/1910.11890

EMRE GURES received the B.Sc. degree in electrical and electronic engineering from Istanbul University, Istanbul, Turkey, in 2017, and the M.Sc. degree in telecommunication engineering from Istanbul Technical University, Istanbul, in 2020, where he is currently pursuing the Ph.D. degree, under the supervision of Dr. I. Shayea. His current research interests include in mobility management, handover, and load balancing.

IBRAHEEM SHAYEA (Member, IEEE) received the B.Sc. degree in electronic engineering from the University of Diyala, Baqubah, Iraq, in 2004, and the M.Sc. degree in computer and communication engineering and the Ph.D. degree in mobile communication engineering from the National University of Malaysia, Universiti Kebangsaan Malaysia (UKM), Malaysia, in 2010 and 2015, respectively. From January 2011 to February 2014, he worked as a Research and Teaching Assistant with UKM. From January 2016 to June 2018, he joined the Wireless Communication Center (WCC), University of Technology Malaysia (UTM), Malaysia, where he worked as a Research Fellow. He has been a Researcher Fellow with Istanbul Technical University (ITU), Istanbul, Turkey, since September 2018. His main research interests include wireless communication systems, mobility management, radio propagation, and the Internet of Things (IoT).

ABDULRAQEB ALHAMMADI (Graduate Student Member, IEEE) received the B.Eng. degree in electronic majoring in telecommunications, the M.S. degree, and the Ph.D. degree in wireless communication from Multimedia University, Malaysia, in 2011, 2015, and 2020, respectively. He served as a Research Assistant/Research Scholar with Multimedia University from 2012 to 2019. He is also the author of more than 20 articles in international journals and conferences. His main research interests include heterogeneous networks, mobility management, D2D communication, cognitive radio networks, localization, and propagation modeling. He is a member of professional institutes and societies, such as IEICE, IACSIT, and IAENG. He is also a member of program committees at international conferences and workshops. He was a recipient of several awards, including the Excellent Researcher Award 2019, Multimedia University.

MUSTAFA ERGEN received the B.S. degree in electrical engineering as a Valedictorian from Orta Doğu Technical University and completed four programs at UC Berkeley: the M.S. and Ph.D. degrees in electrical engineering and the M.A. degree from international studies and MOT program from the HAAS Business School. He is currently a Professor of electrical engineering with Istanbul Technical University, and the President of venture funded Ambeent Inc., focusing 5G and Artificial Intelligence. He also served in the board of trustees of TOBB University. He co-founded Silicon Valley startup WiChorus Inc., to focus on 4G technologies and company is acquired by Tellabs [now Coriant]. Earlier, he was a National Semiconductor Fellow [now TI] with the University of California at Berkeley, where he also co-founded the Distributed Sensing Laboratory, focusing on statistical sensor intelligence and vehicular communication. He has more than 40 patent applications, many publications, and authored three books *Girişimci Kapital: Silikon Vadisi Tarihi ve Startup Ekonomisi* (Third Edition-Alfa, 2018), *Mobile Broadband: Including WiMAX and LTE* (Springer, 2009), and *Multi Carrier Digital Communications: Theory and Applications of OFDM* (Springer, 2004). He is the founding member in 5G Infrastructure Association of European Union and an Advisor at Berkeley Program on Entrepreneurship and Development.

HAFIZAL MOHAMAD (Senior Member, IEEE) received the B.Eng. degree (Hons.) and the Ph.D. degree in electronic engineering from the University of Southampton, U.K., in 1998 and 2003, respectively. He is currently a Professor with the Faculty of Engineering and Built Environment, Universiti Sains Islam Malaysia. He is the co-inventor of 36 filed patents and nine granted patents in the field of wireless communication. He has coauthored more than 100 research articles. His research interests include wireless communication, cognitive radio, mesh networks, and the Internet of Things. He was a recipient of several awards, including the ASEAN Outstanding Scientist and Technologist Award (AOSTA) and Top Research Scientist Malaysia (TRSM) by Academy of Sciences Malaysia. He has served in various leadership roles in the IEEE, including the Vice Chair for IEEE Malaysia Section in 2013 and the Chair for IEEE ComSoc/VT Chapter from 2009 to 2011. He was the Conference Operation Chair of the IEEE ICC 2016 and the Technical Program Chair of the IEEE VTC Spring 2019 and APCC 2011. He is a registered Professional Engineer with the Board of Engineers Malaysia. He is also an Enthusiastic Supporter of industrial and academic liaison. He is also appointed as an Expert panel for various technical committees.