IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 4, 2019, accepted August 18, 2019, date of publication August 26, 2019, date of current version September 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2937405

A Survey on Handover Management: From LTE to NR

MUHAMMAD TAYYAB[1,3], XAVIER GELABERT[2], AND RIKU JÄNTTI[3], (Senior Member, IEEE)

[1]Huawei Technologies Finland Oy, 00180 Helsinki, Finland
[2]Huawei Technologies Sweden AB, 164 94 Kista, Sweden
[3]Department of Communications and Networking, School of Electrical Engineering, Aalto University, 00076 Espoo, Finland

Corresponding author: Muhammad Tayyab (muhammad.tayyab5@huawei.com)

This work was supported by the European Union's Horizon 2020 Research and Innovation Program under Grant H2020-MSCA-ITN-2016 SECRET-722424 [12].

ABSTRACT To satisfy the high data demands in future cellular networks, an ultra-densification approach is introduced to shrink the coverage of base station (BS) and improve the frequency reuse. The gain in capacity is expected but at the expense of increased interference, frequent handovers (HOs), increased HO failure (HOF) rates, increased HO delays, increase in ping pong rate, high energy consumption, increased overheads due to frequent HO, high packet losses and bad user experience mostly in high-speed user equipment (UE) scenarios. This paper presents the general concepts of radio access mobility in cellular networks with possible challenges and current research focus. In this article, we provide an overview of HO management in long-term evolution (LTE) and 5G new radio (NR) to highlight the main differences in basic HO scenarios. A detailed literature survey on radio access mobility in LTE, heterogeneous networks (HetNets) and NR is provided. In addition, this paper suggests HO management challenges and enhancing techniques with a discussion on the key points that need to be considered in formulating an efficient HO scheme.

INDEX TERMS Radio access mobility, cell selection, handover, LTE, 5G, NR, mobility enhancers.

I. INTRODUCTION

There has been an exponential growth in mobile data usage over the last 15 years (over 400 million fold) that is expected to go up nearly 6-fold between 2017–2022 reaching 77 Exabyte per month by 2022 [1]. A key approach supported by 5G is ultra-cell-densification to satisfy the high data traffic demands. The spectral efficiency of the 5G network may be largely improved by shrinking the coverage of base stations (BSs), thus reducing the number of users served by each BS and improving the frequency reuse. However, a clear impact of densification planning is increased handover (HO) rate, i.e. the successive change of handling BS for a moving user. Using this approach, the gain in capacity is achieved but at the cost of increased HO rates and higher signaling overheads caused by HO procedure. The signaling overheads interrupt the data flow and thus reduce the throughput of the users [2].

Future cellular networks need to support data-hungry applications with enhanced data rates possibly via cell densification (small cells). In addition to providing high data rates, it is equally important to provide reliable HO mechanisms as this directly impacts on the perceived quality of experience (QoE) for the end-user. When it comes to small cell studies, the majority of works concentrate solely on capacity and throughput analysis, with fewer devoted to HO management. But a true challenge remains which is reliable HO mechanisms that provides high data-rates for moderate-to-high speed users in an urban environments.

Low cost of the small BSs attract the operators to deploy anywhere and turn them on and off in a coordinated manner to save energy. As a consequence of turning small cells on and off, the radio channel conditions change dramatically for the mobile users and so the neighboring cell list changes rapidly. This may also result in increased interference and high energy consumption of the network, too frequent HOs, unnecessary HOs with Ping-Pong (PP) (back and forth HO) effect, HOF and increased delay. If a device undergoes multiple HO, the HO delay will be accumulate resulting in a severe deterioration to the user experience [3].

The mobility mechanisms in cellular networks enable the users to move within the coverage area anywhere and still be serviced. At large, we can differentiate radio access mobility

The associate editor coordinating the review of this article and approving it for publication was Kai Yang.

into two separate procedures: cell (re-)selection and HO. Cell (re-) selection happens when the UE is in IDLE_MODE, i.e. with no active transmission/reception, and needs to select a suitable cell to camp on and thus be reachable when incoming data is available (subsequently transitioning to active or CONNECTED_MODE). Cell reselection will be triggered whenever a new camping cell is deemed better than the current one. Differently from cell (re-)selection, HO happens when the UE is in CONNECTED_MODE and a better serving cell is deemed better than the current cell. In this paper, the focus will be on the CONNECTED_MODE mobility (i.e. HO) under diverse cell deployments for both LTE and NR.

Pollini published the trends of HO design [4] and recognized, already in 1996, that the demand for increased capacity, leading to smaller cell sizes, would involve an increased number of HOs. The first generation (1G) had a cell size of tens of kilometers. The 1G HO procedure was tested in 1978 [5], consisting of a UE-assisted procedure followed by network decision making. In 2G networks, decreasing cell size increased the HO problems. To speed up the HO process, distributed HO decision making was used [6], which was costly in terms of signaling overheads during HO. The PP problem was addressed by adding a hysteresis value in the decision-making process. Soft HO algorithms were used in 3G code-division multiple access (CDMA) networks [7]. Therein, a single frequency network allowed two BSs to communicate with the user simultaneously, with some signal combination at the receiver end. This soft HO, following the make-before-break idea, improves the link gains but increases the interference to other users, eventually outweighing the gains [8]. As for 4G and 5G technologies, the HO process went back to a break-before-make scheme, with a larger number of BSs than the previous network technologies. The resulting small cells lead to small HO areas and no time for extensive HO signaling. These constraints demand simplified HO procedures for future ultra-dense networks.

In order to meet high mobile data traffic, HetNets are arguably the most promising solution with some notable challenges like inter-cell-interference-coordination (ICIC), mobility management and backhaul provisioning. Out of these issues, mobility management has special importance. In homogeneous deployments, UEs use the same set of HO parameters (such as time to trigger (TTT) and hysteresis margin (HM)) throughout the network. However, using the same set of parameters in HetNets would degrade mobility performance as noted in [9]. When a user tries to connect with an overloaded cell (i.e. may be at the same time the next neighboring cell is less overloaded), it faces poor QoS and a HOF occur due to the deficit in resources. Hence, adaptive management of the cellular network is required to improve mobility performance.

A. 5G TARGETS AND THE IMPORTANCE OF MOBILITY

To meet 5G targets, key technology components include massive Multiple-Input and Multiple-Output (MIMO) antennas, latency-optimized frame structure, high and low-frequency bands interworking, and ultra-lean transmissions. The scope of 5G is much wider than higher system capacity, high data rates and better coverage. One example is the Internet of Things (IoT) devices, where key challenges are low energy consumption, low device cost, handling large numbers of devices and extreme coverage. The second example is ultra-reliable low-latency communication (URLLC) to target critical industry applications. Despite these new and diverse communication paradigms, e.g. narrowband IoT (NB-IoT), the HO management procedure specified by the 3rd generation partnership project (3GPP) remains unchanged. We can then foresee some required HO enhancements to cater for the unique features of NB-IoT and URLLC. 5G will also provide autonomous transportation, mission critical services, access to remote health care, smarter agriculture and public warning systems like Tsunami and earthquake warning messages with the primary notification (short information) and secondary notification (detailed information) [10]. So, mobility management with forward compatibility is an important design principle to enable smooth future optimizations. The first specification was limited to non-standalone NR operation, meaning that NR will rely on LTE for initial access and mobility. A standalone NR operation is the scope of the 3GPP Release 15 that provides a new radio system complemented by a next-generation core network [11]. To meet the 5G targets, mobility management will play an important role, especially in successfully realizing small cell deployments, with its foreseen capacity boost while challenging radio dynamics; but also noting that the demand of mobile services *on-the-move* is increasing with the appearance of new mobility paradigms such as self-driving vehicles, drones, and mobile small cells [12].

B. MOTIVATION FROM AN ENERGY SAVING POINT OF VIEW

Telecommunication is responsible for 0.4 percent of worldwide CO_2 emissions and the data volume is expected to increase ten times every five years that result up to 20 percent increase in energy consumption. This directly impacts the environment through increased global warming. Analyzing the HO solution available in the literature to reduce the number of HOs and make the HO process efficient is a step forward to reduce energy consumption that has a direct impact on CO_2 emissions and operational expenditure of the operator [13].

C. MAIN OBJECTIVES AND SCOPE OF THIS PAPER

This paper provides a survey on HO management in cellular networks, with special emphasis on LTE and NR deployments. Some general concepts of radio access mobility in cellular networks are presented including, IDLE_MODE vs CONNECTED_MODE mobility, HO types (e.g. Inter-/Intra-frequency, Inter-/Intra-cell layer, Inter-/Intra-RAT and Inter-/Intra-operator HO), measurement and reporting, break before make vs make before break HO, and finally the

HO performance metrics. All of these concepts are studied briefly to highlight the challenges and current research focuses in these areas. In addition, an overview of HO management in LTE and NR is presented to highlight the main differences. A brief analysis of the HO techniques available in the literature for the CONNECTED_MODE mobility is shown to see the benefits and drawbacks of each scheme. HO management challenges and enhancing techniques, key points for formulating an efficient HO scheme (that could be considered for 5G NR) and future research directions are also described.

This paper is organized as follows. Radio access mobility is discussed with general concepts in section II. A detailed overview of HO management in LTE and NR are presented in section III and IV respectively. Challenges of HO management and ideas for radio access mobility enhancement with future research directions are described in section V. Finally, section VI concludes the paper.

II. RADIO ACCESS MOBILITY IN CELLULAR NETWORKS: SOME GENERAL CONCEPTS

In this section, we will provide the general concepts behind the radio access mobility procedure in a cellular network.

In a UE-assisted-Network-Controlled HO, the UE will generally perform some kind of signal strength (SS) or signal quality measurement over a specific downlink reference signal (RS) from the S-BS as well as the neighboring BSs. After processing the measurements, if a certain condition is fulfilled (entry condition), a measurement report (MR) is transmitted to the S-BS. Once the MR is correctly received at the S-BS, the HO preparation phase between target BS (T-BS) and serving BS (S-BS) starts and a HO request is transmitted to the T-BS. Upon successful admission, a HO command (HOcmd) is transmitted from the S-BS to the UE. Once the HOcmd is successfully received, the HO execution phase starts, in which the UE accesses the target cell, by means of cell synchronization and a random access procedure. A HO confirmation (HOconf) message is transmitted by the UE once it is able to receive broadcast information from the T-BS. During the HO completion phase, the DL data path is switched towards the target side by the user data gateway (UDGW) and, as a consequence, the T-BS starts receiving packets from the UDGW. Finally, the T-BS transmits a HO complete message to the S-BS to inform the success of the HO. After that, the S-BS releases any allocated resources to that UE.

A. IDLE_MODE VS CONNECTED_MODE MOBILITY

In IDLE_MODE mobility, the UE is not engaged in an active data connection but still needs to be reachable via signaling (paging) through an appropriate cell. Paging is a way to broadcast a brief message over the entire service area usually in a multicast fashion by many BSs at the same time [14]. The UE monitors the paging channel for incoming service requests and for cell (re-) selection process. Such UEs ''wake-up'' periodically to synchronize and check the paging messages from the network. Upon reception of a paging message, the UE establishes the connection with the BS controlling the cell on which the UE is camped. Upon successful connection, involving a random access procedure, the UE transitions to a CONNECTED state [15]. Cell (re-) selection happens only when the UE is in IDLE_MODE and needs to change for a more appropriate cell in order to not compromise the successful reception of future paging messages. Cell reselection is therefore a way to change the current selected cell to camp on, to a better cell. The UE seeks to identify a suitable cell based on, so-called, IDLE_MODE measurements and cell selection criteria. Suitable cells are those whose measured attribute meets the SS/quality selection criteria (coined as *s-criteria*) for the cell selection procedure. If a suitable cell is not available, it will identify an *acceptable* cell. In this last case, the UE will camp on an acceptable cell and will start the cell reselection procedure. In addition, a UE in IDLE_MODE will strive for reducing battery power consumption. This is achieved by a technique called Discontinuous Reception (DRX), whereby the terminal disconnects its receiver and enters a low power state. The terminal will periodically ''wake-up'' to receive paging indications, with typical wake-up periods (so-called DRX cycles) of, typically, 0.32s, 0.64s, 1.28s and 2.56s in LTE [16] (see section §22.5.1). The goal is to minimize interruption in paging reception during the cell reselection procedure. Noteworthy, and unlike in CONNECTED_MODE, the UE can make the cell reselection decision on its own and it is not required to report the measurements or events to the network. When camped on a cell, the UE shall regularly search for a better cell according to the cell reselection criteria. If a better cell is found, that cell is selected.

At every DRX cycle, at least, the UE shall measure RSRP and RSRQ levels to evaluate the so-called cell selection criteria (s-criteria, in short) which is defined in [17] (see section §5.2.3.2 therein). The UE shall filter the RSRP and RSRQ measurements of the serving cell using at least 2 measurements. Failure to fulfill the s-criteria means that the UE needs to identify the new serving cell. If the s-criteria is not satisfied for a specific number of consecutive DRX cycles, the UE initiates measurements of all neighboring cells regardless of the measurement/priority criteria provided to the UE by the network.

In IDLE_MODE, a centralized entity (which for LTE is the mobility management entity (MME)) has crude location information of the UE. This location is known per group of cells denoted as a tracking area (TA). To prevent the frequent registration during ping pong effect and reduce the signaling overheads of location management, a TA list (TAL) was introduced in 3GPP Release 8. In this scheme, one UE can have a list of TAs instead of one TA per UE. The UE keeps the TAL until it moves to a cell that is not included in TAL [18]. Research efforts in this area are mainly devoted to provide dynamic TAL configuration and to reduce the paging delays for IDLE_MODE mobility.

Differently from cell (re-)selection, HO happens when a user is active on a data session or phone call while moving from one BS to another. The user changes its serving cell while it has an active connection. The HO procedure seems to be a simple process of switching from one tower to another tower but in reality, it contains a set of processes running beyond the radio access network as the UE moves from one BS to another BS seamlessly. The UE measures the SS from the serving and neighboring BSs. If the SS of the neighboring cell becomes stronger than the serving cell, a MR is sent to the serving BS to inform about the current state of the UE. Because of the possible fluctuations in the radio channel, the UE will not trigger the MR based on the instantaneous radio channel measurement compared to a threshold, but it will rather perform some local measurement processing to average-off those fluctuations and prevent from unnecessary HOs (e.g. ping pong effects). The UE will generally perform some kind of measurement averaging over some measurement bandwidth, and it will implement a hysteresis loop whereby the average measurement from the neighboring cell should be larger than a given offset during a specified amount of time. In LTE, for example, both L1 and L3 filtering is implemented to introduce a certain level of averaging. In addition, several offset values (cell-specific, frequency-specific etc.) are introduced to determine entering and leaving conditions for MR transmission. Furthermore, these conditions must be fulfilled during a specified amount of time, which for LTE is the Time-To-Trigger (TTT) time.

All these parameters ensure a steady SS of neighboring cells before the UE sends the MR to the serving cell requesting HO. The cost of seamless connectivity comes from high signaling overheads among the BSs and small cells [19]. So, mobility has a direct impact on the performance of the data session since the data transmission may be subject to interruption due to the change of serving BS and high signaling overheads, thus affecting throughput, latency, etc. Main research directions within CONNECTED_MODE mobility are looking into HO parameters optimization to improve the HO process and reduce the radio link failures (RLFs), but sometimes this increases the power consumption of the UEs located at the cell edge. These parameters avoid link failures before HO thus engaging the serving BS for a long time that increases the power consumption. Recent research is focused on an early initiation of HO that will help to improve the energy efficiency by switching OFF the serving BS resources earlier [13]. Another interesting future research direction is to reduce the signaling overheads during HO in order to improve the energy efficiency of the HO procedure.

B. HANDOVER TYPES

Different HO types, owing to the different involved domains, are discussed in the following subsections, highlighting the main challenges.

1) INTER-/INTRA-FREQUENCY HANDOVER

If a serving and target BS operates on the same carrier frequency, we refer to them as intra-frequency neighbors. The UE will then perform measurements over different downlink RSs, each of them originating from different neighboring cells. It is usually the network which conveys the information to the UE on which RS to measure at any given time. If a serving and target eNB operate on a different carrier frequency, they are called inter-frequency neighbors. In this case, the UE needs measurement gaps to perform the measurements on different frequencies [21]. A measurement gap indicates the time period when no DL/UL transmissions are performed. The objective of this time gap is to enable UE to switch its carrier frequency and perform measurements from the neighboring BSs when they operate on frequencies other than that of the S-BS [15]. Intra-frequency measurements are based on the ranking of the cells with the same carrier frequency while inter-frequency measurements are based on the ranking of the quality of other carrier frequencies. A UE must measure both Intra and Inter frequency in the order of priority indicated by the BS. The more the measurement frequencies, the greater is the UE battery power consumption, which poses a big challenge [15]. Another main challenge in future RATs is to reduce the measurement gaps and thus improve the throughput for inter-frequency HO.

2) INTER-/INTRA-CELL LAYER HANDOVER

HetNets provide a multiplicity of cellular layers (including macro, micro and pico cell layers) aiming to fulfill the capacity requirement of the users in a cost-effective way. To provide high capacity and fill the ''coverage holes'' of the macro cell layer, small cells (micro or pico) are added to improve the network service quality and performance by offloading the users from a large macro cell. Intra-frequency macro-macro/pico-pico cell HO is called Intra-layer HO and it is based on the received SS/signal quality from cells belonging to the same layer. To optimize the UE power consumption, periodic inter-frequency measurements are taken at specific measurement gaps. When the pico cell quality becomes better than a threshold, the UE offloads to the small cell layer, this being called a macro to pico/Inter-layer HO. Another reason for HO to other layer is due to load balancing principles. If a UE is crossing two pico cells with an overlapping coverage region, an intra-frequency pico-pico HO occurs [22]. An Inter-layer HO can be both Intra- and Inter-frequency HO. In Intra-frequency Inter-layer HO, the main challenge is the interference as both macro and small cell are working on the same carrier frequency while Inter-frequency Inter-layer HO has high UE battery power consumption as it needs to measure other frequencies [15].

In HetNets, the main challenges and future research directions are listed as follows. Firstly, as the coverage of pico cells is small, Inter-layer HOs are more frequent than the Intra-layer HOs and this generates high ping-pong and signaling overheads which will in turn deteriorate the network QoS. Secondly, energy efficient schemes such as power control and dynamic sleeping decrease the range of pico cells that will affect cell selection decision and trigger more HOs. Lastly, during Inter-layer HetNet HO, the joint optimization of resource allocation, interference cancellation, power control and load balancing (that affects the HO decision) is a big challenge [23].

3) INTER-/INTRA-RAT HANDOVER

Horizontal HO occurs in the same radio access technology (RAT), and it is also called an intra-RAT HO. This HO process happens between various cells of the same network mainly due to preserving the connectivity of the UE with the network. On the other hand, the HO that occurs between different RATs is called a vertical or inter-RAT HO. It is the network that decides and instructs the UE to perform an inter-RAT HO. In heterogeneous access networks, the need for inter-RAT HO can be initiated as a special case of convenience rather than connectivity (e.g., according to a particular service choice of the user). Network switching automation and seamlessness are the main challenges with inter-RAT HO [24]. In the Inter-/Intra-RAT HO field, researchers are trying to improve the load balancing between RATs, among others.

4) INTER-/INTRA-OPERATOR HANDOVER

Different operators provide an opportunity to the users to use one or many types of systems and technologies. A HO that occurs between technologies/networks and systems of the same operator is called intra-operator HO. On the contrary, inter-operator HO occurs between technologies/networks and systems that are not provided by the same operator [25]. A good example of inter-operator HO is roaming, which refers to the possibility for a user outside the range of its home network to connect to another available network [14]. Device equipment and SIM card functionalities instruct the UE to measure other operator reference signals and perform inter-operator HO when the UE is out of the range of its home network. Connectivity on the move to other countries is one of the biggest benefits of inter-operator HO. The main challenge for the operators is to support other mobile operator's frequency bands, interfaces, protocols, network elements and roaming agreement for seamless roaming. Another roaming challenge is the different network standards and mobile data services among countries. Current research is attempting at improving the roaming quality through refined management to guarantee efficiency and thus increase the revenue.

C. MEASUREMENTS AND REPORTING

The UE is configured by the BS to perform the SS and/or signal quality measurements over a set of received RSs sent by the S-BS and the neighboring BSs. The time-frequency location of these signals and their design are known by both the UE and the BS. As an example, in LTE, the Reference Signal Received Power (RSRP) is defined as the average received power without interference and noise components. Also in LTE, the Reference Signal Received Quality (RSRQ) is defined as the ratio between the RSRP and the Received Signal Strength Indicator (RSSI), where RSSI is the total received power including noise and interference. When the RSRP measurement falls below a threshold specified in the s-criteria, the UE starts measuring the neighboring BSs. Different values of s-criteria thresholds decide on how to perform intra/inter-frequency measurements. To facilitate the measurements from the neighboring cells, the UE stops the DL data transmission for a duration specified by Measurement Gap [15].

After processing the measurements, including filtering at layers L1 and L3, if an entry condition is fulfilled, a MR is triggered to the S-BS. Specifically for LTE, the A3 event [20] is used as entry condition to assess if the filtered SS of the T-BS is better than that of the serving cell plus a hysteresis margin (called A3 offset). The entry condition has to be maintained during a time defined by the TTT timer. In order to achieve a good compromise between HO reliability and HO frequency, HO optimization deals with the adjustment of the TTT, A3 offset, and the L3 filter coefficient K [26]. The measurements are performed with specific measurement gaps and these are valid till identifying a better neighboring BS according to the specified reporting criteria. In general, three reporting criteria are used for HO-related measurements:

- Event-triggered reporting: the UE reports the measurement after a specified event has occurred.
- Periodic reporting: the UE reports the measurements at specified time intervals.
- On-demand/blind reporting: the UE reports the measurement immediately after it receives a request from the BS.

An event is triggered when an entry condition is satisfied. Such conditions are signaled by the BS in the form of threshold, hysteresis and offset parameters. Blind redirection (or blind HO procedure) is used to support load balancing. In this case, the UE receives a HO command to detect and access the target cell indicated in that command, without having received a measurement report from the UE [20].

D. BREAK-BEFORE-MAKE VS MAKE-BEFORE-BREAK

Hard HO refers to the case when a UE is connected to only one BS at a time, which causes the connection with the S-BS to be momentarily interrupted before a new connection is made towards the T-BS. This is sometimes referred also as a break-before-make HO. On the contrary, a soft HO occurs when the UE is able to be simultaneously connected to two BSs for a while, thus minimizing and practically eliminating, any interruption time during the communication. This is sometimes referred to as a make-before-break (MBB) HO [24]. In LTE's break-before-make HO, data interruption of 40ms is experienced at each HO [27], which deteriorates the user experience. Zero HO execution (HOE) time, low latency, high reliability, road vehicle safety, and efficiency are the use cases of CONNECTED_MODE mobility in NR. HOE time can be reduced in NR by using the MBB strategy, along with multi-cell connectivity and synchronized HO [27]. It has to be kept in mind however that, under NR at high frequencies, the mm-wave signal can be blocked by building material, mortar, bricks, and the human body. So, an NR standalone architecture with traditional HO methods cannot react quickly enough. A possible solution to this problem is multi-connectivity [28]. The main goals for designing a measurement and reporting scheme is the reduction of signaling overheads and ping-pongs. However, multi-cell connectivity increases the UE measurement and reporting complexities, simultaneous utilization of resources in multiple cells and power consumption. Research efforts are addressing concepts that allow the UE to autonomously release and add various radio links to reduce the signaling overheads [27].

E. HANDOVER PERFORMANCE METRICS

There are different performance metrics to measure the HO performance including, HOF rates, HO delays, HO frequency, gain in average throughput, HOE time, PP rate, energy consumption, signaling overheads due to frequent HO, HO success rate, data latencies, packet loss and HO interruption time. A brief definition of these performance metrics is provided next:

- HOF rate: HO failures may occur at different stages during the HO process. These may include failures at the radio link due to poor radio conditions in both the uplink and downlink, failure to convey particular messages after a given number of retransmission attempts, synchronization failures, failures during random access procedures and others. The HOF rate is the total number of HOFs divided by the sum of the total number of HOFs and the total number of successful HOs [29].
- HO frequency (or HO rate): is the number of HOs per second. HO frequency usually increases by increasing the UE speed and decreasing cell size. In LTE, lowering parameters such as the TTT and A3 offset, favors HO to be triggered earlier at the cost of an overall increased number of HOs. So, there is a tradeoff between reducing the HO frequency and HOFs [29].
- PP rate: is the number of ping pong events during a given period of time. A ping pong event is the occurrence of a HO between a S-BS and a T-BS, followed by another HO to the original S-BS, all this happening under a predefined, and generally short, time [26].
- HO Energy consumption: is the amount of energy consumed during a HO procedure [13].
- HO success rate: is the total number of successful HOs divided by the total number of triggered HOs [29]. A HO process is successful if it is completed before the measured SS from the S-BS drops below the minimum acceptable SS level [15].
- Data Latency: a period between the reception (or transmission) of the last data packet from the S-BS and the first packet through the T-BS is called the data latency [30].
- HO interruption time: is a time period during a HO procedure when the UE cannot exchange user plane packets with any of the BS [31].

The optimum HO procedure should try to minimize the HOF rate, HO delay, HO frequency, HOE time, PP rate, energy consumption, overheads, latencies, packet loss and HO interruption time. Also, it should maximize the gain in average throughput and HO success rate along with the use of adaptive TTT, A3 offset and HM values (depending upon different mobility scenarios).

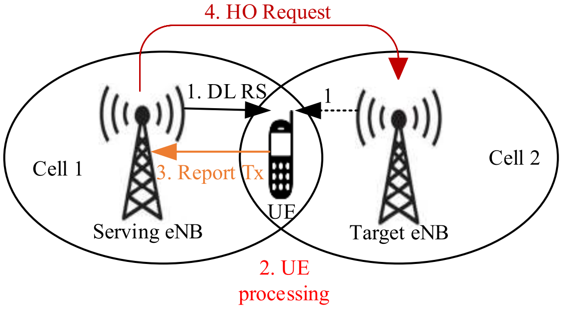

FIGURE 1. HO process in 3GPP LTE.

III. HANDOVER MANAGEMENT IN LTE

This section describes the LTE HO with a brief introduction of key features and the entities involved in LTE mobility. A step by step HO procedure is provided with a detailed literature survey.

A. KEY FEATURES OF LTE HANDOVER MANAGEMENT

The HO process in 3GPP LTE can be succinctly described in four steps for CONNECTED_MODE mobility as shown by Fig. 1 [30],

1. The serving and neighbor eNBs transmit DL RS.
2. The UE performs and processes SS measurements.
3. The UE sends the measurement report (MR) to the serving eNB.
4. Based on the received MR, the serving eNB makes the HO decision and sends a HO request to the target eNB.

The entities and interfaces involved in LTE mobility are the following as shown in the network architecture provided in Fig. 2:

- **eNB:** The eNB handles some aspects of user mobility in a CONNECTED state. For example, receiving the MR sent by the UE, deciding whether a HO is needed, requesting the target eNB for admission control, and others.
- **Mobility Management Entity (MME):** Handles mobility management during session establishment, IDLE_MODE, setting up of bearers and security procedures. It is the controlling node in the core network.
- **Serving Gateway (SGW):** The SGW is in charge to set up a user plane and act as a local anchor for CONNECTED_MODE user mobility. It forwards the user data to the UE through the eNB.
- **Packet Data Network Gateway (PGW):** It is responsible for connection from LTE to external network and allocation of IP address to the UE.
- **S1:** The S1 is an interface between the RAN and the LTE core network side. The S1-U interface carries user

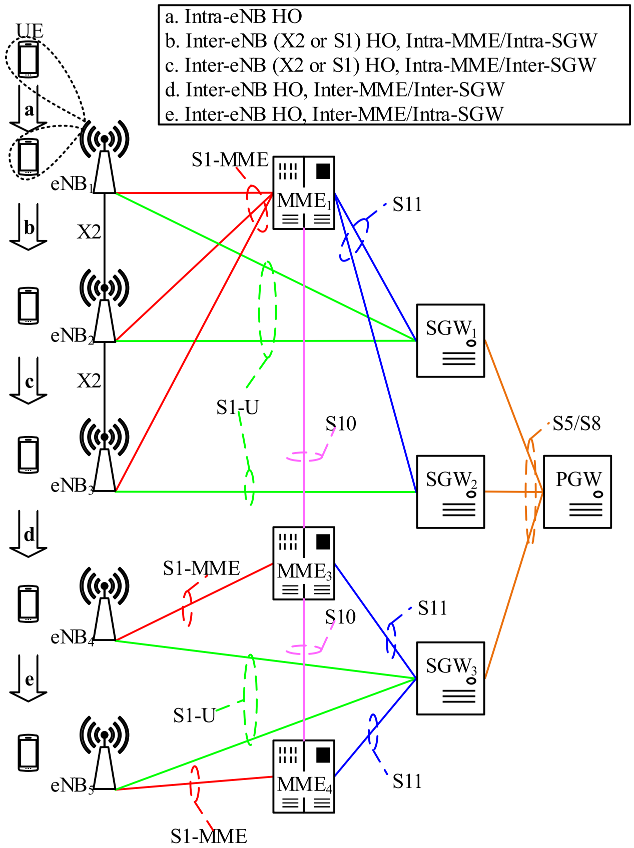
a. Intra-eNB HO
b. Inter-eNB (X2 or S1) HO, Intra-MME/Intra-SGW
c. Inter-eNB (X2 or S1) HO, Intra-MME/Inter-SGW
d. Inter-eNB HO, Inter-MME/Inter-SGW
e. Inter-eNB HO, Inter-MME/Intra-SGW
UE
a
b
c
d
e
eNB1
eNB2
eNB3
eNB4
eNB5
X2
X2
S1-MME
S1-MME
S1-MME
MME1
MME3
MME4
S11
S11
S11
SGW1
SGW2
SGW3
PGW
S5/S8
S1-U
S1-U
S10
S10

FIGURE 2. LTE mobility architecture along with relevant interfaces and HO use cases.

data between eNB and SGW, and the S1-MME interface carries control signaling messages between the eNB and the MME. The S1 can carry HO signaling and HO preparation (HOP) messages between two eNBs as shown in Fig. 2. In such a case, the HO is called S1-based HO.

- **X2:** The X2 provides an interface between two eNBs which carries control information messages. During an X2-based HO, the X2 interface carries HO signaling and HOP messages between the source and target eNBs.

With the above architecture in mind, the following HO types in LTE can be defined:

- **Intra-eNB HO:** When both the source and target cells belong to the same eNB, see Fig. 2 (a).
- **Inter-eNB HO:** When both the source cell and target cell belong to different eNBs. In this particular case, we assume that the MME will not change as a consequence of the HO (i.e. intra-MME). In addition, the SGW may (intra-SGW) or may not (inter-SGW) be relocated, as determined by the implemented physical deployment, see Fig. 2 (c) and see Fig. 2 (b) respectively. This inter-eNB HO is either supported by the X2 interface or, in its absence, by the S1 interface. We then refer to X2 HO and S1 HO respectively. Specification for both interfaces and their functionalities can be found in [32] and [33], along with references therein.
- **Inter-eNB HO with MME Change:** In this case, the HO involves a change of MME via signaling messages through the S10 interface between the source and target MMEs. Only the S1 interface (not X2) will be used in this case. In addition, the SGW may (intra-SGW) or may not (inter-SGW) be relocated, as determined by the implemented physical deployment, see Fig. 2 (e) and see Fig. 2 (d) respectively.

In general, the S1 HO procedure is more complex than the X2 HO because the MME has to act as an intermediary message relay and coordinate between the source and target eNBs. The HO process in X2 and S1 is completed alike but the user will face a longer data interruption in the S1 case.

As an example, a more detailed description of the intra-MME/Serving Gateway (SG) HO procedure is shown in Fig. 3 and described hereafter [21]. The HO procedure is divided into three phases: the HOP phase (steps 0–6), the HOE phase (steps 7–11) and the HO completion phase (steps 12–18):

- **Step 1:** The UE measurement procedure is configured by the serving eNB according to access restriction and roaming information.
- **Step 2:** The UE sends a MR to the serving eNB.
- **Step 3:** Based on the MR, the serving eNB makes a HO decision.
- **Step 4:** The serving eNB sends a HO request message (containing the necessary information to prepare the HO at the target side) to the target eNB.
- **Step 5:** The target eNB performs admission control to find out if enough resources can be granted by the target eNB.
- **Step 6:** The target eNB sends the HO request acknowledgement to the serving eNB. As soon as it is received, data forwarding between source and target eNBs may start.
- **Step 7:** The target eNB generates the RRC message and transmits it to the UE with the necessary information to perform the HO.
- **Step 8:** The serving eNB sends the SN (Sequence Number) Status Transfer message to the target eNB in order to keep track of packet ordering.
- **Step 9:** The UE detaches from the old cell and synchronizes with the target cell (new cell).
- **Step 10:** After successfully completing the random access channel (RACH) procedure, the target eNB responds with uplink (UL) allocation and timing advance (TA) information for the UE.
- **Step 11:** After UE successfully accesses the target cell, the UE sends the RRC connection reconfiguration complete message along with an uplink buffer status report to the target eNB which indicates that the HO procedure is completed for the UE. Then the target eNB starts sending data to the UE.
- **Step 12:** The target eNB informs the MME that the UE has changed cell and that the DL path from the SGW should be changed.

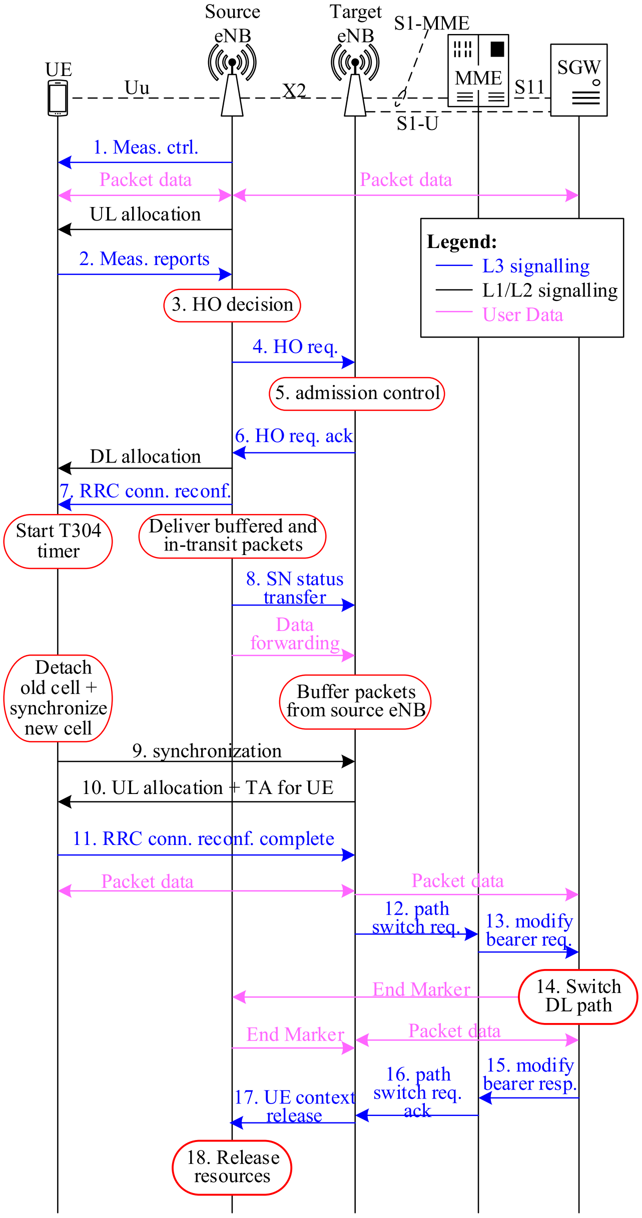

FIGURE 3. Intra-MME/Intra-SGW HO procedure, based on [21].

- **Step 13:** The MME sends a modify bearer request message to the SGW.
- **Step 14:** The DL data path is switched to the target side by the SGW.
- **Step 15:** Then SGW sends a modify bearer response message to the MME.
- **Step 16:** The MME acknowledges the path switch request.
- **Step 17:** The target eNB informs the success of HO to the serving eNB and triggers the release of resources used by the source eNB by sending a UE context release message.
- **Step 18:** The serving eNB release radio resources associated with the UE.

TABLE 1. Some examples of HO failure types [26].

Failure types	Category	Description
F0	Too late HO	T310 expiry before measurement report triggered
F1	Too late HO	T310 expiry before measurement report received
F2	Too late HO	Radio Link Control (RLC) measurement report transmission error
F3	Too late HO	T310 expiry before HO command transmission
F4	Too late HO	T310 expiry before HO command reception
F5	Too late HO	RACH failure after T304 expiry
F6	HO to a wrong cell	T310 expiry before HO confirm received
F7	Too early HO or HO to a wrong cell	RLC HO confirm transmission error

In the next subsections, we will provide the main causes of HO failures and how we can evaluate the RLF.

1) HANDOVER FAILURES

In general, HO failures can be categorized as a result of “wrong” HO decisions, which can be described as:

- **Too early HO:** After successful HO, UE connects back to the serving cell (result in high PP rate).
- **Too late HO:** When serving cell channel quality drops too low before HO to the target cell is completed.
- **HO to a wrong cell:** UE disconnects from the target cell and connects to a new cell.

Examples of different HO failure types are described in Table 1, [26], where T310 is a timer which starts upon indications of problems in the radio link, and timer T304 is used as a safeguard for the HO execution phase, and starts after the HO command is received (see step 7 in Fig. 3). At the expiry of timer T304, it initiates the RRC connection re-establishment procedure otherwise it stops at the successful completion of HO.

Radio Link Failure:

In RRC_CONNECTED state, radio link monitoring enables the UE to determine whether it is in-sync or out-of-sync with respect to its serving cell. On getting a consecutive number of out-of-sync indications, UE stats a RLF timer ‘T310’ as shown in Fig. 4. Both in-sync and out-of-sync (N311 and N310) counters are configured by the network. The T310 timer will be stopped when a number of consecutive in-sync (N311) indications are reported (case 2 in Fig. 4). If T310 expires, a RLF is declared and the UE turns off its transmission to avoid interfering with other UEs and subsequently tries to re-establish a connection within a measured delay called UE connection re-establishment delay (case 1 in Fig. 4). The delay when UE detects the need for RRC-connection Re-establishment until it transmits a random access signal to the target cell is called RRC connection re-establishment delay. In-sync threshold Q_{in} corresponds

)

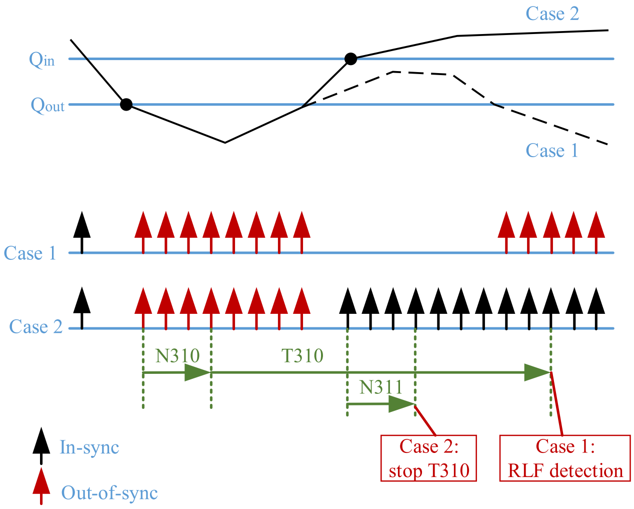

FIGURE 4. **RLF detection [16].**

to 2% Block Error Rate (BLER) and out-of-sync Q_{out} corresponds to 10% BLER [16] (see section §22.7.1). With no DRX configuration, when the DL radio link quality in 200ms period becomes worse than Q_{out}, an out-of-sync occurs. If the DL radio link quality in 100ms period becomes better than the Q_{in}, an in-sync occurs [16] (see section §22.7.2). These occurrences are reported by the UE to the physical layer and the higher layers which may apply L3 filtering for evaluation of the RLF. With DRX configuration, in-sync and out-of-sync evaluation periods are extended matching the DRX length [16].

B. LITERATURE SURVEY

In this section, we identify and categorize the proposed LTE HO techniques found in the literature, for both single-layer and multi-layer (HetNets) LTE networks.

When addressing HO in LTE, existing works are usually divided into targeting the reduction of RLF [34], reduction of ping-pongs (PPs) [35], or a combination of both [36]–[38], reduction of HO rate [23], [30], [39], and reduction of HOFs [23], [40], [41]. Other works are devoted to finding the signaling overheads and the power consumption during HO [42], assess the energy-saving [13], [36], load balancing [43], the impact of inter-site distance (ISD) [26], [44], and UE speed [26] on the HO process.

RLF is one of the main causes for HOF. The details of the RLF are presented in section §III. A. 1. A distributed mobility robustness optimization algorithm was proposed in [34] to minimize HOFs due to RLFs by adjusting TTT and offset parameters. The algorithm classifies too early, too late and wrong cell HOF categories and optimize the HO parameters according to the dominant failure. To reduce the PPs, an orientation matching based fast HO approach for LTE advance was proposed in [35] to choose the target eNB based on orientation match, received SS and current load.

The techniques to reduce both, the RLFs and PPs are presented in [36]–[38]. A reduced early HO scheme was proposed in [36] to reduce the RLF and PPs and achieve high energy efficiencies, at the cost of keeping the other performance parameters within acceptable limits. A fuzzy multiple criteria cell selection scheme was proposed in [37] that consider UE UL conditions, resource block utilization in addition to s-criteria of LTE conventional cell selection method. This technique provides a highly reliable cell selection with reduced HO PPs and failures that lead to high throughput. A HO detection with self-organizing HO parameters algorithm was proposed in [38] that was based on reinforcement learning concept. The algorithm improves the user mobility performance by reducing call drops, PPs and HOFs.

The works in [23], [30], [39] are devoted to reducing the HOR. A distance-based HO scheme for macro and femto cells was proposed in [23] to minimize the HOFs and unnecessary HOs. A reactive HO technique was proposed in [30] to postpone the HO process until UE lose its serving cell or the most probable location of a UE arrives. This method reduced the HO overheads, HO rate, and latencies. Using user mobility information, time the user stays in femto cell was calculated in [39]. HO from macro to femto cell was proposed only when the available data volume of femto is larger than the macro cell and the time user stay in femto cell is greater than a certain threshold.

The works in [23], [40], [41] are dedicated to reducing the HOFs. Based on radio link conditions of a UE, two different HO schemes were proposed in [40]. For poor radio link conditions, radio link proactive HO was proposed. Using this method, HO can be completed before RLF by lowering the TTT and A3 offset values. For UEs in good radio link conditions, HO was divided into two parts, early HO preparation (EHOP) and HOE. EHOP reduces the HOF rates and HOE was triggered only when HO is needed, to avoid the PPs. An adaptive beamforming scheme for LTE with dynamic adjustment of the HO HM was proposed in [41] to improve the HO performance of the high-speed railways. Beamforming with various gain factors was used to improve the received signal quality in the overlapping region. The proposed scheme effectively improves the HO trigger and HO success probabilities.

The work in [42] is dedicated to finding the signaling overheads and power consumption that results from the transmission of HO related signaling during the HO procedure. The results depict that MR transmission is the main contributor to air-interface signaling overheads. A simplified power consumption model (for both UE and eNB) is also presented that finds RACH transmission as the highest power consuming part out of the UE signaling message transmission.

To make the HO process energy efficient, the works in [13], [36] are devoted to reducing energy consumption that has a direct impact on CO_2 emissions and operational expenditure of the operator. A reduced early HO with bandwidth expansion scheme was proposed in [13] to enhance energy saving. The proposed scheme initiate HO early and turned off freed resource blocks when compared to conventional LTE at the cost of increased RLFs up to 5 percent. While a reduced early HO scheme was proposed in [36] to provide high energy efficiencies and thus reduce the operational expenses of a mobile operator.

When a user tries to connect with an overloaded cell, it faces poor QoS and HOF due to a deficit in resources. At the same time, may be the neighbor cell resources remained unused. Hence, adaptive HO management of the cellular network is required through load balancing. The work in [43] is dedicated to improving load balancing. Moving distance of a UE was calculated using a change of RSS. The HO was triggered only by comparing the moving distance with a threshold. The proposed technique shifts the load released from a heavily loaded cell to the nearest neighboring cells.

The impact of ISD on the HO process is presented in [26], [44] and the impact of UE speed in [26]. In [26], it is found that certain cell size can be found around which any increase or decrease of cell size brings the performance degradation. For large cell sizes, low-speed UE HOFs may increase due to the inability to escape from poor radio condition areas. It is also investigated that high-speed UEs degrade the performance, especially for small cell sizes. In [44], the impact of ISD on HO success rate in an LTE network was discussed. It is found that by increasing the ISD, we should decrease the TTT, A3 offset and HM values.

When addressing HO in LTE HetNets, existing works are usually divided in targeting the HOR reduction [45], joint reduction of PPs and HOFs [9], joint reduction of HOR, HOFs and PPs [46], and reduction of HO signaling [19]. Other works are devoted to assess the energy saving [29], [47], [48], load balancing [49], [50], improving user capacity or QoE [51]–[56], and user Mobility State Estimation (MSE) (for example, categorizing into low, medium or high speed cases) for HO decision [57], [58].

An SINR-based HOR analysis is presented in [45] that considers the impact of SINR on HO because of the strong interference caused by the dense small cell deployment. SINR effect at the user locations before and after it moves are used to find the interference and correlated association regions in various time slots. On this basis, the mathematical expressions of SINR-based vertical and horizontal HO rate are derived under maximum average received power strategy. The results show that SINR-based HO has a HOR which is ‘‘gap’’ lower than the SINR-free case. This gap indicates the effect of interference on the HO procedure.

Release 10 idea of range expansion was used in [9] to improve the signal to interference and noise ratio (SINR) for HO completion purpose and increase pico cell DL coverage. But poor PP performance was observed. Then mobility based ICIC (MB-ICIC) was proposed for HetNets which combines HO parameter optimization along with enhanced ICIC to reduce PPs and HOF rates. For low-speed UEs, large TTT was proposed to reduce PPs. A Control/Data Separation Architecture (CDSA) in ultra-dense HetNets was proposed in [19] with several analytical models to achieve promising gains in term of HO performance and reduce the HO signaling overheads when compared to conventional networks.

A fuzzy-logic based scheme is proposed in [46] for a dense small cell deployment to jointly reduce the HOR, HOFs and PPs. The radio channel quality and user speed are used to choose a hysteresis margin for HO decision in a self-optimization manner. The results show that the proposed scheme has superior HO performance than the existing solutions.

The energy-saving techniques for LTE HetNets are proposed in [29], [47], [48]. HO procedure with the suitability of macro cell power off was proposed in [29]. It was found that macro cell power off procedure is suitable only for low-speed users and high small cell densities. One valuable approach to improve energy efficiency is to minimize the unnecessary HOs in overlapping areas of the cells without compromising the QoS of the mobile terminal [47]. The author formulated the HO problem as a constrained Markov decision process and proposed the HO only if the expected energy saving amount through the HO is greater than the energy loss at BS during the HOE phase to avoid frequent HOs. The main idea was to consider HO energy consumption under the expectation on stochastic behaviors of the HO parameter to make a HO decision, which results in a reduction of frequent HOs occurring within the areas covered by many BSs. Another energy-centric HO decision algorithm that minimizes the power consumption at the UE side in an integrated LTE macro-femto network was proposed in [48]. It suitably adopts the HO HM that reduce the power consumption and interference at the cost of increased HOE events.

The load balancing techniques are proposed in [49], [50]. The load balancing algorithm proposed in [49] was for small-cell, macro-cell, and HetNets. The algorithm limits the load released from a heavily loaded cell to the nearest neighboring cells. A centralized Self-Organizing Network (SON) was introduced to continuously estimates the load status of the cells and decide a HO for a user to provide load balancing and avoid performance oscillations. A resource block utilization ratio was defined to measure the cell load and a threshold was also introduced to determine the overloaded cells. To reduce PPs of the load between the cells, moving the load impact on the neighboring cell was also considered. Another HetNet model with randomized Radio Resource Management (RRM) in femto cells and a proactive macro to femto offloading scheme (to improve data rates in the congested networks) was proposed in [50]. A hybrid access control approach at femto cells was also proposed to allow up to 20 times higher data rates for the femto BSs authorized users, compared to non-authorized users in the network. The femto tire operating resources were determined by random fragment allocation and simple spectrum fragmentation. To implement an intelligent RRM, the fragment size was dynamically adjusted according to the congestion of the network. It is beneficial for the operators to guarantee a minimum average data rate for the authorized users and introduced new services to promote the usage of femto cells especially in indoor scenarios such as offices and homes.

A context-aware Markov-based HO model is reported in [51] to improve the user capacity. In this work, the performance of the mobile user in a HetNet scenario has been characterized as a function of the neighboring cell power profile, user mobility, traffic load and HO related parameters. In [52], a machine learning based HO management scheme is presented to improve the QoE of the user. The scheme learns from the past experiences considering how the QoE of the user is affected when HO is completed to a certain eNB. To predict the most appropriate cell for HO, a neural network-based supervised learning is used. Another algorithm to improve the QoE of the user was proposed in [53] that uses the QoS metrics to obtain a utility function for each macro and famto cell. The resulting utility determines the HO necessity. An analytical model for the cross-tier HO in HetNets is proposed in [54] to improve the QoE of the user by reducing the HOR, HOFs and PP rate. Closed-form expressions for HO performance metrics are obtained as a function of TTT, BS density and user mobility that can provide guidance to deploy HetNets. A frequent HO mitigation algorithm is proposed in [55] to improve the QoE of the user especially by reducing the HORs and improving the throughput. Based on the proposed algorithm, frequent HO experience users are categorized as either PP or fast-moving user. The PP users are managed by adjusting HO related parameters and the fast-moving users are handed to the macro cells. The impact of fading and shadowing on the HO performance is analyzed in [56]. Therein, an analytical model is used to analyze PP rate and HOFs under varying channel conditions for HetNets with vehicular users. The results show that the fading can significantly degrade the HO performance.

A user MSE for HO decision are presented in [57], [58]. UE reselection count threshold and fading frequency threshold was defined to specify low, medium and high mobility state of a user in [57]. This method was very attractive as UE already measure the Doppler frequencies for channel estimation purpose. Another user mobility pattern bases HO decision-making approach for HetNets was proposed in [58] to improve the user performance and provide the user preferences. This approach performed better than the conventional vertical HO algorithms.

IV. HANDOVER MANAGEMENT IN NR

Some of the key features of NR include [11]: high-frequency operation, spectrum flexibility, forward compatibility, and Ultra-Lean design. NR operates between 1 GHz to 52.6 GHz in both licensed and unlicensed spectrum. Forward compatibility ensures enabling new services and introducing new technologies in the radio interface design in the future. The ultra-lean design principle aims to minimize the always-on transmissions (like broadcasting of system information, signals for BS detection and always-on RSs for channel estimation), to achieve high data rates and high energy performance of the network.

The transmission scheme in NR is the same as LTE, Orthogonal Frequency Division Multiplex (OFDM). NR supports multiple subcarrier spacings (called numerologies) ranging from 15kHz up to 240kHz, with some limitations on the supported bandwidths, these ranging between 5MHz and 400MHz. NR allows bandwidth adaptation to reduce the device energy consumption because not all the devices need to use the same bandwidth. The frame structure of NR includes a 10ms radio frame which is divided into ten 1ms subframes. Then, a subframe is further divided into slots and each slot consist of 14 OFDM symbols. The length of the slot depends on the aforementioned numerology configuration.

NR support Frequency Division Duplex (FDD), Time Division Duplex (TDD) and a new duplex scheme called dynamic TDD. Dynamic TDD is a key technology in NR whereby symbols within a slot can be dynamically allocated to either DL or UL as part of some scheduler decision.

Another key feature of NR is the massive number of steerable antenna elements (more than 64, for beam steering purpose) for both reception and transmission. The beam steering technique avoids interference and provides the targeted coverage for NR.

In NR, the main challenge is the limited network coverage due to higher radio channel attenuation. The difficulty of providing full coverage at higher frequencies is eliminated using interworking with the system operating at lower frequencies. The DL data rates are achieved using wider bandwidth at higher frequencies while the UL data rates are achieved on the lower frequency spectrum (because UL is power limited). Although lower frequencies have fewer bandwidths, the UL data rates are still achieved due to low channel attenuation.

Next section describes the NR HO with a brief introduction of key features and the entities involved in NR mobility. Also, a step by step HO procedure is provided with a detailed literature survey.

A. KEY FEATURES OF HANDOVER MANAGEMENT IN NR

A key challenge in NR is providing mobility robustness and minimizing service interruption. Due to shrinking the cell size, NR introduces three big challenges. First is the frequent HO resulting in increased HOF rate; second is the increased number of intra/inter-frequency measurements which reduces the battery life of the mobile user; and, third is the increased overheads due to frequent HO at microwave/mm-wave frequencies which may limit the frequency resources for static users.

A NG Radio Access Network (NG-RAN) node is either a gNB which provides NR user plane and control plane protocol terminations towards the UE or an ng-eNB which provide E-UTRAN user plane and control plane protocol terminations towards the UE [59]. Overall architecture of NR is shown in Fig. 5, where Xn interface interconnects the gNBs and ng-eNBs while NG interface connects gNBs and ng-eNBs to the NR core network. More specifically, NG-C interface connects gNBs and ng-eNBs to the Access and Mobility Management Function (AMF) and NG-U interface to the User Plane Function (UPF).

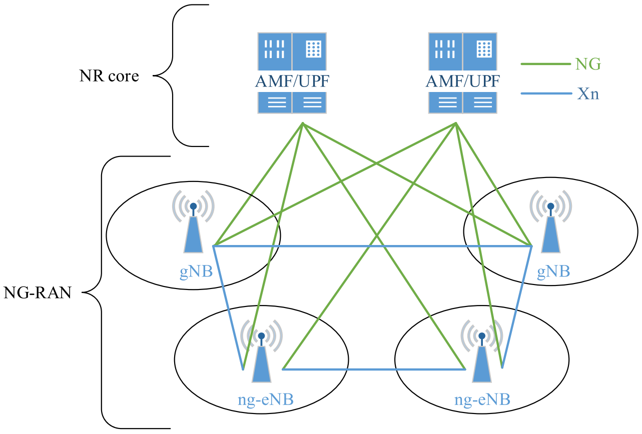

FIGURE 5. Overall Architecture of NR (adapted from [59]).

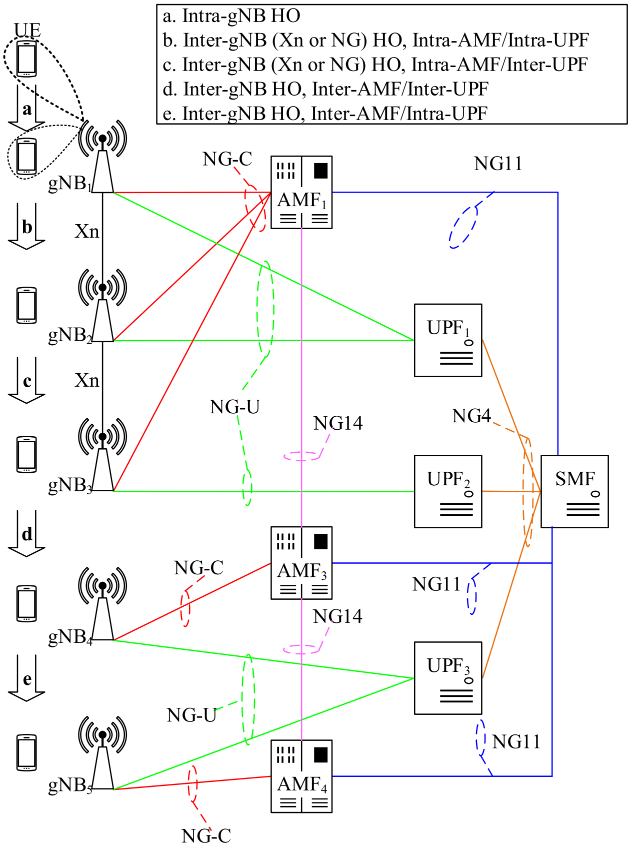
a. Intra-gNB HO
b. Inter-gNB (Xn or NG) HO, Intra-AMF/Intra-UPF
c. Inter-gNB (Xn or NG) HO, Intra-AMF/Inter-UPF
d. Inter-gNB HO, Inter-AMF/Inter-UPF
e. Inter-gNB HO, Inter-AMF/Intra-UPF
UE
a
b
c
d
e
gNB1
gNB2
gNB3
gNB4
gNB5
Xn
Xn
NG-C
NG-C
NG-C
NG-U
NG-U
AMF1
AMF3
AMF4
UPF1
UPF2
UPF3
SMF
NG11
NG11
NG11
NG14
NG14
NG4

FIGURE 6. NR mobility architecture along with relevant interfaces and HO use cases.

Fig. 6 shows the NR mobility architecture along with relevant interfaces (as specified in 3GPP TS 23.501 [60]) and HO use cases. With the Fig. 6 architecture in mind, the following HO types in NR can be defined:

- **Intra-gNB HO:** When both the source and target cells belong to the same gNB, see Fig. 6 (a).
- **Inter-gNB HO:** When both the source cell and target cell belong to different gNBs. In this particular case, we assume that the AMF will not change as a consequence of the HO (i.e. intra-AMF). In addition, the UPF may (intra-UPF) or may not (inter-UPF) be relocated, as determined by the implemented physical deployment, see Fig. 6 (c) and see Fig. 6 (b) respectively.
- **Inter-gNB HO with AMF Change:** In this case, the HO involves a change of AMF via signaling messages through the NG14 interface between the source and target AMFs. Only the NG interface (not Xn) will be used in this case. In addition, the UPF may (intra-UPF) or may not (inter-UPF) be relocated, as determined by the implemented physical deployment, see Fig. 6 (e) and see Fig. 6 (d) respectively.

Multi-RAT dual connectivity (DC) operation is supported in NG-RAN while a UE is in RRC-CONNECTED mode. DC operation allows the UE to connect with two gNBs simultaneously (the serving and the target gNBs) and thus reduce the HO interruption time to 0ms.

NR introduced a UE third state called INACTIVE and this falls between IDLE and CONNECTED state of LTE. The purpose of this state is to reduce the time to bring the UE in the CONNECTED state while reducing the signaling overheads and improving the UE battery life. A UE in this state can go to the CONNECTED state without the involvement of Non-Access Stratum (NAS) signaling as both the UE and the gNB store AS context in this state. In the RRC_CONNECTED mode, network-controlled mobility applies to UEs being categorized into two types of mobility: beam level mobility and cell level mobility [59]. In the presence of beamforming feature with multiple antennas, beam level mobility handles the procedures that ensure good alignment between transmitting and receiving antenna beams. Beam level mobility can only be performed based on the Channel State Information Reference Signal (CSI-RS). It deals with at the lower layers using physical layer and Medium Access Control (MAC) layer control signaling. It does not require explicit RRC signaling to be triggered. Also, RRC is not required to know which beam is being used at a given point in time. On the other hand, cell level mobility requires explicit RRC signaling to be triggered (i.e. intra/inter-gNB HO). During the HO procedure, data forwarding, duplication avoidance and in-sequence delivery is ensured. As with LTE, NR support timer-based HOFs and to recover from a HOF, an RRC connection reestablishment procedure is also used.

The basic HO procedure in NR (which is similar to LTE HO procedure) is shown in Fig. 7 [59], containing HOP (0-5), HOE (6-8) and HO completion (9-12) phases:

- **Step 1:** The UE measurement procedure is configured by the serving gNB according to access restriction and roaming information and the UE sends a MR to the serving gNB.
- **Step 2:** Based on the MR and RRM information, the serving gNB makes a HO decision.
- **Step 3:** The serving gNB sends a HO REQUEST message (containing the necessary information to prepare the HO at the target side) to the target gNB.

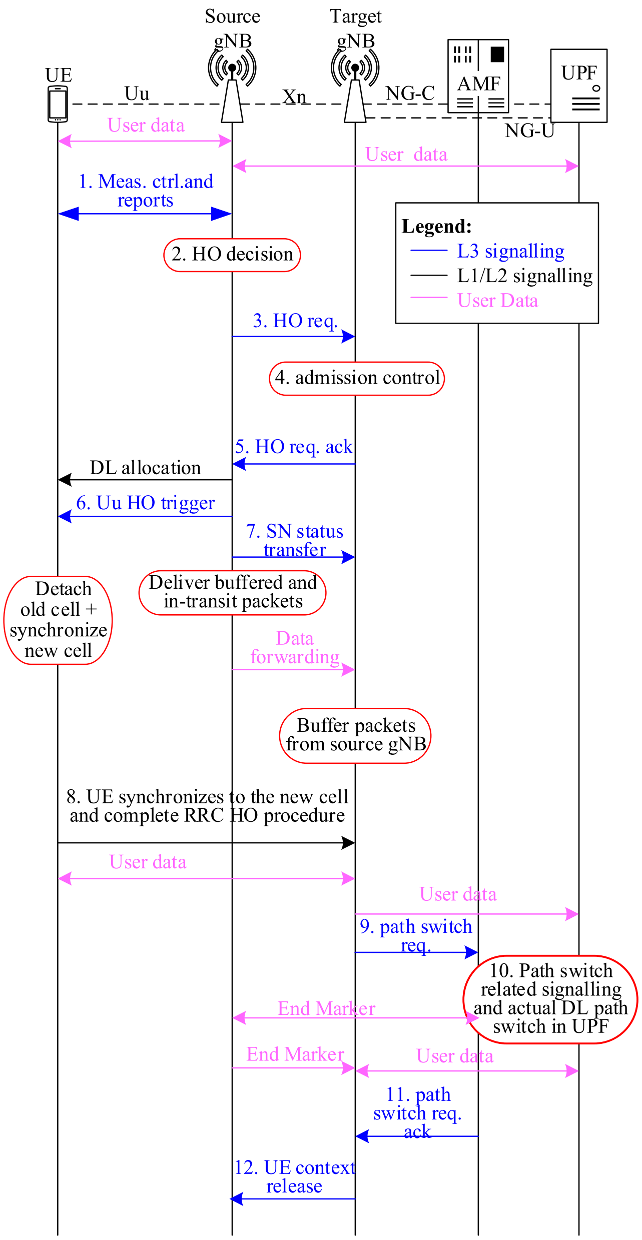

FIGURE 7. Intra-AMF/UPF HO in NR [59].

- **Step 4:** If the resources can be granted by the target gNB, the target gNB performs Admission Control procedure.
- **Step 5:** The target gNB sends a HO Request Acknowledgement to the serving gNB. As soon as the serving gNB receives the HO Request Acknowledgement message, data forwarding may be initiated.
- **Step 6:** Serving gNB sends a HO command to the UE.
- **Step 7:** Serving gNB sends the SN Status Transfer message to the target gNB.
- **Step 8:** UE detach from the old cell and synchronize with the target cell.
- **Step 9:** The target gNB informs AMF that UE has changed the cell, through the Path Switch Request message.

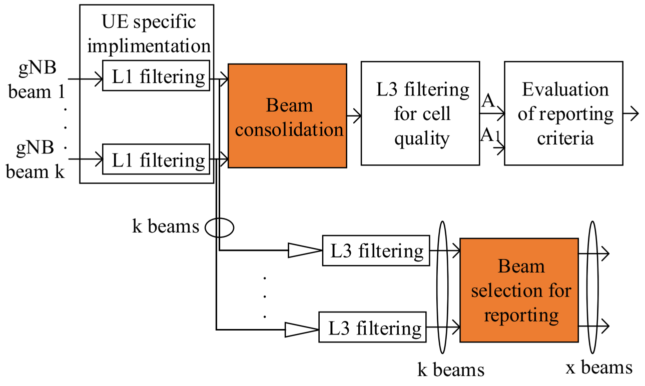
UE specific implimentation
gNB beam 1
gNB beam k
L1 filtering
L1 filtering
Beam consolidation
L3 filtering for cell quality
A
A1
Evaluation of reporting criteria
k beams
L3 filtering
L3 filtering
Beam selection for reporting
k beams
x beams

FIGURE 8. NR measurement model [59].

- **Step 10:** The DL data path switch towards the target side by NR core.
- **Step 11:** The AMF acknowledges the Path Switch Request.
- **Step 12:** The target gNB informs the success of HO to the serving gNB and triggers the release of resources by the serving gNB by sending a UE Context Release message. Finally, the serving gNB release the radio resources associated with the UE.

The UE measures multiple beams of a cell in RRC_ CONNECTED mode and, to derive the cell quality, the measurements results (power values) are averaged. The UE is configured to consider a subset of N best-detected beams above a certain threshold. To derive the beam quality, filtering takes place at the physical layer. To derive the cell quality from multiple beams, the second phase of filtering takes place at the RRC level [59]. Cell quality is derived in the same way for the serving cell(s) and for the non-serving cell(s). The gNB may configure the UE to include measurements for the X "best" beams in the MR. In order to provide better user-centric (beam level) mobility experience, UE specific CSI-RS can be configured. For example, to achieve better measurement accuracy with improved SINR, narrow beam CSI-RS can be configured for the cell-edge users in high-frequency scenario.

A brief NR measurement model is shown in Fig. 8 [59] where k beam specific samples are filtered at L1 and reported to L1 and L3. The beam consolidation block derives the cell quality and reports it to L3 for filtering. The evaluation of reporting criteria block checks whether measurement reporting information is necessary to send on the radio interface based on different measurements illustrated by A and A_1. In the second stage, for the non-serving cells, L3 filtering is performed and X measurements are selected out of k to send on the radio interface as a measurement report.

Using Supplementary UL (SUL) [61], the UE can be configured with two UL bands for one DL band of the same cell as shown in Fig. 9. Usually, the cell coverage in UL direction is lower because UE transmit power is less than the gNB. The idea is to use a low-frequency secondary

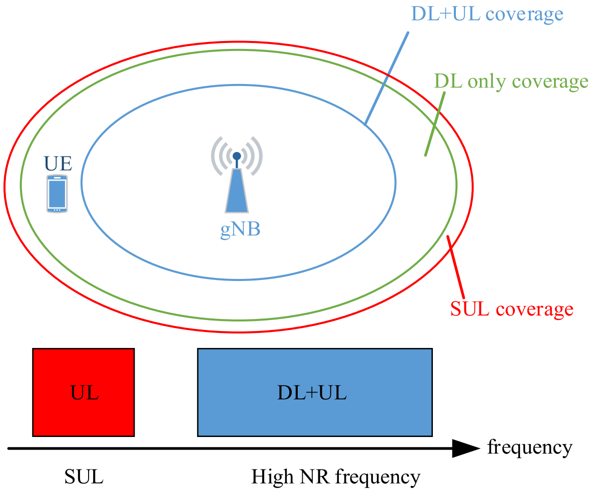

FIGURE 9. SUL example [59].

UL band in addition to the original higher UL frequency band, for the cell-edge users to increase the cell coverage and thus overcome adverse channel conditions. Therefore, the SUL concept can be utilized to reduce the HO failure types due to UL transmission errors (for example F0, F5 and F7 in Table 1).

B. LITERATURE SURVEY

When addressing the HO in 5G, existing works are usually divided in targeting the reduction of HOFs [3], [31], [62], reduction of HO rate [1], [63], [64], [65], or a combination of both [66]–[70], improving throughput or QoE of the user [1], [31], [71]–[77] and reduction of HO signaling [78], [79]. Other works are devoted to assess the energy saving [62], [80] and load balancing [62], [64].

The works in [3], [31], [62] are devoted to reduce the HOFs. A Markov chain based HO management strategy was proposed in [3]. This technique chooses and assigns the optimal next eNB to OpenFlow tables of the mobile node virtually (before the actual connection) considering transition and available resource probability estimation. Compared to the conventional LTE approach, HOFs and delays were reduced to 21 and 52 percent respectively. Estimating the mobility state (low, medium and high speed) of a 5G user for scaling TTT was proposed in [31] to reduce the HO interruption time, HOFs and thus increase the throughput. The proposed method also activates the DC mode if a UE surpasses a certain threshold of velocity. In [62], the author proposed cashing technique for 5G which was used to store the future data contents in advance, to use when wireless resources are not sufficient. Offloading UEs from the heavily loaded cell was also proposed for load balancing.

The works in [1], [63]–[65] are dedicated to reducing the HOR. HO skipping is a technique that skips some HOs along a high-speed user trajectory to reduce the number of HOs and improve user throughput at high speeds. Three different topology-aware HO skipping techniques were proposed in [1] including, location-aware, cell size aware and hybrid HO skipping. It was assumed that the serving area of each eNB is known. An anchor-based multi-connectivity architecture with HO probabilities expression was proposed in [63] to reduce the HO cost. Through multi-connectivity, the best access point is chosen as a HO anchor to provide control plane and thus reduce the HO rate. In [64], the author proposed mobility-aware user association strategy for the mmWave 5G network with attractive features including, after mobility channel condition tracking, load balancing and stopping the recurrent HOs. An algorithm is proposed in [65] to minimize the unnecessary HOs for HetNets with dense small cell deployment. The distance between the UE and the small cell along with the angle of movement of the UE is used to construct a short candidate list that helps to reduce the unnecessary HOs and signaling overheads.

The works in [66]–[68] are focused to reduce both the HOR and HOFs for URLLC. In [66], the authors use mmWave links for data transmission and μWave links for handling paging and control information. Using this technique, large data can be cached via high-speed mmWave links. The mobile user can mute a HO towards a small cell and only maintain macro-cell connection for communicating control signal. This technique reduces the HOFs by avoiding unnecessary HO towards small cells. A similar procedure is introduced in [67] for 5G HetNets that decouple user and control plane to reduce the HO frequency. The work in [68] targets at low HOR with zero HOFs that is achieved using LTE-NR DC (i.e. Dual Connectivity) feature with signaling and data radio bearers duplication. The work in [69] maintains the HOFs to an operator-predefined acceptable level and reduces the unnecessary HOs for small cells based HetNets. A tradeoff between HOFs and unnecessary HOs is found using a time metric and estimated time-of-stay of the user is used to avoid long neighbor list. A grey rational analysis (GRA) based algorithm is proposed in [70] to reduce both the HORs and the HOFs in dense small cell HetNets. The GRA method is used to rank the best available target cells for HO and an analytical hierarchy process is utilized to obtain the weight of HO metrics. These weight are further utilized to reduce the link failures and frequent HOs.

The works in [31], [71]–[77] are targeted at maximizing the throughput during the HO process. In [71], [72], context-aware RAT selection for the 5G network is proposed to improve the QoE, mainly the user throughput and delays. A fuzzy logic controller is modeled to extract the RAT suitability metric based on the BS load, UE mobility, backhaul load, traffic flow type and RSS. To improve the QoE of the user in 5G HetNets, machine learning-based algorithms (namely k-nearest neighbor, KNN, and support vector machine, SVM) can be applied to find the optimal HO solution [73]. Two modified methods based on Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) are proposed in [74] to improve the QoE of the user (mainly by reducing HORs, HOFs, and improving the user throughput). The first method uses an entropy weighting technique and the second uses the standard deviation weighting technique to improve the HO metrics. Novel HO methods for 5G HetNets are proposed in [75], [76] to improve the throughput, load balancing and HO related metrics. The influence of the interference from both the macro cell and small cell is taken into consideration to offload the user from the congested cell for a forced HO to the best small cell. The best small cells are obtained from a reduced the neighboring cell list (NCL) which is optimized considering the time-of-stay of the user and the SINR threshold. Cell load and the interference are taken into consideration for a modified A3 HO initiation event. A user-velocity-aware HO skipping scheme is proposed in [77] to improve the average throughput of the mobile user in a two-tier cellular network. The proposed scheme sacrifices the best BS connectivity to reduce the HO rate and maintain a longer connection duration. To quantify the performance of the proposed HO schemes in terms of user throughput, a mathematical model is presented using stochastic geometry.

The works in [78], [79] are focused on the HO signaling reduction. A mobility management scheme based on UE location tracking was proposed in [78] to provide proactive and seamless HOs. The sounding reference signals (SRSs) transmitted by the UE are utilized to determine the angle of arrival and line of sight path and thus track its location. This technique eliminates the HO signaling overheads and provides seamless mobility at the cost of increased computational complexity. Performance of an ultra-dense network millimeter-wave network with control and user plane separation architecture was compared with a conventional architecture in [79]. To minimize the HO cost with a certain coverage probability requirement, an analytical framework was also proposed. The HO cost and coverage probabilities outperformed than a conventional architecture. It was also found that the conventional architecture HO cost can be reduced by adding more macro BSs while it is more beneficial to add small cells with the proposed architecture. This makes the proposed architecture a key enabling architecture for 5G small cells.

A novel HO technique was introduced in [80] for 5G URLLC to provide perspective on different trade-offs between user plane delay and energy efficiency. The author proposed a direct HO request from UE to target gNB based on UE measurements and bypassed the role of source gNB. If the request is accepted from the target gNB, data transfer can begin more quickly. The UE will also provide an alert to its source cell once the target gNB respond to its availability, to avoid unnecessary resource allocation from the source cell. Both energy efficiency and delay metrics can be improved due to faster HO to the better targeted gNB.

A comparison table of the literature survey is shown in Table 3. Overall, location estimation errors are found in [1], [9], [23], [30], [35], [39], [43], [64], [81], the works proposed in [3], [13], [19], [29], [30], [36], [40], [43], [48], [49], [79], [80] effect other HO performance related parameters, the algorithm proposed in [34], [35], [37], [38], [47], [57], [58], [78] increased the computational complexity, the multi-connectivity in [27], [28], [31] [62], [63], [82] introduces high UE complexity and high utilization of resources, and efficient HO algorithm were not shown in [28], [50], [64].

V. HANDOVER MANAGEMENT FROM LTE TO NR: DIFFERENCES, ENHANCERS AND CHALLENGES

In this section, we will highlight the differences between LTE and NR handover management, and we will elaborate on possible HO enhancements and future challenges. Specifically, we will analyze the mobility enhancement in LTE and NR with some techniques and procedures available in the literature. We will discuss the mobility enhancement techniques considering the followings, HO management techniques for high-speed scenarios, beam management and beam level mobility and UL RS based mobility. This section also shows the discussion for formulating an efficient HO scheme with future research directions based on the conducted study.

The HO procedure in NR is very similar to the procedure in LTE, in which the network controls UE mobility based on UE measurement reporting. NR has both the beam level and the cell level mobility while LTE has just cell level mobility. This being said, the adoption of high frequency bands with beamforming may increase interruption time in NR as compared to LTE due to beam sweep delay. In Rel-15 NR, 0ms interruption time can be achievable by using intra-cell using beam mobility and addition/release of small cell for DC operation, whereby simultaneous connections with the source cell and the target cell are maintained. With newly introduced RACH-less HO in NR, the RACH phase to the target cell is skipped thus reducing the interruption time during HO, avoiding random access at every HO and reducing the HO delay.

A. OVERVIEW OF MOBILITY ENHANCEMENT IN LTE AND NR

An overview of mobility enhancement solutions for both LTE and NR is given in Table 2. Some of the solutions like RACH-less HO and MBB have been adapted in LTE R14 and R15 whereas dual connection and UE based mobility are currently under discussion in NR.

To target zero mobility interruption time and zero HOF rates even at 120 km/h speed for NR, a conditional MBB HO was proposed in [83] for URLLC. The solution achieved zero HOF rate by receiving a HO command when the source cell has good radio link conditions and execute the HO command at the preferable location of the target cell. A zero interruption time was achieved through DL reception from the target cell first and then releasing the resources of the source cell.

The random access procedure can be avoided in some synchronized network deployments, where the timing advance between the source cell and the target cell is the same, or the value of timing advance is zero (e.g., in small cells). With combined MBB+RACH-less HO, the HO interruption time

TABLE 2. Overview of mobility enhancements [84].

Solutions	Description	Interruption	Technology
LTE HO	Conventional HO (break before make).	~ 50ms	
MBB	Serving eNB link is maintained until RACH preamble transmission in the target eNB starts.	~ 30ms	LTE Rel-14, Rel-15
RACH-less HO	Corresponding RACH phase is skipped completely in the target cell during HO.	~ 20ms	LTE Rel-14, Rel-15
MBB + RACH-less	Source link is maintained until the first Physical Uplink Shared Channel (PUSCH) transmission starts at the target.	~ 5ms	LTE Rel-14, Rel-15
Dual-connection based mobility	Simultaneous connections with the source cell and the target cell are maintained.	~ 0ms	Under discussion for NR
UE based HO	HO command is provided earlier (in good radio link conditions) and UE decides when to perform the HO.	~ 50ms	Under discussion for NR

(HIT) can be decreased up to 6 ms. If the target gNB sends the DL data earlier to the UE without receiving an HO complete message, the HIT can further be reduced close to 0 ms. However, by considering the misalignment of the subframe boundary between the serving gNB and the target gNB, the HIT would be around 1 ms [83]. If a UE at the cell edge stays logically connected to both the serving and the target cells, then the DC have higher reliability with minimum interruption time up to 0ms [84]. This solution is currently under discussion for NR.

B. HANDOVER MANAGEMENT IN HIGH-SPEED SCENARIOS

The main challenge in 5G is to provide 50 Mbps data rate in DL and 25 Mbps in UL for the users moving at very high-speeds up to 500km/h for a high-speed train and up to 250 Km/h for highway deployment scenarios with high reliability and availability [85]. There are four main challenges for high mobility wireless communications [86],

1. Optimized network deployments
2. Advanced signal processing
3. Accurate channel estimation
4. Effective mobility management

To meet the 5G requirements in high-speed train scenario, optimized network deployment has an important role (see adaptive optimization in [87] and enhancing video QoE in [88]). In advance signal processing, new coding, modulation, precoding, diversity techniques, and waveforms are required to overcome the challenges in 5G. Due to the large Doppler spread, it is difficult to accurately predict the channel, estimate and track the fast time-varying fading coefficients. Several channel models for high-speed train scenario are proposed in literature such as [89]–[91] but still, more research is required in this area. For effective mobility management, efficient HO algorithms are required to reduce the number of HOs, latency, and HOF rates.

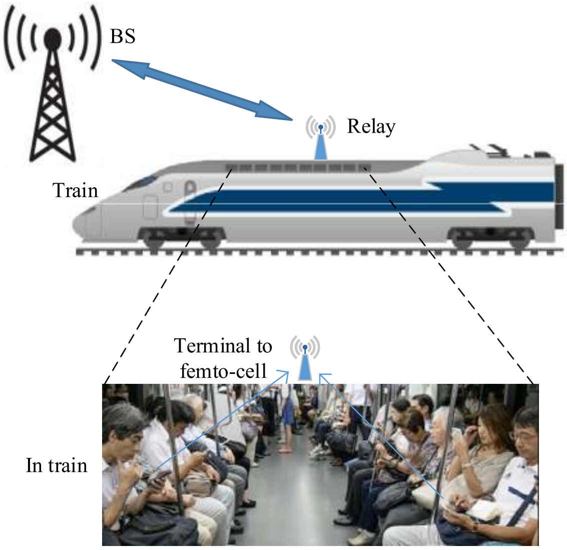

FIGURE 10. Two Hop network Architecture.

Two hop networks provide reliable and consistent QoS in high-speed train scenarios. A relay on the rooftop of the train is used to relay the signal between BS and in train users as shown in Fig. 10 [86]. HOF rates, HO probability, outage probability and HO latencies can be reduced using group HO [92], [93], and collective channel state information can be received from users [94]. Multiple antennas in a linear array can also be installed for high-speed trains [95]. In good radio link conditions, massive MIMO increases the multiplexing gains and reduces the diversity gains to increase the throughput. With poor radio link conditions, massive MIMO reduces the multiplexing gains and turns on the diversity gains to increase the SINR for more reliable transmission and better QoS.

The main challenges for high-speed users' mobility management are high penetration loss, frequent HOs, cell selection, and heavy signaling overheads if users are moving in a group [86]. Techniques for solving mobility management problem in high-speed scenario include deployment optimization with Radio Remote Unit (RRU), mobile relay, multi-connection, mobile cell, and geo-aided fast HO. These are described hereafter.

We can optimize the deployment scenario by connecting multiple RRUs to a signal Baseband Unit (BBU) to improve the coverage, reduce the dropped calls and the number of HOs [96]. It is good to plan the overlapping coverage between the neighboring cells so that enough time will be available for MRs that reduces the HOF rates effectively [97].

TABLE 3. Literature survey comparison.

Ref.	Technology	Scheme/Proposed Method	Advantages	Challenges/Drawbacks
[34]	LTE	distributed mobility robustness optimization. NS-3 simulator.	minimize HOFs due to RLF	increased the computational complexity
[35]	LTE	orientation matching based fast HO approach. Python based simulator.	enhance the HO performance, reduce the PPs and leads to the reduction of HO time	increase the measurement processing load
[36]	LTE	reduced early HO scheme. System-level simulation.	high energy efficiencies, reduce the operational expenses of mobile operator. A superlative value to TTT was provided for an unbiased RLF and PPs	effect other HO performance parameters
[37]	LTE	fuzzy multiple criteria cell selection scheme. System-level simulation.	provides a highly reliable cell selection with reduced HO PPs and failures that leads to high throughput	increases the computational load
[38]	LTE Advance	HO detection with self-organizing HO parameters algorithm. System-level simulation.	improve the user mobility performance by reducing call drops, PPs and HOFs	increased computational complexity
[30]	LTE	Reactive and proactive HO strategy.	reduced the HO overheads, HO rate and latencies	postponing the HO process introduces the data interruption and getting the most probable location for HO is challenging
[23]	LTE	distance based HO scheme for macro and femto cells.	minimize the HOFs and unnecessary HOs	estimating mobile user moving distance accurately is a challenge
[39]	LTE	macro to femto cell HO. HO decision based on data volume. System-level simulation.	HO from macro to femto cell is proposed only when the available data volume of femto is larger than macro cell and the time user stay in femto cell is greater than a certain threshold	estimation of mobility state of a user is the main challenge
[40]	LTE	radio link proactive HO and early HO preparation. OPNET Modeler simulator.	reduced number of HOFs	more overheads during HO
[41]	LTE	adaptive beamforming scheme with dynamic adjustment of HM.	improve the HO performance of the high-speed railways, effectively improve the HO trigger and HO success probabilities	estimating the serving area of gNB is challenging due to beamforming feature
[13]	LTE	reduced early HO with bandwidth expansion scheme. System-level simulation.	enhance energy saving, initiate HO early and turned off freed resource blocks	increased RLFs up to 5 percent
[43]	LTE	UE moving distance calculation and load balancing.	moving distance of a UE was calculated using the change of RSS, shift the load released from a heavily loaded cell to the nearest neighboring cells	cannot estimate the actual moving distance of a UE due to unpredictable clutter and fading, increased number of HOs
[26]	LTE	impact of cell size and user speed on the HO procedure. System-level simulation.	a certain cell size can be found around which any increase or decrease of cell size brings the performance degradation, high-speed UEs degrade the performance especially for small cell sizes	-
[44]	LTE	ISD impact on the HO procedure.	increasing the ISD: TTT, A3 offset and HM values decrease	-
[9]	HetNets	mobility based ICIC. System-level simulation.	combines HO parameter optimization along with enhanced ICIC to reduce PP and HOF rate	estimating the mobility state of a user is challenging
[19]	Ultra-dense HetNets	CDSA technique with several analytical models.	achieve promising gains in term of HO performance and reduced HO signaling overheads when compared to conventional networks	this work assumed only too-late HO as a cause of HOF
[29]	HetNets	HO with the suitability of Macro cell power off. System-level simulation.	saves energy	suitable only for low-speed users and high small cell densities
[45]	HetNets	SINR based HOR analysis.	Gap between SINR-free and SINR-based HOR indicates the effect of interference on the HO procedure	-

TABLE 3. ***(Continued.)*** **Literature survey comparison.**

[46]	HetNets	fuzzy-logic based scheme.	reduce the HORs, HOFs and PPs	-
[47]	HetNets	formulated the HO problem as a constrained Markov decision process.	save energy by proposing the HO only if the expected energy saving amount through the HO is greater than the energy loss at BS during the HOE phase to avoid frequent HOs	increase the complexity
[48]	HetNets	Energy-centric HO decision algorithm.	suitably adopt the HO HM to reduce the power consumption and interference	increased HOE events
[49]	HetNets	load balancing algorithm. System-level simulation.	shifts the load released from a heavily loaded cell to the nearest neighboring cells	increased the number of HOs
[50]	HetNets	randomized radio resource management with proactive macro to femto offloading scheme. System-level simulation.	hybrid access control approach allow up to 20 times higher data rates for the femto BSs authorized users, compared to non-authorized users in the network, beneficial for the operators to guarantee a minimum average data rates for the authorized users and introduced new services to promote the usage of the femto cells	an efficient HO mechanism for the proposed network is not discussed
[51]	HetNets	context-aware Markov-based HO model. Monte Carlo simulations.	improve the user capacity	-
[52]	HetNets	machine learning based HO management. LENA LTE-EPC simulator.	The scheme learns from the past experiences and improves the QoE of the user	high computational complexity
[53]	HetNets	QoS-based multi-criteria handoff algorithm. Monte Carlo simulations.	improve the QoE of the user	may increase the computational complexity
[54]	HetNets	analytical model for the cross-tier HO.	improve the QoE of the user by reducing the HOR, HOFs and PP rate	-
[55]	HetNets	frequent HO mitigation algorithm. NS-3 simulator.	improve the QoE of the user especially by reducing the HORs and improving the throughput	may increase the computational complexity
[56]	HetNets	Analytical HO model. System-level simulation.	The results show that the fading can significantly degrade the HO performance	-
[57]	HetNets	estimating the user mobility state (low, medium or high mobility).	UE reselection count threshold and fading frequency threshold was defined to specify low, medium and high mobility state of a user. This method was very attractive as UE already measure the Doppler frequencies for channel estimation purpose	high computational complexity
[58]	HetNets	user mobility pattern bases HO decision making. Trace-driven simulation.	to improve the user performance and provide the user preferences	high computational complexity
[27]	LTE, 5G	MBB HO strategy along with multi-cell connectivity.	reduced the number of HOFs	multi-cell connectivity increases the UE complexity and utilization of resources
[82]	4G, 5G	hybrid HO forecasting mechanism.	improves the HO decision mechanisms and reduce the HOF and PPs	high utilization of resources
[28]	LTE, 5G	Improved HO through dual connectivity. NS-3 based system-level simulation.	fast HO procedure, fast detection of RLFs and dynamic TTT adaptation	an accurate analytical model for mobility sceneries was not developed. Also, multi-cell connectivity increases the UE complexity with high utilization of resources
[81]	4G, 5G	HO skipping techniques.	reduce the HO rate and outperform from moderate to high velocities	estimating the serving area of a cell and user trajectory is challenging
[3]	5G	Markov chain based HO management strategy.	HOFs and delays were reduced to 21 and 52 percent respectively	the effect of this technique on PP, unnecessary and frequent HO is not considered
[31]	5G	estimating the mobility state along with DC. System-level simulation.	reduced HO interruption time, increased throughput and low HOFs	correct tuning of MSE is required, DC increases the UE complexity and utilization of resources
[62]	5G, HetNets	caching technique, DC and load balancing.	store the future data contents in advance, to use when wireless resources are not sufficient, Offload UEs from heavily loaded cells, reduced HOFs and energy consumption	multi-cell connectivity increases the UE complexity and utilization of resources
[1]	NR	topology aware HO	Location-aware, cell size aware and hybrid HO skipping	estimating the serving area of gNB is

TABLE 3. ***(Continued.)*** **Literature survey comparison.**

		skipping techniques. Monte Carlo simulations.	techniques reduce the number of HOs	challenging due to beamforming features, estimating the user trajectory is another challenge
[63]	5G	anchor-based multi-connectivity architecture.	reduce HO rate	multi-cell connectivity increases the UE complexity and utilization of resources
[64]	5G	mobility-aware user association strategy. Monte Carlo simulations.	after mobility, channel condition tracking, load balancing and stopping recurrent HOs	HO procedure for mmWave band has not been addressed properly and accuracy of GPS is a big concern for indoor situations
[65]	5G, HetNets	Algorithm to minimize unnecessary HOs. System-level simulation.	reduce the unnecessary HOs and signaling overheads of scanning	may increase the computational complexity
[66]	5G	Technique to reduce HORs and HOFs for URLLC.	reduces the HOFs by avoiding unnecessary HO with small cells	-
[67]	5G, HetNets	decouple user and control plane.	reduce the HO frequency	high computational complexity
[68]	5G, HetNets	LTE-NR DC feature with signaling and data radio bearers duplication.	low HOR with zero HOFs	multi-cell connectivity increases the UE complexity and utilization of resources
[69]	5G, HetNets	trade-off between HOFs and unnecessary HOs. MATLAB based simulation.	maintains the HOFs to an operator predefined acceptable level and reduces the unnecessary HOs for small cells based HetNets	-
[70]	5G, HetNets	grey rational analysis (GRA) based algorithm.	reduce both the HORs and the HOFs in dense small cell HetNets	may increase the computational complexity
[71][72]	5G, HetNets	context-aware RAT selection. System-level simulation.	improves the QoE, mainly the user throughput and delays	may increase the computational complexity
[73]	5G, HetNets	machine learning-based algorithm.	find the optimal HO solution	may increase the computational complexity
[74]	5G, HetNets	TOPSIS based methods. MATLAB based simulation.	improve the QoE of the user (mainly by reducing HORs, HOFs, and improving the user throughput)	may increase the computational complexity
[75][76]	5G, HetNets	Novel HO methods. MATLAB based simulation.	improve the throughput, load balancing and HO related metrics	
[77]	5G, HetNets	user velocity aware HO skipping scheme. Monte Carlo simulations.	improve the average throughput of the mobile user in a two-tier cellular network	estimating the user trajectory is challenging
[78]	5G	mobility management scheme based on UE location tracking.	eliminates the HO signaling overheads, provide proactive and seamless HOs	increased the computational complexity
[79]	5G	mm-Wave network with control and user plane separation architecture. Monte Carlo simulations.	HO cost and coverage probabilities outperforms than a conventional architecture, adding small cells with the proposed architecture make it more beneficial for 5G	may degrade some other HO performance parameters
[80]	5G	HO technique for ultra-reliable low latency communication.	a direct HO request from UE to target gNB based is proposed on UE measurements that bypassed the role of source gNB. Energy efficiency and delay metrics were improved due to faster HO to the better target gNB	if the HO request is rejected by the target cell, the system fallback to the existing gNB assisted HO which further increases the HO delays

Mobile relay maintains the connection with all UEs within a bus, car, train, and metros and provides the UEs an unaware HO process. The multi-connection process improves mobility performance at the cost of an increased number of signaling overheads and greater UE complexity [86]. The train geographical information (speed and location) can be tracked using a sensor and GPS. The RRU will be switched ON only when the train is approaching, to make the process energy efficient and reduce the interference [98].

There are two major problems caused by the Doppler Effect in high-speed scenario including, carrier frequency offset and fast fading. There are three strategies to deal with these problems, Doppler compensation, Doppler planning, and Doppler diversity [86]. Consider Doppler planning at the early stage of the network design as it is the best option to reduce the Doppler Effect for fast fading. Doppler estimation and compensation are effective means to deal with auto frequency correction (AFC) at the speed up to 350 km/h [99]. Doppler diversity in the receivers is used to improve the reliability of the transmissions, even at high mobility with imperfect channel estimation.

To improve the HO performance of the high-speed railways, an adaptive beamforming scheme for LTE with dynamic adjustment of the HO HM was proposed in [41]. A concept of moving cell was proposed for high-speed train scenarios at 60 GHz in [100]. Therein, a radio-over-fiber technique was employed to switch the serving cell in unison with the train and provide uninterrupted transmission to the users. However, the direction of the train and velocity needs to be known for synchronization and providing adjustment to the passenger's speed.

C. BEAM MANAGEMENT AND BEAM MOBILITY

Multi-beam operation is one of the key features in 5G that differentiate 5G from LTE and assist to fulfill 5G requirements [101]. A survey of beam-related techniques for mmWave can be found, for example in [102], [103]. A UE measures a set of analog beams for each digital port and reports the beam quality to BS which then assigns one or a small number of analog beams to the UE [104]. Beam management is required for above 6GHz to establish a seamless and low latency link with the UE. Beam management procedures are categorized into beam measurement and reporting, beam determination, beam switching, and beam recovery [80]. A radio connection between a UE with **N** analog beams and a BS with **M** analog beams has a total of **MN** TX-RX beam pairs. As the number of TX/RX beams are typically large, it is really important to ensure low overhead and UE complexity during an efficient beam measurement. In beam determination procedure, the BS and the UE find a beam direction to ensure good radio channel quality. Beam switch procedure is performed when the quality of the current beam degrades. Then the UE and the BS switch to another beam with better radio channel quality. When a UE is suffering from poor radio link conditions, it will get it as a beam failure. In the case of beam failure, another beam from the same cell can be used (e.g. through beam recovery) that avoids the frequent declaration of the RLF and cell reselection. If the beam recovery procedure will not be successful, the UE then initiates a RLF procedure and starts the cell reselection process. As the propagation losses increase on the higher frequencies, a higher number of beams are required to increase the coverage. Beam level mobility is managed at PHY and/or MAC layer without RRC signaling. The main challenge in beam management is finding the best combination of the TX and RX beams (that both UE and BS jointly select), and this increases the UE complexity.

D. UPLINK REFERENCE SIGNAL BASED MOBILITY

Unlike current HO schemes, the network could track and locate the mobile user measuring an UL RS [105] instead of having the UE measuring DL RS and reporting them back (as noted in LTE). Uplink RS measurements are processed in the network side to decide which BS or cell is the most appropriate to serve the user. Using the proposed method in [105], the transmission of MRs between user and network is not required. Thus, it improves the mobility performance through the reduction of HO signaling overheads.

When UEs are moving together, they can jointly be tracked in a group using one UL RS (instead of multiple RSs) potentially minimizing the resource consumption and improving the overall HO efficiency [106]. Since massive MIMO uses the UL channel measurement of transmitted RS (e.g. Sounding Reference Signal (SRS), in LTE), the same measurements can be reused for mobility purposes to improve the network performance with very low impact on the UE. DL data is transmitted to each UE without grouping as the grouping was just used to track the location of the UEs moving together.

The following steps are used for group-based mobility in [106]. The first step is to identify the UEs which are moving in a cluster/group. In the second step, one UE out of the group is configured to send group UL RS and other UEs in that group stop transmitting the individual UL RSs. The control of a group was handed to the target gNB to make the HO hidden from the UEs. When a UE is not a part of the group, it starts its own UL RS transmissions. Using group-based UL RS transmission, interference reduces that allows the RS to be transmitted more frequently and thus reduces the miss detection rate. But UL beacons lead to higher UL messages which might drain the user battery that contradicts with main 5G prerequisites.

The UE position information can be utilized in proactive RRM, network-enabled Device to Device (D2D) communications, self-driving cars, the positioning of a large number of IoT sensors, Intelligent Transportation Systems (ITSs), and mobility management. For proactive HO, the SRSs are utilized in [78] to determine the UE location using Angle of Arrival (AoA) and line of sight path. To provide an energy efficient solution from the user device perspective, positioning algorithms are carried out at the network side. An Extended Kalman Filter (EKF) based solution is formulated in [107] for joint estimation and tracking of the Direction of Arrival (DoA) and Time of Arrival (ToA) of the User Nodes (UNs) using UL RS. In order to fuse the individual DoA and ToA estimates across one or more Access Nodes (ANs) into an accurate UN position estimate, a second EKF stage is used. The additional EKF stage provides an accurate clock offset estimate and reliable clock synchronization of the access link. It is assumed that the locations of the ANs are fully known for 2D positioning (xy-plane only). In future, this work can be extended to 3D positioning, location-based beamforming and mobility management.

E. DISCUSSION AND FUTURE RESEARCH DIRECTIONS

In this section, discussion on the formulation of an efficient HO scheme with some future research directions is provided based on the conducted study. While formulating an efficient HO scheme, the following key points are needed to be considered,

- MSE has great importance to improve the HO performance. Through an accurate MSE, we can choose the HM, A3 event and TTT adaptively. Using the information of the following three time-varying parameters, the number of cell reselections, fading frequency and RSS, user with low, medium or high mobility state can be estimated with improved accuracy. When the users are moving in a cluster with low speed and high micro cell density, consider to power off the macro cell in this cluster for energy saving purpose.
- When a user is in good radio link conditions, start EHOP and execute the HO only when a HO procedure is required. This reduces the cases of RLFs before the HO.
- Estimate the cell edge from RSS and start caching to store the future data contents so that it can be used when the wireless resources are not available. This improves the QoS of the network.
- During cell selection, compare the top three cell utilization and select the cell with low utilization. This procedure will provide an automatic load balancing.
- For high-speed cases, activate the DC mode and connect the user with a macro cell (preferably). Estimate the time user stay at the micro cell from the speed of the user. When a micro cell having good radio link conditions is less utilized and the time user stay at the micro cell is greater than the minimum threshold time, switch the user towards the micro cell.

This survey suggests that the current mobility mechanisms suffer from the degraded performance when the size of cells is diminished and the speed of users is moderate to high. So, a new set of solutions is required that enhances the network performance while at the same time reduces the energy expenditures both at the network and the terminal side. The following ideas suggest the future work to improve the mobility performance in future cellular networks (especially the small cell deployments),

- Implementation of UL RS based mobility in a system-level simulations to compare the performance of HO in small cell deployment scenario with DL based measurement (for example, LTE mobility case). Also, find the potential benefits of using UL RS based mobility for the mobile small cell deployment scenario.
- Realization of the mobile small cells in a simulation scenario with group-based mobility schemes to reduce the signaling overheads during HO and improve energy efficiency.
- Deployment of the small cells with overlapping coverage to perform the dynamic power management (i.e. turning off a small cell) by considering the current load factor.

VI. CONCLUSION

To meet future cellular networks targets, efficient HO management techniques are required possibly with 0ms service interruption time. To make this possible, we considered the HO techniques available in the literature and developed the key points (based on the conducted study) need to be considered while formulating a HO problem.

Initially, we introduced some general concepts of radio access mobility in cellular networks and highlighted the current research focuses and major challenges in these areas. We found that for IDLE_MODE mobility, research efforts are devoted to reduce the paging delays and to provide dynamic TA list (TAL) configuration while for CONNECTED_MODE mobility, HO parameter optimization, early initiation of HO, and reducing signaling overheads are the main research focuses. HO types related challenges are, saving UE battery power, load balancing among RATs, and improving the roaming quality. Current research efforts are also addressing the concepts that allow the UE to add and release different radio links autonomously to reduce the signaling overheads produced during the HO procedure.

Then, we provided key features and entities involved in LTE and NR mobility with a detailed literature survey. It is found that the basic HO scenario in NR is very similar to the LTE except the entities involved and a slight change in HO steps. For example, the basic HO procedure in NR is completed in twelve steps while in LTE it goes through eighteen steps.

As a next step, we found the HO management challenges and techniques to overcome these hurdles. Different solutions for mobility enhancement are elaborated for reducing the service interruption time and it is found that the dual connection based mobility has the lowest service interruption time. Also, HO management in high-speed train scenario can be improved using two-hop network with geo-aided fast HO and considering the Doppler planning at the early stage of the network design. In addition, the UL RS based mobility method improves the network overall performance through the reduction of HO signaling messages because this method does not require MRs between user and network. Also, the group UL RS based mobility improves HO efficiency through minimizing resource consumption.

Lastly, we found key points need to be considered while designing an efficient HO scheme and research directions for future evolution.

REFERENCES

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update 2017–2022," Cisco, San Jose, CA, USA, White Paper c11-738429, Feb. 2019.

[2] R. Arshad, H. Elsawy, S. Sorour, T. Y. Al-Naffouri, and M.-S. Alouini, "Handover management in 5G and beyond: A topology aware skipping approach," *IEEE Access*, vol. 4, pp. 9073–9081, 2016.

[3] 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./Aug. 2017.

[4] G. P. Pollini, "Trends in handover design," *IEEE Commun. Mag.*, vol. 34, no. 3, pp. 82–90, Mar. 1996.

[5] D. L. Huff, "Advanced mobile phone service: The developmental system," *Bell Syst. Tech. J.*, vol. 58, no. 1, pp. 249–269, Jan. 1979.

[6] M. Gudmundson, "Analysis of handover algorithms (microcellular radio)," in *Proc. 41st IEEE Veh. Technol. Conf.*, May 1991, pp. 537–542.

[7] X. Yang, S. Ghaheri-Niri, and R. Tafazolli, "Evaluatfon of soft handover algorithms for UMTS," in *Proc. 11th IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, vol. 2, Sep. 2000, pp. 772–776.

[8] C.-C. Lee and R. Steele, "Effect of soft and softer handoffs on CDMA system capacity," *IEEE Trans. Veh. Technol.*, vol. 47, no. 3, pp. 830–841, Aug. 1998.

[9] D. Lopez-Perez, I. Guvenc, and X. Chu, "Mobility management challenges in 3GPP heterogeneous networks," *IEEE Commun. Mag.*, vol. 50, no. 12, pp. 70–78, Dec. 2012.

[10] *5G NR*. Accessed: Mar. 2018. [Online]. Available: https://www.qualcomm.com/invention/5g/5g-nr

[11] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Commun. Standards Mag.*, vol. 1, no. 4, pp. 24–30, Dec. 2017.

[12] J. Rodriguez, A. Radwan, C. Barbosa, F. H. P. Fitzek, R. A. Abd-Alhameed, J. M. Noras, S. M. R. Jones, I. Politis, P. Galiotos, G. Schulte, A. Rayit, M. Sousa, R. Alheiro, X. Gelabert, and G. P. Koudouridis, "SECRET—Secure network coding for reduced energy next generation mobile small cells: A European training network in wireless communications and networking for 5G," in *Proc. IEEE Internet Technol. Appl. (ITA) Conf.*, Sep. 2017, pp. 329–333.

[13] K. Kanwal, "Increased energy efficiency in LTE networks through reduced early handover," Ph.D. dissertation, Dept. Comput. Sci. Technol., Univ. Bedfordshire, Bedford, U.K., 2017.

[14] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2013, ch. 1, p. 10.

[15] A. Karandikar, N. Akhtar, and M. Mehta, *Mobility Management in LTE Heterogeneous Networks*, vol. 2. Singapore: Springer, 2017, pp. 13–32.

[16] S. Sesia, I. Toufik, and M. Baker, *LTE—The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. Chichester, U.K.: Wiley, 2011, pp. 503–529.

[17] *Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Procedures in Idle Mode*, document 3GPP 36.304, Mar. 2019.

[18] S. M. Razavi, D. Yuan, F. Gunnarsson, and J. Moe, "Dynamic tracking area list configuration and performance evaluation in LTE," in *Proc. IEEE Globecom Workshops*, Dec. 2010, pp. 49–53.

[19] A. Taufique, "On analytical modeling of mobility signaling in ultra dense HetNets," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Oklahoma, Norman, OK, USA, 2018.

[20] *5G;NR; Radio Resource Control (RRC); Protocol Specification*, document 3GPP TS 38.331 Version 15.3.0 Release 15, Jan. 2018.

[21] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2*, document ETSI TS 136 300 V15.3.0 (2018-10), Mar. 2019, pp. 105–158.

[22] S. Barbera, K. Pedersen, P. H. Michaelsen, and C. Rosa, "Mobility analysis for inter-site carrier aggregation in LTE heterogeneous networks," in *Proc. IEEE 78th VTC Conf.*, Sep. 2013, pp. 1–5.

[23] Y. Li, B. Cao, and C. Wang, "Handover schemes in heterogeneous LTE networks: Challenges and opportunities," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 112–117, Apr. 2016.

[24] M. Kassar, B. Kervella, and G. Pujolle, "An overview of vertical handover decision strategies in heterogeneous wireless networks," *Comput. Commun.*, vol. 31, no. 10, pp. 2607–2620, Jun. 2008.

[25] H. Persson and J. M. Karlsson, "An approach to structure the handover terminology," in *Proc. SNCNW*, Halmstad, Sweden, Nov. 2005.

[26] M. Tayyab, G. P. Koudouridis, and X. Gelabert, "A simulation study on LTE handover and the impact of cell size," in *Proc. Broadband Commun., Netw., Syst. (BROADNETS)*, 2018, pp. 398–408.

[27] M. Lauridsen, L. C. Giménez, I. Rodriguez, T. B. Sorensen, and P. Mogensen, "From LTE to 5G for connected mobility," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 156–162, Mar. 2017.

[28] 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.

[29] X. Gelabert, G. Zhou, and P. Legg, "Mobility performance and suitability of macro cell power-off in LTE dense small cell HetNets," in *Proc. IEEE 18th Int. Workshop Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, Sep. 2013, pp. 99–103.

[30] A. Ulvan, R. Bestak, and M. Ulvan, "The study of handover procedure in LTE-based femtocell network," in *Proc. 3rd Joint IFIP IEEE Wireless Mobile Netw. Conf. (WMNC)*, Oct. 2010, pp. 1–6.

[31] M. Joud, M. García-Lozano, and S. Ruiz, "User specific cell clustering to improve mobility robustness in 5G ultra-dense cellular networks," in *Proc. 14th Annu. Conf. Wireless On-Demand Netw. Syst. Services (WONS)*, Feb. 2018, pp. 45–50.

[32] *LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 General Aspects and Principles*, document ETSI TS 136 420 V10.2.0 (2011-10), Jun. 2018.

[33] *LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 General Aspects and Principles*, document ETSI TS 136 410 V10.0.1, Jun. 2018.

[34] M. T. Nguyen, S. Kwon, and H. Kim, "Mobility robustness optimization for handover failure reduction in LTE small-cell networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4672–4676, May 2018.

[35] S. K. Ray, H. Sirisena, and D. Deka, "LTE-advanced handover: An orientation matching- based fast and reliable approach," in *Proc. IEEE Conf. Local Comput. Netw.*, Oct. 2013, pp. 280–283.

[36] K. Kanwal and G. A. Safdar, "Energy efficiency and superlative TTT for equitable RLF and ping pong in LTE networks," *Mobile Netw. Appl.*, vol. 23, no. 6, pp. 1682–1692, 2018.

[37] Y. S. Hussein, B. M. F. A. Ali, A. Sali, and A. M. Mansoor, "A novel cell-selection optimization handover for long-term evolution (LTE) macrocellusing fuzzy TOPSIS," *J. Comput. Commun.*, vol. 73, pp. 22–33, Jan. 2016.

[38] S. Chaudhuri, I. Baig, and D. Das, "Self organizing method for handover performance optimization in LTE-advanced network," *Comput. Commun.*, vol. 10, pp. 151–163, Sep. 2017.

[39] J.-I. Choi, W.-K. Seo, J.-C. Nam, I.-S. Park, and Y.-Z. Cho, "Handover decision algorithm based on available data volume in hierarchical macro/femto-cell networks," in *Proc. 14th Int. Conf. Commun. Electron.*, Aug. 2012, pp. 145–149.

[40] H.-S. Park, Y.-S. Choi, B.-C. Kim, and J.-Y. Lee, "LTE mobility enhancements for evolution into 5G," *ETRI J.*, vol. 37, pp. 1065–1076, Dec. 2015.

[41] J. Zhao, Y. Liu, C. Wang, L. Xiong, and L. Fan, "High-speed based adaptive beamforming handover scheme in LTE-R," *IET Commun.*, vol. 12, no. 10, pp. 1215–1222, 2018.

[42] M. Tayyab, G. P. Koudouridis, X. Gelabert, and R. Jäntti, "Signaling overhead and power consumption during handover in LTE," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Marrakech, Morocco, 2019, pp. 1–6.

[43] C. Liu, J. Wei, S. Huang, and Y. Cao, "A distance-based handover scheme for femtocell and macrocell overlaid networks," in *Proc. 8th Int. Conf. Wireless Commun. Netw. Mobile Comput.*, Shanghai, China, Sep. 2012, pp. 1–4.

[44] A. S. Priyadharshini and P. T. V. Bhuvaneswari, "A study on handover parameter optimization in LTE-A networks," in *Proc. IEEE Conf. Microelectron., Comput. Commun. (MicroCom)*, Jan. 2016, pp. 1–5.

[45] X. Zhang, Y. Xie, Y. Cui, Q. Cui, and X. Tao, "Multi-slot coverage probability and SINR-based handover rate analysis for mobile user in Hetnet," *IEEE Access*, vol. 6, pp. 17868–17879, 2018.

[46] K. Da Costa Silva, Z. Becvar, and C. R. L. Frances, "Adaptive hysteresis margin based on fuzzy logic for handover in mobile networks with dense small cells," *IEEE Access*, vol. 6, pp. 17178–17189, 2018.

[47] Y. Song, P.-Y. Kong, and Y. Han, "Potential of network energy saving through handover in HetNets," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10198–10204, Dec. 2016.

[48] D. Xenakis, N. Passas, and C. Verikoukis, "An energy-centric handover decision algorithm for the integrated LTE macrocell–femtocell network," *J. Comput. Commun.*, vol. 35, pp. 1684–1694, Aug. 2012.

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

[50] A. Ichkov, V. Atanasovski, and L. Gavrilovska, "Analysis of two-tier LTE network with randomized resource allocation and proactive offloading," *Mobile Netw Appl.*, vol. 22, pp. 806–813, Oct. 2017.

[51] F. Guidolin, I. Pappalardo, A. Zanella, and M. Zorzi, "Context-aware handover policies in HetNets," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1895–1906, Mar. 2016.

[52] Z. Ali, N. Baldo, J. Mangues-Bafalluy, and L. Giupponi, "Machine learning based handover management for improved QoE in LTE," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp.*, Apr. 2016, pp. 794–798.

[53] H. Kalbkhani, S. Jafarpour-Alamdari, M. G. Shayesteh, and V. Solouk, "QoS-based multi-criteria handoff algorithm for femto-macro cellular networks," *Wireless Pers. Commun.*, vol. 98, pp. 1435–1460, Jan. 2018.

[54] X. Xu, Z. Sun, X. Dai, T. Svensson, and X. Tao, "Modeling and analyzing the cross-tier handover in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7859–7869, Dec. 2017.

[55] 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.

[56] K. Vasudeva, M. Simsek, D. López-Pérez, and I. Güvenç, "Analysis of handover failures in heterogeneous networks with fading," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6060–6074, Jul. 2017.

[57] *Enhanced Mobility State Estimation by Doppler Frequency Measurements*, document 3GPP R2-124007, NTT DOCOMO, Qingdao, China, Aug. 2012.

[58] Y. Zhu, L. Ni, and B. Li, "Exploiting mobility patterns for inter-technology handover in mobile environments," *Comput. Commun.*, vol. 36, no. 2, pp. 203–210, 2013.

[59] *5G New Radio—NR and NG-RAN Overall Description-Stage 2*, document 3GPP TS 38.300 Release 15, Version 15.5.0, Mar. 2019.

[60] *5G; System Architecture for the 5G System*, document 3GPP TS 23.501 version 15.2.0 Release 15, Jun. 2019.

[61] *Supplementary Uplink (SUL) and LTE-NR Co-Existence*, document 3GPP TR 37.872 Version 15.1.0, Jan. 2019.

[62] O. Semiari, W. Saad, M. Bennis, and B. Maham, "Caching meets millimeter wave communications for enhanced mobility management in 5G networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 779–793, Feb. 2018.

[63] H. Zhang, W. Huang, and Y. Liu, "Handover probability analysis of anchor-based multi-connectivity in 5G user-centric network," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 396–399, Apr. 2019.

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

[65] M. Alhabo and L. Zhang, "Unnecessary handover minimization in two-tier heterogeneous networks," in *Proc. 13th Annu. Conf. Wireless Demand Netw. Syst. Services (WONS)*, Jackson, WY, USA, Feb. 2017, pp. 160–164.

[66] O. Semiari, W. Saad, M. Bennis, and M. Debbah, "Integrated millimeter wave and sub-6 GHz wireless networks: A roadmap for joint mobile broadband and ultra-reliable low-latency communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 109–115, Apr. 2019.

[67] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, "Achieving sustainable ultra-dense heterogeneous networks for 5G," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 84–90, Dec. 2017.

[68] I. Viering, H. Martikainen, A. Lobinger, and B. Wegmann, "Zero-zero mobility: Intra-frequency handovers with zero interruption and zero failures," *IEEE Netw.*, vol. 32, no. 2, pp. 48–54, Mar./Apr. 2018.

[69] M. Alhabo, L. Zhang, and N. Nawaz, "A trade-off between unnecessary handover and handover failure for heterogeneous networks," in *Proc. 23rd Eur. Wireless Conf.*, May 2017, pp. 1–6.

[70] 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, 2019.

[71] S. Barmpounakis, A. Kaloxylos, P. Spapis, and N. Alonistioti, "Context-aware, user-driven, network-controlled RAT selection for 5G networks," *Comput. Netw.*, vol. 113, pp. 124–147, Feb. 2017.

[72] D. Calabuig, S. Barmpounakis, S. Gimenez, A. Kousaridas, T. R. Lakshmana, J. Lorca, P. Lunden, Z. Ren, P. Sroka, E. Ternon, V. Venkatasubramanian, and M. Maternia, "Resource and mobility management in the network layer of 5G cellular ultra-dense networks," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 162–169, Jun. 2017.

[73] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[74] M. Alhabo and L. Zhang, "Multi-criteria handover using modified weighted TOPSIS methods for heterogeneous networks," *IEEE Access*, vol. 6, pp. 40547–40558, 2018.

[75] M. Alhabo and L. Zhang, "Load-dependent handover margin for throughput enhancement and load balancing in HetNets," *IEEE Access*, vol. 6, pp. 67718–67731, 2018.

[76] M. Alhabo, L. Zhang, and O. Oguejiofor, "Inbound handover interference-based margin for load balancing in heterogeneous networks," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Bologna, Italy, Aug. 2017, pp. 146–151.

[77] R. Arshad, H. ElSawy, S. Sorour, T. Y. Al-Naffouri, and M.-S. Alouini, "Velocity-aware handover management in two-tier cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1851–1867, Mar. 2017.

[78] N. Malm, L. Zhou, E. Menta, K. Ruttik, R. Jäntti, O. Tirkkonen, M. Costa, and K. Leppänen, "User localization enabled ultra-dense network testbed," in *Proc. IEEE 5G-WF Conf.*, Jul. 2018, pp. 405–409.

[79] B. Yang, X. Yang, X. Ge, and Q. Li, "Coverage and handover analysis of ultra-dense millimeter-wave networks with control and user plane separation architecture," *IEEE Access*, vol. 6, pp. 54739–54750, 2018.

[80] 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./Apr. 2018.

[81] E. Demarchou, C. Psomas, and I. Krikidis, "Mobility management in ultra-dense networks: Handover skipping techniques," *IEEE Access*, vol. 6, pp. 11921–11930, 2018.

[82] H. Qu, Y. Zhang, J. Zhao, G. Ren, and W. Wang, "A hybrid handover forecasting mechanism based on fuzzy forecasting model in cellular networks," *China Commun.*, vol. 15, no. 6, pp. 84–97, Jun. 2018.

[83] H. S. Park, Y. Lee, T. J. Kim, B. C. Kim, and J. Y. Lee, "Handover mechanism in NR for ultra-reliable low-latency communications," *IEEE Netw.*, vol. 32, no. 2, pp. 41–47, Mar. 2018.

[84] *Mobility Enhancements for NR*, document 3GPP TSG-RAN WG2 #101, R2-1801883, Samsung, Mar. 2018.

[85] *Service Requirements for the 5G System; Stage 1 (Release 16)*, document 3GPP TS 22.261 Version 16.8.0, Jun. 2019.

[86] P. Fan, J. Zhao, and C.-L. I, "5G high mobility wireless communications: Challenges and solutions," *China Commun.*, vol. 13, no. 2, pp. 1–13, 2016.

[87] X.-P. Ma, H.-H. Dong, P. Li, L.-M. Jia, X. Liu, Y. Qin, and J.-Q. Tang, "Adaptive optimization of multi-hop communication protocol for linear wireless monitoring networks on high-speed railways," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2313–2327, Jun. 2019.

[88] Y. Cao, N. Wang, C. Wu, X. Zhang, and C. Suthaputchakun "Enhancing video QoE over high-speed train using segment-based prefetching and caching," *IEEE MultiMedia*, to be published.

[89] Y. Liu, C.-X. Wang, J. Huang, J. Sun, and W. Zhang, "Novel 3-D non-stationary mmWave massive MIMO channel models for 5G high-speed train wireless communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2077–2086, Mar. 2019.

[90] Y. Liu, C.-X. Wang, C. F. Lopez, G. Goussetis, Y. Yang, and G. K. Karagiannidis, "3D non-stationary wideband tunnel channel models for 5G high-speed train wireless communications," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[91] J. Yang, B. Ai, S. Salous, K. Guan, D. He, G. Shi, and Z. Zhong, "An efficient MIMO channel model for LTE-R network in high-speed train environment," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3189–3200, Apr. 2019.

[92] M. Munjal and N. P. Singh, "Group mobility by cooperative communication for high speed railway," *Wireless Netw.*, pp. 1–10, Jan. 2019.

[93] W. Li, C. Zhang, X. Duan, S. Jia, Y. Liu, and L. Zhang, "Performance evaluation and analysis on group mobility of mobile relay for LTE advanced system," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2012, pp. 1–5.

[94] M. Sternad, M. Grieger, R. Apelfröjd, T. Svensson, D. Aronsson, and A. B. Martinez, "Using 'predictor antennas' for long-range prediction of fast fading for moving relays," in *Proc. IEEE WCNC Workshop*, Apr. 2012, pp. 253–257.

[95] D. Oliva and J. I. Alonso, "A two-hop MIMO relay architecture using LTE and millimeter wave bands in high speed trains," in *Proc. IEEE WCNC Conf.*, Apr. 2018, pp. 1–6.

[96] P. T. Dat, A. Kanno, K. Inagaki, F. Rottenberg, N. Yamamoto, and T. Kawanishi, "High-speed and uninterrupted communication for high-speed trains by ultrafast WDM fiber–wireless backhaul system," *J. Lightw. Technol.*, vol. 37, no. 1, pp. 205–217, Jan. 1, 2019.

[97] W. Luo, R. Zhang, and X. Fang, "A CoMP soft handover scheme for LTE systems in high speed railway," *EURASIP J. Wireless Commun. Netw.*, vol. 2012, Jun. 2012, Art. no. 196.

[98] Q. Luo, W. Fang, J. Wu, and Q. Chen, "Reliable broadband wireless communication for high speed trains using baseband cloud," *EURASIP J. Wireless Commun. Netw.*, vol. 2012, Sep. 2012, Art. no. 285.

[99] W. Guo, W. Zhang, P. Mu, F. Gao, and H. Lin, "High-mobility wideband massive MIMO communications: Doppler compensation, analysis and scaling laws," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3177–3191, Jun. 2019.

[100] B. Lannoo, D. Colle, M. Pickavet, and P. Demeester, "Radio-over-fiber-based solution to provide broadband Internet access to train passengers [topics in optical communications]," *IEEE Commun. Mag.*, vol. 45, no. 2, pp. 56–62, Feb. 2007.

[101] J. Liu, K. Au, A. Maaref, J. Luo, H. Baligh, H. Tong, A. Chassaigne, and J. Lorca, "Initial access, mobility, and user-centric multi-beam operation in 5G new radio," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 35–41, Mar. 2018.

[102] S. Kutty and D. Sen, "Beamforming for millimeter wave communications: An inclusive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 949–973, 2nd Quart., 2016.

[103] 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., 2018.

[104] E. Onggosanusi, S. Rahman, L. Guo, Y. Kwak, H. Noh, Y. Kim, S. Faxer, M. Harrison, M. Frenne, S. Grant, R. Chen, R. Tamrakar, and 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.

[105] X. Gelabert, C. Qvarfordt, M. Costa, P. Kela, and K. Leppänen, "Uplink reference signals enabling user-transparent mobility in ultra dense networks," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1–6.

[106] H. Lundqvist, G. P. Koudouridis, and X. Gelabert, "Joint tracking of groups of users with uplink reference signals," in *Proc. IEEE 22nd Int. Workshop Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, Jun. 2017, pp. 1–5.

[107] M. Koivisto, M. Costa, J. Werner, K. Heiska, J. Talvitie, K. Leppänen, V. Koivunen, and M. Valkama, "Joint device positioning and clock synchronization in 5G ultra-dense networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2866–2881, May 2017.

MUHAMMAD TAYYAB received the B.Sc. degree from the University of the Punjab, Lahore, Pakistan, in 2012, and the M.Sc. degree (Hons.) from the King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, in 2017, both in electrical engineering. He is currently pursuing the Ph.D. degree in electrical engineering with Aalto University, Finland. He has more than two years, from 2013 to 2015, of professional experience as an RF Planning and Optimization Executive at Wi-Tribe Pakistan Ltd. (Ooredoo Group). During his stay at KFUPM, he was associated with the King Abdullah University of Science and Technology (KAUST), Saudi Arabia, as a Visiting Student and a KFUPM–KAUST Joint Research Initiative. He has been a Researcher with the Helsinki Research Center, Huawei Technologies Finland Oy, since February 2018. His current research interest includes energy-efficient mobility for small-cell overlaid cellular networks. He received the Gold Medal Award for obtaining the first position in the B.Sc. degree.

XAVIER GELABERT received the M.Sc. degree in electrical engineering from the KTH Royal Institute of Technology, Stockholm, in 2003, and the joint B.Sc. and M.Sc. degrees in telecom engineering and the Ph.D. degree (Hons.) from the Technical University of Catalonia (UPC), Barcelona, in 2004 and 2010, respectively. Over the past 15 years, he has held research positions in academia (UPC, Barcelona, GATech, Atlanta, UPV, Valencia, and KCL, London), a non-profit research center (iTEAM, Valencia), a Telco Operator (Orange Labs, Paris), and a Telecom Vendor (Huawei, Stockholm). He has been a Senior Research Engineer with Huawei Technologies Sweden AB, since 2012.

RIKU JÄNTTI (M'02–SM'07) received the M.Sc. degree (Hons.) in electrical engineering and the D.Sc. degree (Hons.) in automation and systems technology from the Helsinki University of Technology (TKK), in 1997 and 2001, respectively. He was a Professor Pro Tem with the Department of Computer Science, University of Vaasa. In August 2006, he joined Aalto University, Finland, where he is currently a Professor in communications engineering and the Head of the Department of Communications and Networking, School of Electrical Engineering. His research interests include radio resource control, spectrum management, and performance optimization of wireless communication systems. He is an Associate Editor of the IEEE Transactions on Vehicular Technology.