Beyond 5G: Reducing the Handover Rate for High Mobility Communications

Naor Zohar

***Abstract*—The fifth-generation (5G) and beyond cellular networks are expected to support a huge number of mobile devices, roaming seamlessly across very small cells. Consequently, the handover rate for these extremely dense networks is expected to be very high. To reduce the burden caused by rapid handover requests, and to support a massive number of highly mobile devices in 5G and beyond networks, this study suggests using proximity-based clusters as nomadic cells integrated with Aerial Access Networks (AANs). These nomadic cells are formed by two-levels hierarchical partitioning of the mobile devices into proximity-based clusters.**

Previous distributed mobility management schemes are not sufficiently efficient to support the handover rate expected for 5G and beyond networks. Due to their high computational complexity, previous group-based methods are not applicable for real-time services. In contrast to these schemes, the proposed scheme is scalable with the number of devices. Moreover, the creation of a mobility group raises practical as well as security and privacy issues that were overlooked by previous schemes. These issues are addressed in this study.

***Index Terms*—5G and beyond networks, aerial access networks, cellular networks, mobility management.**

I. Introduction

5G and beyond cellular networks are required to support a massive number of Internet of things (IoT) devices and provide seamless access with Quality-of-Service (QoS) guarantees for all of them. 5G cellular networks are expected to support communications with high mobility. The term "high mobility" does not necessarily refer only to the velocity of the mobile devices. Rather, this term refers to the challenges caused by mobility. For instance, the rate of network disconnection events caused by handover. In general, the challenges caused by mobility depend, among other things, on the users' velocity and density, the cell size, and the required network latency and QoS. Examples for applications in high mobility scenarios are high-speed railways, vehicular ad hoc networks, and unnamed aerial vehicle (UAV) communications.

5G and beyond networks are expected to support real-time services for devices such as autonomous cars, drones, and other smart vehicles. These services require precise knowledge and a very low latency about the exact location of these devices, to react in real-time [1]–[3]. That implies that the exact location of the mobile device (e.g., a UAV) must be known accurately. While existing cellular networks have the time to search for a mobile user upon request, due to their short latency constraint [2], [3], 5G and beyond networks may not always have this privilege.

It follows from the above discussion that 5G networks will have to support tenths of billions of highly mobile devices, subject to a low latency [2]–[4], [6] and high location accuracy [1], [6] constraints. Since there is a clear trade-off between the rate of the mobile user location update and the uncertainty in its location whereabouts, the signaling cost associated with mobility in 5G and beyond networks is expected to be significantly higher than the equivalent cost in existing cellular networks. This cost increment is expected for both network cost, as well as for each mobile node.

Moreover, since beyond 5G networks are expected to use smaller cells arranged hierarchically based on macro-cells, micro-cells and femtocells, the signaling cost associated with mobility management should be significantly higher for beyond 5G networks, in comparison with the equivalent cost in 3G and 4G networks. Therefore, there is a need to reduce the signaling cost associated with mobility, especially for supporting a massive number of highly mobile devices.

A. Background and Related Work

IP mobility support is provided for IPv4 by MIPv4 [7], [8], and for IPv6 by MIPv6 and its derivatives, such as PMIPv6 [9], fast proxy mobile IPv6 (FPMIPv6) [10], and FH-PMIPv6 [11]. However, these protocols are not sufficiently efficient to support real-time applications, in terms of high handover latency, and packet loss ratio [11], [12], [13]. Therefore, none of these protocols can support devices in high mobility scenarios.

Recently, it was shown in [14] that due to the small dense cells architecture of 5G networks, physical layer methods used in existing cellular networks for detecting handover may not work properly for 5G extreme high mobility scenarios. The authors in [14] extended the cross band channel prediction proposed in [15] to mobility scenarios, to suggest a new physical layer method for handover detection. The focus of this study is on the networking layer. Therefore, the proposed method can be integrated with the scheme suggested in [14].

Several studies attempted to reduce the signaling cost associated with mobility support. Distributed mobility management (DMM) was proposed in [16], and described in [17]. A Software defined network (SDN)-based version of DMM was suggested in [18]. However, DMM is a network-based scheme, aiming to suggest a solution only for the core network. Hence, the problem of handling frequent handover requests at the network edge remains open.

Group-based mobility support methods are based on the proposal to apply the network mobility (NEMO) [19] concept.

Manuscript received August 22, 2021; revised January 3, 2022; approved for publication by Hongliang Zhang, Guest Editor, December 30, 2021.

N. Zohar is with University of Haifa, Israel, email: zohar@math.haifa.ac.il.
N. Zohar is the corresponding author.

Digital Object Identifier: 10.23919/JCN.2022.000001

1229-2370/22/$10.00 © 2022 KICS

The key idea of the NEMO proposal is very simple: In the case where several nodes move as a group, it is possible to select the node with the largest computational capability as the group leader. This leader is used as a mobile router that delivers the packets for all the other nodes in the group. The leader also performs the mobility management signaling on behalf of all the nodes in the group.

NEMO proposal [19] suggested an extension to MIPv6 protocol which enables a node to perform network mobility on behalf of other nodes, using IP-in-IP encapsulation. Other aspects of network connectivity, such as privacy and security, though mentioned in [19], are not considered. Besides, the criterion of how to select the "leader" which performs the mobility on behalf of other nodes, is not well defined. Several schemes were proposed to extend PMIPv6 and its derivatives to support the network mobility NEMO scheme [20]–[24]. All these schemes form the mobility groups based on their mobility patterns.

A group-based approach was suggested in [25]. The key idea is to use a group mobility management (GMM), in which the central database partitions the mobile devices into groups, based on the similarity of their mobility patterns, as recorded at the central database. The incentive for this grouping is to reduce the congestion on the signaling random access channel (RACH). However, this reduction is achieved at the expense of complicating the central database. Moreover, each mobile device must initially bind itself to the network. Therefore, the time required to identify and create the mobility group may violate the network short-latency constraint [2]–[4], [6]. For this reason, the concept of GMM is not applicable for real-time applications.

Previous studies based on NEMO proposal [19] used a network-centralized approach, in which a network element partitions the mobile nodes into disjoint groups, based on their mobility patterns [20]–[25]. For each group, the node with the largest computational capability is selected as the group leader, which performs the mobility on behalf of all the other nodes in its group. As it is shown in this study, the problem of partitioning a given set of independent nodes into mobility groups is NP-hard. Therefore, the previous group-based methods attempted to handle mobility are not scalable. They cannot support the expected huge number of billions of IoT devices. Furthermore, the concept of external partition of independent nodes into groups raises practical, as well as privacy and security issues, which must be considered. The issue of privacy and security was not considered in previous group mobility management schemes. In reality, NEMO [19] -based schemes are not implemented in real networks.

Previous group-based methods [20]–[25] do not apply to real-world networks for two main reasons: i) Due to their high computational complexity, they cannot support real-time and delay-sensitive services. ii) They do not address the issue of bandwidth limitation. The usage of one user as a gateway for many peers makes this user (the group leader) a bottleneck, in terms of bandwidth consumption. Both issues are addressed in this study. Besides, this study addresses practical, latency, and privacy and security issues that were overlooked by the studies cited above.

This study is an enhanced version of [26], presented in IEEE SMDS 2021, in conjunction with IEEE SERVICES 2021.

B. Contributions of This Work

This study suggests a mobility management scheme that is scalable and therefore feasible, that can support highly mobile devices subjected to the short-latency constraint imposed on 5G and beyond networks. As opposed to previous group-mobility schemes which are not scalable, and therefore they have never been implemented in real networks, the scheme proposed in this study can be easily implemented.

NEMO-based mobility groups are formed based on the similarity of the mobility patterns of their members. As it is shown in this study, this approach is not scalable. This study suggests using proximity-based clusters formed locally by the users and communicating with AANs as nomadic cells, to support a massive number of highly mobile devices in 5G and beyond cellular networks. The goal is to provide continuous network connectivity services by reducing the congestion on the physical random access channel (PRACH), and the rate of handover requests, subject to the short-latency constraint imposed on 5G networks [2]–[4], [6]. As opposed to previous mobility management methods which are network-centric, our scheme is user-centric, and therefore scalable.

The key idea of our scheme is to exploit the capability of 5G networks to provide computing and storage resources within the edge of the radio access network (RAN). While traditionally base stations are network elements used only for transmission, we use proximity-based clusters formed by the users as nomadic cells. To support these nomadic clusters, the LTE-advance 3GPP releases 9 and 10 specifications with the support for the combination of large macro cells with small cells, and release 12 specifications with the support for small cells deployment in dense areas can be adapted to 5G networks [27]. These nomadic cells are used as dual-mode BSs that integrate millimeter-wave and microwave frequencies, as suggested in [29]. Thus, they can communicate either directly with the RAN, or with a large macro cell. Consequently, the rate of handover requests is expected to be significantly reduced. This approach is especially suitable for highly mobile devices, such as high-speed railways, or buses moving in an urban area, etc.

To achieve sufficiently high scalability that can support tenths of billions of deployed IoT devices, our solution is based on utilizing the concept of home agent (HA), similarly to mobile IP-like protocols. In terms of scalability, this concept is preferable over a solution based on home location register (HLR) and visitor location register (VLR), which are widely used in cellular networks.

The main contributions of this study are:

1) A scalable user-based distributed scheme, while previous studies used a centralized approach which is not scalable.
2) A significant reduction in the handover rate handled by the network. This contribution is crucial for highly mobile devices moving in 5G and beyond networks.
3) Privacy and security issues, that were overlooked by previous group mobility schemes can be potentially

considered. Mobile devices are more protected against faraway hostile attacks, in comparison with previous group-based schemes, for which the mobile nodes are vulnerable to such attacks.

4) Packet loss ratio can be significantly reduced, in comparison with proxy-based schemes, since the network proxy is very close to the users.

C. Paper Organization

The rest of this paper is organized as follows: Model and problem formulation are given in Section II. Our scheme is introduced in Section III, and analyzed in Section IV. Performance comparison with other methods is given in Section V. Simulation results are described in Section VI. Finally, summary and concluding remarks are provided in Section VII.

II. Model and Problem Formulation

The problem we face is to support high mobility communication in highly congested dense cellular networks. The focus of this study is on the mobility management challenges caused by high mobility. For instance, the need to handle simultaneously rapid handover requests for many independent mobile devices.

We consider an all-IP wireless network, consisting of a set of base stations (BSs), and a set of roaming mobile devices, referred to as mobile nodes (MNs). An MN can move seamlessly across the network. The MN can be, for instance, a smartphone, a wearable device, or any mobile device. A BS (defined as eNB in LTE-A) is the network interface to the mobile devices, via a wireless link that connects the BS to the mobile devices within its coverage area. The network service area is partitioned into zones, based on the coverage (service) area of each BS. An aerial access network (AAN) is a heterogeneous network that is engineered to utilize an airborne platform, such as a drone, or a satellite, to build a network access platform that enables "connectivity from the sky". The service area of an AAN considered in this study is partitioned into zones that are typically significantly larger than the average cell size in 5G and beyond networks, due to the usage of long-range communication between the AAN and the devices it communicates with. Therefore, a BS can be either terrestrial or airborne. It is assumed that time is slotted. This assumption holds for all standardized cellular networks. For the sake of simplicity, this study uses a 5G network as the model. However, it can be easily applied to any IP network, and specifically for 3G and beyond cellular networks.

Given that there is an average of σ sessions exist simultaneously in a cell c, and that the user probability to move from c to another cell during a time slot is ϕ, the expected rate of handover requests originated from c is given by:

$$Handover_{rate} = \sigma\phi, \tag{1}$$

events per time slot. Since the cell size in 5G networks is significantly smaller, and the users' density is expected to be significantly larger than the equivalent cell size and density in existing cellular networks, the rate of handover requests is expected to increase for 5G networks. Our goal is to reduce the expected rate of handover requests $Handover_{rate}$, and the signaling load on the PRACH.

III. The Mobility Management Scheme

The proposed mobility management scheme is based on two-levels hierarchical partitioning of the MNs into proximity-based clusters formed (at the bottom hierarchical level) by the human users. Each cluster is managed by its cluster head (CH), which is used as a server that provides network connectivity services and manages the mobility of its clients, i.e., the MNs within its proximity, on behalf of these devices. The client-server connection is established by a proximity-based authentication process. The CH is defined as an MN which can support IP, has sufficient computational capability and power capacity to manage the MNs within its cluster, and has the sufficient bandwidth required to support them. In terms of NEMO [19] proposal, the CH is the group leader, which serves all the members in its group. However, as opposed to NEMO-based schemes, the CH must be significantly superior to its cluster members, in terms of bandwidth, computational, and power capacities. The CH has sufficient bandwidth and processing power to support its cluster members without becoming a bottleneck. The CH can be either a smartphone or an AAN interface, as will be explained later.

The clusters are formed hierarchically. In the first (bottom) level, human user equipment, for instance, a smartphone, is used as the CH of the same user wearable devices such as his/her smartwatch, as long as these devices are within its proximity. Many smartphone producers leverage their smartphones to detect IoT devices. For instance, in [30] this feature was used for IoT device authentication, in which the human user is required to perform one of two hand gestures. A smartphone - smartwatch communication is already used commercially. Upon entering a car, this cluster becomes a sub-cluster of the second (upper) level, managed by a car-installed device as its CH. Using a proximity-based authentication process, the persons sharing the same car can attach their smartphones to the car-installed CH, thus reducing the amount of radio signaling messages associated with mobility even further. This car is represented by its CH as an MN. As opposed to previous group-based schemes, this car-installed CH should support the bandwidth required by up to a pre-defined number of smartphones. A group of persons sharing the same public transportation, for instance, a bus, or a train, can use a vehicle-installed device as their CH. To form such a cluster there is a need to equip this vehicle with such a device, which acts as a "mini" BS (i.e., a nomadic cell), entitled to a dynamic bandwidth allocation up to a pre-defined number of smartphones. This nomadic cell can communicate either with an AAN, or as a small nomadic cell moving in a large macro cell, or directly with the RAN.

Implementation: These nomadic cells can be incrementally deployed by adapting the LTE-A releases 9,10 and 12 specifications with the proposal of small cell deployment in dense areas to support the combination of small nomadic cells with large macro cells by 5G networks [27].

The goal of this vehicle-installed device is to solve two issues that were overlooked by previous group-based schemes: i) The need to provide the CH a sufficient bandwidth that can support its group members and ii) the need to reduce the rate of handover requests. These goals can be achieved by using the CH as a "mini" BS which can communicate either with an AAN or directly with the RAN, or with a large macrocell.

The CH initiates the signaling required to bind the MNs within its cluster to the network and is responsible for handling the network connectivity of the MNs within its proximity. The client MN uses a short-range communication with its CH, which is used also as its gateway to the cellular network. Therefore, the client MN is more protected against hostile attacks, as explained in [30].

The size of the proximity zone in which the CH is responsible for the mobility management of its clients depends on the CH. For instance, a smartphone used as a CH can sense and handle the mobility of wearable devices carried by the same person who owns the smartphone.

The mobility support provided by the CH to its clients consists of a) Authenticating the nearby MN. b) Register the MN in its list and update the relevant HA database. c) Routing the information in and from the MN. d) Updating the MN HA on any location update of the CH. Upon receiving a registration message from the local CH, the MN HA sends a de-registration message to the previous CH, which deletes the MN from its list.

In addition to its usage as a "mini" BS, the vehicle-installed CH functions similarly to the local mobility anchor (LMA) and the mobility access gateway (MAG) defined in PMIPv6, with two major differences: i) The CH is user equipment, while the MAG and the LMA are network elements. ii) The CH actively manages the network connectivity of its clients, while LMA and MAG just respond to the messages transmitted by the MN. That is, as opposed to previous group-mobility schemes [20]–[25], the initial process that binds the MN to the network is initiated and conducted solely by the CH, not by the MN itself. Consequently, the time duration required to establish the mobility group, and the signaling cost (required from both the network and the MN) are both significantly reduced, in comparison with the previous group mobility schemes cited above. Since the partition of the MNs into clusters is conducted hierarchically, and since each cluster is created based on proximity-based authentication, we refer to the proposed mobility management scheme as hierarchical proximity-based consolidation (HPC).

A. HPC - A Formal Description

The HPC scheme is as follows:

1) Initialization: For privacy and security reasons, a proximity-based an authentication mechanism is used. The human user is required to bind its MN to the CH, using a short-range communication (e.g., Bluetooth), similarly to the existing binding process that binds a smartphone to a car-installed device. A smartphone can be used as a CH for the IoT devices carried by the same person who owns the smartphone, by performing hand gestures in front of these devices. Here, the authentication mechanism is similar to the one described in [30]. Once the authentication process is completed, the CH initiates a network binding process on behalf of the MN.
2) The CH sends a registration message to the HA of the MN, on behalf of the MN.
3) As long as the CH can sense the MN, the CH is responsible for maintaining the MN address reachable, by using the CH IP address as the MN address. The MN ID is used internally in a table maintained by the CH, while externally every message directed to or from the MN uses the CH IP address, as described in detail in NEMO proposal [19].
4) The HA of the MN updates the CH address as the MN current address, and sends a de-registration message to the previous CH, that informs the previous CH that the MN is no longer under its responsibility.
5) The previous CH updates its list of MNs, and the MN is deleted from this list. The previous CH sends a de-registration acknowledge message to the MN HA.

The binding process that connects the MN to the network is as follows:

1) Whenever the CH authenticates a new MN within its service area, it reads the MN ID and HA.
2) The CH sends a registration message to the MN HA, which informs the HA that the MN is now residing within its service area.
3) The HA updates its associated database and sends a registration acknowledge message to the CH.
4) The CH updates its records, and the MN is added to the list of MNs handled by the CH.
5) The CH sends a registration completion message to the MN HA, which confirms that now the MN is handled by this CH.
6) The MN HA sends a registration cancelation message to the previous CH, which handled the MN until now.
7) The previous CH sends a registration cancelation acknowledge message to the MN HA and deletes the MN record from its list.
8) Upon receiving the registration cancelation message acknowledge, the MN HA updates its associated database and deletes the previous CH from its database.

B. HPC - An Illustrative Example

Fig. 1 illustrates the HPC feasibility in a real-world scenario. The handoff rate reduction is conducted in two steps. In the first step, at the bottom hierarchical level, the human user smartphone handles the mobility of all the wearable devices carried by this user as their CH. Consequently, the number of MNs is reduced to the number of human users. This is still a large number, but feasible, since existing cellular networks handle their human users very efficiently. In the second step, a shared vehicle (e.g., a bus) is used as the CH of its passengers. As illustrated in Fig. 1, this step reduces the number of MNs further to the number of vehicles. Moreover, the shared vehicle, being recognized as a nomadic "mini" BS, uses an AAN for network access. Thus, the handoff rate is significantly

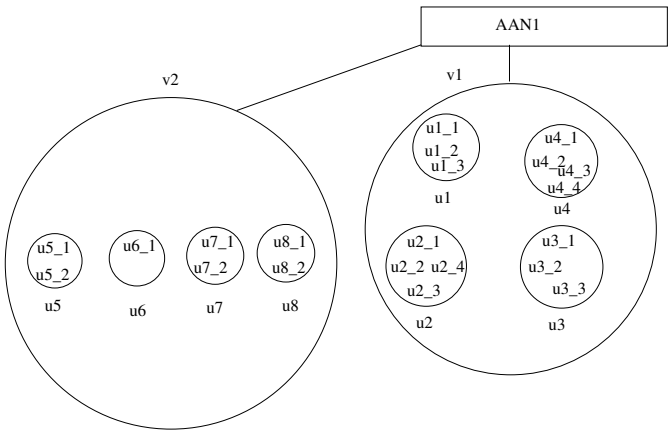

Fig. 1. System description: An illustrative example for an AAN which supports two vehicles and their passengers.

reduced, since the number of the AAN cells is significantly smaller, and their typical size is significantly larger, than the number and typical size of 5G cells. Note that the AAN cells are used to support highly mobile devices, such as trains, buses, etc. Therefore, the network is partitioned into two portions - the conventional portion supports lowly mobile users and devices, while the other portion is composed of AANs - dedicated to supporting highly mobile devices. Fig. 1 describes 8 human users traveling in two shared vehicles, $v1$ and $v2$. The users $u1$ - $u4$ use $v1$, while users $u5$ - u_8 use $v2$. Each user wears several wearable IoT devices, as shown in Fig. 1. For instance, the user $u1$ wears the devices $u1_1, u1_2, u1_3$. Each vehicle communicates directly with the AAN $AAN1$, which handles 8 human users carrying 21 wearable IoT devices, arranged in two levels: In the top level we have $v1$ and $v2$, each of them is the CH of its cluster, while in the bottom(first) level we have 8 smartphones, each of which is the CH of the wearable devices carried by the human user who owns the smartphone.

IV. HPC Analysis

The potential capability of HPC to reduce the handover rate depends on two parameters: (i) The ratio of the cell size of the AAN with which the CH communicates, to the average cell size in the area in which the CH is moving, and (ii) the number of MNs handled by the CH. That is the average cluster size. Since the cell size is expected to shrink for 5G and beyond networks, the potential benefit of using AANs with large service areas and large cell size to support highly mobile devices is expected to be very significant.

Indeed, using a vehicle-installed device as a "mini" nomadic BS implies additional equipment. Since HPC is a proximity-based scheme, there is a tradeoff between the cluster size and the proximity requirement. This tradeoff is inherent to the cellular structure of HPC, for which a large AAN cell supports many mobile clusters. Exactly as existing cellular networks overcome this tradeoff by splitting congested cells, so does HPC. Since an AAN is a complementary network, integrated with the terrestrial cellular network, to accommodate with larger number of highly mobile nodes in denser networks, there is a need for additional equipment. The vehicle-installed CH devices are used to access the AAN.

In this section, the HPC scheme is analyzed and compared with mobility management schemes, such as DMM [16]–[18], and previous NEMO-based [19] methods, from aspects of practicality, scalability, the radio signaling cost associated with mobility, privacy and security, and packet loss ratio.

Since the clusters established by HPC are formed by a short-range communication (e.g., Bluetooth) between the CH and its cluster members, the CH is the only entity revealed to the network. Therefore, the HPC scheme can be integrated with a network-based mobility management scheme, such as DMM and its derivatives [16]–[18]. That is, the CH can implement any network-based strategy on behalf of itself and its cluster members. Therefore, the performance of the HPC scheme should be no worse than DMM, or any network-based scheme. However, due to the consolidation of the cluster members, the HPC signaling cost associated with mobility over the wireless link must be reduced, in comparison with these schemes.

The first aspect to be considered is scalability. Below it is shown that the NEMO-based approach that consolidates similar mobility patterns into mobility groups is not scalable. The reason for this claim is that given a set of independent mobile devices, the problem of partitioning this set into a minimal number of mobility groups is NP-hard.

Theorem IV-.1: Given a set of independent elements (MNs), the problem of partitioning the set into a minimal number of (mobility) groups is NP-hard.

Proof: We prove that this problem is NP-hard by a reduction from the set cover problem, which is known to be NP-hard [31]. Given a universe U and a family S of subsets of U, we define a cover of U as a collection of sets s, $s \in S$, such that the union of the collection is U. Our goal is to find the minimal number of sets in S which cover U. Given a universe U and a family S of subsets of U, we reduce the original set cover problem to the following problem: We substitute each element $i \in U$ with an MN x. We substitute each set $s \in S$ with a group of MNs to be referred to as "mobility group". Without loss of generality, we define a mobility group as a subset of U. Note that an MN x can belong to several mobility groups of devices, moving in the same direction at the same time and in the same place. The partition of the group of all MNs into a minimal number of mobility groups is the same partition that solves the original set cover problem. ■

It follows from Theorem IV-.1 that the NEMO-based approach used by the studies cited above is not scalable with the number of devices. Moreover, even a heuristic algorithm is not practical, since it still must identify the users' mobility patterns of a huge number of MNs, and then consolidate the devices having the same mobility pattern into (not necessarily minimal) disjoint groups, and finally to announce this partition to the participating MNs, subject to a short network latency constraint. For instance, drones and autonomous cars cannot afford the time latency required to establish a NEMO [19] based mobility group. On the other hand, the user-initiated clusters established by HPC are formed instantly. While ex-

isting cellular mobility management schemes (over the wireless link) are distributed and scalable, previous group-based schemes are centralized and they are not scalable.

In contrast to the NEMO-based approach, this study uses a distributed scheme. The MNs are NOT independent, and their partitioning is based on proximity-based clusters, that communicate with AANs. There is no attempt to find an optimal solution. HPC is a heuristic algorithm that finds an approximated solution. As follows from Section III, HPC has linear computational complexity. It should be noted that although it seems that the mobility group partitioning problem is a special case of the set cover problem, in the real world it is NOT. The reason for that observation is that, as opposed to the set cover problem, the mobility groups must be disjoint. Therefore, in reality, we cannot take the opposite direction and solve the mobility group partitioning problem by a reduction to the set cover problem. Given a partition of the set of all mobile devices into a minimal number of (not necessarily disjoint) mobility groups that cover U, which is the solution to the set cover problem, for each element x which belongs to several sets in the minimal solution which covers U, there is a need to delete x from all the sets in this partition but one set. In practice, that means that the network must select for each such element the mobility group which is the most suitable. Given the tight delay constraints imposed by 5G and beyond networks, this requirement is not realistic for the expected huge number of mobile devices that these networks are expected to serve.

NEMO [19] based approach is not scalable with the number of users in the mobility group, in terms of bandwidth consumption. Since the group leader is a user device, elected externally by the network, it may form a bottleneck that cannot provide the bandwidth demands of its peers. For instance, one smartphone cannot provide the bandwidth consumed by many smartphones running simultaneously and independently several video applications. This situation is avoided by HPC. As opposed to NEMO architecture, the CH is defined as such that it can always provide network connectivity services to its cluster members. For instance, at the bottom level, a smartphone serves the wearable IoT devices that belong to the same person who owns the smartphone. Thus, we have a single human user controlling its own devices with its most powerful device. Vehicle-installed devices in cars, buses, or trains are dedicated "mini" BSs designed as nomadic cells to support a pre-defined number of devices, based on the passengers' capacity of each vehicle. Hence, the CH is expected to support the bandwidth demands of its cluster members.

The HPC scheme offers two mechanisms to overcome the limitations of previous group-based methods. The first mechanism is the distributed user-based mechanism for creating proximity-based clusters (in contrast to NEMO-based schemes, that rely on the users' mobility patterns). The second mechanism is the usage of the CH installed in buses, or trains, as a nomadic cell ("mini" BS). A promising approach to enhance mobility and handover in highly mobile networks is to deploy dual-mode BSs that integrate millimeter-wave and microwave frequencies [29]. A vehicle-installed CH can exploit this approach to improve mobility and reduce the handover rate. Thus, the CH can communicate either with an AAN, or directly with the RAN, or with a large macro-cell. This mechanism enables a reduction of the rate of handover requests. This improvement is very significant for 5G and beyond networks, in which the cell size is much smaller, in comparison with previous cellular generations. It should be noted that NEMO-based schemes cannot offer this advantage since these schemes just (externally) consolidate several MNs moving in the same direction into one mobility group. An attempt to integrate a NEMO [19] based scheme with vehicle-to-vehicle (V2V) communication raises several problems. For instance, nearby cars moving in the same direction should be considered by this approach as one mobility group. The establishment of such a group contradicts the basic concept of V2V communication: The vehicles must be considered as independent individual items communicating and exchanging information with their neighbors. NOT as one group managed by a single node. The dependency of the group members on their group leader is not consistent with the concept of V2V communication. As opposed to this scheme, HPC considers each vehicle as a cluster that remains independent of the other vehicles. Therefore, it can be integrated with ad-hoc networking such as V2V communication.

A. Signaling Cost

The management of a mobility group must handle group disconnection events. Each time a person leaves the group (for instance - get out from the bus/train/subway), we have an event of temporary loss of network connection. Moreover, whenever the group leader is disconnected from his/her group (e.g., because of moving to another direction), all the group members are disconnected from the network. Let α_i, α' denote the probability of the group leaving by the MN client i and the CH, respectively. Then, given that there are N MNs managed by the CH (including the CH itself), comparing HPC with existing mobility management schemes that are not NEMO-based, the CH behaves the same, while the $(N-1)$ MNs behave differently. Therefore, the condition under which the wireless signaling cost associated with HPC mobility management is less than the equivalent cost associated with existing distributed mobility management schemes (e.g., DMM), i.e., the extra signaling caused by frequent location update events for DMM is greater than the extra signaling caused by group leaving events for a NEMO-based scheme is given by:

$$\beta(N-1)C_l > \sum_{i=1}^{N-1} \alpha_i C_b + \alpha'(N-1)C_b. \tag{2}$$

Where β is the group location update probability, C_l is the wireless signaling cost of MN location update and C_b is the wireless signaling cost of network binding procedure. Note that since the CH can use DMM on behalf of its cluster members, the associated non-wireless cost is the same for both HPC and DMM.

For vehicle-installed CH, $\alpha' = 0$ when HPC is used. Therefore, subject to this condition, substitute $\alpha' = 0$ in (2), we get that the condition under which the wireless signaling

cost associated with HPC is less than the equivalent cost associated with DMM is given by:

$$\beta(N-1) > \frac{C_b}{C_l}\sum_{i=1}^{N-1}\alpha_i. \tag{3}$$

Assuming that $\forall i, \alpha_i = \alpha$, it follows from (3) that HPC outperforms DMM for wireless signaling cost when:

$$\frac{\beta}{\alpha} > \frac{C_b}{C_l}. \tag{4}$$

Independently of the size of the mobility cluster N ($N \geq 2$). Using (4), the wireless cost reduction caused by using HPC instead of DMM for vehicle-installed CH (assuming that the CH uses DMM, and that $\forall i, \alpha_i = \alpha$), is given by:

$$wcost_{DMM} - wcost_{HPC} = (N-1)(\beta C_l - \alpha C_b). \tag{5}$$

Since NEMO-based schemes used a network element to select the group leader externally and arbitrarily, based only on its computational power, they are more vulnerable to the group leaving scenarios than HPC. For 5G networks, for which we have a short latency constraint, this situation may violate the requirement for short latency. Thus, NEMO-based schemes cannot support real-time applications. Whenever the rate of the group leaving events becomes sufficiently high, it is not clear if the signaling cost-saving justifies the extra signaling associated with frequent group leaving events, and the need for network binding procedures that follow these events. The situation in which the CH leaves its group is more likely to happen for the previous group-based studies than for HPC, for which the clusters are formed by the users based on proximity. Each HPC cluster is carefully formed for the MNs that are expected to remain within their CH proximity. Thus, group leaving events are expected to occur relatively rarely.

In contrast to previous group mobility schemes, HPC does not force the MN to initiate the network binding process. Thus, the wireless signaling cost associated with mobility is eliminated from the MN, and the CH takes the burden of binding the MN to the network. The CH is responsible to route the information to and from its cluster members using IP-in-IP encapsulation, as described in detail in NEMO proposal [19]. Since the client MN must be located within the vicinity of its CH, the information and signaling exchange between the CH and its client MN is transmitted over a short range (typically, less than a few meters), using short-range communication. Another cost associated with HPC is the delay incurred by the need to route the messages to/from the MN via the CH. Using an anchor point as a gateway to another device is not always recommended since this usage forces a triangular routing to this device. However, since the CH must reside nearby its MNs clients, the cost associated with a triangular routing is negligible.

It should be noted that the CH does not increase its mobility-associated signaling. Each MN has a digital representative in its CH which represents it. Therefore, any corresponding node needs only to interact with this digital representative.

B. Privacy and Security

Previous group-mobility schemes enabled unauthorized access to the MNs via their group leader. This mechanism requires privacy and security consideration. As opposed to this approach, HPC enables a PSK-based authentication process for accessing the MNs. The usage of a smartphone to protect its nearby IoT devices from hostile attacks was described in [30].

The issue of privacy and security was not addressed in any of the previous group mobility schemes. The possibility of hacking to many independent MNs using their "group leader" (which is used also as their mobile router) should be carefully considered. By selecting one individual as a mobile router to all the other members in the mobility group, we give this individual unauthorized access to the MNs which belong to another person. In addition, we increase the load on this individual. It is not clear if this selected group leader would accept this duty. In contrast to these schemes, since the HPC server-client connection is established only after performing a proximity-based authentication process, initiated by the human user, the MNs are more protected against a middle-man attack. To defend the MNs from such an attack, a vehicle-installed CH should be defined as a "mini" BS, such that it will be difficult for a smartphone to pretend as such. Due to paper length limitation, the implementation of this defense mechanism is not considered here.

C. Packet Loss Ratio

Another aspect of mobility management cost is the packet loss ratio. HPC requires the MN to transmit and receive packets via its CH. Since HPC is a proximity-based mobility management scheme, the MN-CH distance must be relatively short. For most practical cases, this distance should be no more than a few meters. Therefore, the packet loss ratio expected from HPC should be reduced, in comparison with existing mobility management schemes, in which the proxy is a network element. The reason for this observation is as follows. Since for short-range point-to-point communication along a line of sight (which is the case for MN-CH communication), the packet loss ratio depends mainly on the signal-to-noise-ratio, which depends (among other things) on the distance between the transmitter and the receiver, the packet loss ratio expected for HPC must be reduced. This is in comparison with existing proxy-based mobility management schemes (e.g., DMM), in which the proxy is a network element, whose distance from the MN is significantly larger than the distance from the MN to its CH. Indeed, the CH must still communicate with the network. However, since the CH should be a much more powerful device than its clients (in terms of radio signal strength), the expected packet loss ratio should be significantly reduced, in comparison with the alternative in which the MN communicates directly with the network.

NEMO [19] based schemes cannot guaranty either a short distance, nor a line of sight, between the group leader (which is the equivalent to the CH) and the MNs it manages. The group leader defined in NEMO [19] proposal is selected by a network element (previous NEMO-based schemes have chosen

different solutions for this network element) based only on its computational capability.

V. Performance Comparison

In this section, the performance of HPC is analyzed and compared with the performance of DMM [16], [17], and GMM [25], which at the time of writing this paper, is the most cited group-mobility scheme for cellular networks. Since no mobility client is requested by DMM on the MN, the CH can use DMM. Therefore, HPC can be integrated with DMM. For this scenario, the performance comparison of HPC with DMM is straightforward: Given that there are N members in the mobility group (including the CH), it follows from (5) that for vehicle-installed CH, the handover rate should be reduced by N times by using HPC, if the rate of handover requests initiated by the CH remains the same. In reality, the performance improvement expected from HPC should be even larger than N times, since as explained above, by roaming between large macro cells, handled by AANs, (for instance, by using a dual-mode as suggested in [29]), the handover rate initiated by the CH should be significantly reduced, in comparison with DMM. Consequently, the load on the PRACH should be significantly reduced by integrating HPC with DMM.

The major performance metric in this section is the handover rate associated with the rate of changing the MN location. To analyze the rate of changes in the MN location, a random walk model is used. An undirected graph $G = (V, E)$ is used to model the network topology. It is assumed that the CH and all the members in its mobility group are using PMIPv6 for mobility support. The MAG nodes are represented by the vertices V. An edge $e \in E$ represents a connection between two neighboring MAGs. The service area of each MAG may contain several BSs. An MN associated with a MAG i can move to any MAG j connected to i (i.e., there exists an edge in G connecting the vertices i and j) with a probability $p(i, j)$. The probability to move from the current MAG i to any other MAG is given by:

$$p_i = \sum_{j \neq i} p(i, j). \tag{6}$$

Given a uniform probability to move from any MAG i to another MAG, we get that p_i is independent of i. Therefore, from now on p_i is denoted by p. Thus, given that i and j are neighbors in G, the probability $p(i, j)$ to move from MAG i to MAG j during a time slot is given by:

$$p(i, j) = \frac{p_i}{N_i} = \frac{p}{N_i}. \tag{7}$$

Where N_i is the number of the nearest neighbors of i. That is, the group of all nodes in G, such that there exists an edge in G that connects i to each one of them. If j is not a nearest neighbor of i then $p(i, j) = 0$. The transition probability matrix P represents the transition probabilities $p(i, j)$ for all $i, j \in V$. There exists a unique vector Π which describes the steady state location probability distribution $\Pi = (\pi_1, \pi_2, ..)$. Each element $\pi_i \in \Pi$ describes the steady state probability to reside in location i. It is shown in [32] that the vector Π is obtained by solving the equation:

$$\Pi = \Pi P. \tag{8}$$

It follows from (8) that the steady state location probability vector Π depends on both the network topology as well as on the user mobility. Thus, no general closed form expression can be obtained in general for mobility cost analysis.

Let us consider an MN during a session. Given that the probability p to switch MAG during a time slot is constant for all the MNs and all the locations, then using PMIPv6, the rate S of handover events during t time slots is given by:

$$S_{PMIPv6} = tp. \tag{9}$$

As long as the MN remains in the same domain. Group mobility schemes, such as GMM, rely on mobility patterns similarity in order to create the mobility groups. That implies that the initial network binding procedure must be performed by each MN, independently of other MNs. Thus, for PMIPv6-based group mobility schemes, such as GMM, during a session, each MN must initiate p handover requests per time slot as long as the LMA remains the same until the mobility group is created. On the other hand, using HPC, from the very first beginning the rate of p handover requests per time slot holds only for the CH. The information of all the MNs handled by the CH is encapsulated in the tables handled by the CH. The HA of each MN forwards the messages directed to the MN to its associated CH.

As follows from the above analysis, in general, no closed term expression can be obtained for mobility cost analysis. The mobility cost depends on both the network topology and the MN mobility pattern. Therefore, from now on we consider a random walk model in a metropolitan area. The system consists of an infinite two-dimensional grid topology. Each BS has exactly 4 nearest neighbors to which the MN can move with an equal probability. That is, the MN can move either right, left, up, or down to another BS. The distance is computed using the Manhattan metric, which is commonly used for a metropolitan area. The distance is measured in terms of the number of BSs traveled by the MN. That is, given that the MN has traveled a distance d_x (in terms of BSs) along the x axis in one direction and a distance d_y along the y axis, then the distance d is computed as $d = |d_x| + |d_y|$. It is assumed that during a time slot the MN can move from its associated BS to at most one of its nearest neighbors, or remains at its location. The network binding strategy is such that upon traveling a pre-defined distance D, in terms of the number of BSs, from its current MAG, the MN must bind itself to a new MAG, and therefore, must perform a network binding process. It is further assumed that the mobility pattern is local within the same domain, with no LMA switch. Given the probability p to switch BS, the probability to make m movements from one BS to another during t time slots is given by:

$$\mu(m, t) = \binom{t}{m} p^m (1 - p)^{t-m}, \tag{10}$$

if $m \leq t$, and 0 if $m > t$. Let us consider first a one dimensional motion. Given that the MN has made m_x movements

from one BS to another along one axis, say x, the probability to travel a distance of d BSs along this direction, either right, left, up, or down, is given by:

$$\rho_{1D}(d, m_x) = \binom{m_x}{\frac{m_x-d}{2}} p_{1D}^{\frac{m_x+d}{2}} p_{1D}^{\frac{m_x-d}{2}}, \tag{11}$$

if $m_x \geq d \geq 1$ and $m_x - d$ is even, and 0 otherwise. The explanation for (11) is that to travel a distance d, in terms of the number of BSs, along one direction, the MN must travel along this direction a distance that is greater by exactly d than the distance along the opposite direction. That is, given that the MN has made m_x movements along the x axis, $(m_x + d)/2$ movements were made along this direction, while $(m_x - d)/2$ movements were made along the opposite direction. Since the probability to move to any of the four possible directions is the same, we get that:

$$p_{1D} = \frac{1}{4}p. \tag{12}$$

Substitute (12) in (11), we get that:

$$\rho_{1D}(d, m_x) = 2^{-2m_x} \binom{m_x}{\frac{m_x-d}{2}} p^{m_x}, \tag{13}$$

if $m_x \geq d \geq 1$ and $m_x - d$ is even, and 0 otherwise. Thus, for the one-dimensional motion case, given that the MN has made m_x movements only along one axis, either x or y, the probability to travel a distance d along this axis, for any of the four possible directions, is given by:

$$\rho_{any1D}(d, m_x) = 4\rho_{1D}(d, m_x) = 2^{-2(m_x-1)} \binom{m_x}{\frac{m_x-d}{2}} p^{m_x}, \tag{14}$$

if $m_x \geq d \geq 1$ and $m_x - d$ is even, and 0 otherwise. The probability to travel a distance d is the sum over all the probabilities to travel a distance d_x along the x axis and a distance d_y along the y axis, such that $|d_x| + |d_y| = d$.

Since both d_x and $d_y = d - d_x$ can be along either one among two opposite directions, the probability to travel a distance d, given that the MN has made m movements, from which m_x movements along the x axis, and given that the MN has traveled a distance d_x along the x axis, is given by:

$$\theta(d, m, d_x, m_x) = 4 \binom{m_x}{\frac{m_x-d_x}{2}} \binom{m - m_x}{\frac{(m-m_x)-(d-d_x)}{2}}. \tag{15}$$

For $m_x - d_x$ even, $m - d$ even, and $d > d_x > 0$. On the other hand, if d_x equals either zero or d, we have only two possibilities that satisfy the condition $|d_x| + |d_y| = d$. In this case we get:

$$\theta(d, m, d_x, m_x) = 2 \binom{m_x}{\frac{m_x-d_x}{2}} \binom{m - m_x}{\frac{(m-m_x)-(d-d_x)}{2}}. \tag{16}$$

For any other case $\theta(d, m, d_x, m_x) = 0$. The probability to travel a distance d, given that the MN has made m movements, is the sum over all possible values of m_x that satisfy the constraints $m \geq d \geq d_x$ and $m_x \geq d_x$:

$$\Theta(d, m) = \sum_{d_x=0}^{d} \sum_{m_x=d_x}^{m} \theta(d, m, d_x, m_x). \tag{17}$$

Using again the constraints $m \geq m_x \geq d_x$ and $d \geq d_x$, and $m - m_x = m_y \geq d_y = d - d_x$ we get that:

$$d_x = d - d_y \geq d - m_y = d - (m - m_x). \tag{18}$$

Hence, it follows from (17) and (18) that:

$$\Theta(d, m) = \frac{1}{4^m} \sum_{m_x=0}^{m} \binom{m}{m_x} \sum_{d_x=max\{0,d-(m-m_x)\}}^{min\{m_x,d\}} \theta(d, m, d_x, m_x). \tag{19}$$

The probability to travel a distance d during exactly t time slots is the sum over all probabilities $\Theta(d, m)\mu(m, t)$, where m ranges from d to t:

$$\eta(d, t) = \sum_{m=d}^{t} \Theta(d, m)\mu(m, t). \tag{20}$$

Hence, the expected number of location update events during T time slots is given by

$$\xi(D, T) = \sum_{d=D}^{T} \eta(d, T) \lfloor \frac{d}{D} \rfloor. \tag{21}$$

The explanation of (21) is as follows: Given that the MN has traveled a distance d during T time slots, the number of location update events during this time period is

$$\nu(d, D) = \lfloor \frac{d}{D} \rfloor. \tag{22}$$

Using the fact that $\eta(d, t) = 0$ if $d > t$, we get that the expected number of location update events during time interval of T time slots is the sum over all probabilities to travel a distance d during T time slots, multiplied by $\lfloor d/D \rfloor$, where d ranges from $d = D$ to $d = T$. Substitute (20) in (21) we get that

$$\xi(D, T) = \sum_{d=D}^{T} \sum_{m=d}^{T} \Theta(d, m)\mu(m, T) \lfloor \frac{d}{D} \rfloor, \tag{23}$$

if $D \leq T$, and 0 if $D > T$.

VI. Simulation Results

In this section numerical experiments were used for comparing the HPC performance with that of DMM [16], [17] and GMM [25]. The performance metric considered in the simulation is the rate of handover requests. This parameter is critical for high mobility communication since the time duration during which the network must allocate the required bandwidth decreases with the MN mobility. Note that due to bandwidth limitation, upon a handover event during a video session, GMM must handle each MN independently of the other group members. The reason for this behavior is that since GMM [25] selects one MN to serve its peers as a router, there is no guaranty that the group leader can provide the bandwidth required by the group members. For instance, a smartphone selected by GMM [25] as a group leader cannot support video sessions for many smartphones in its mobility group. Therefore, whenever there is a need to establish a video session for more than one group member, the required

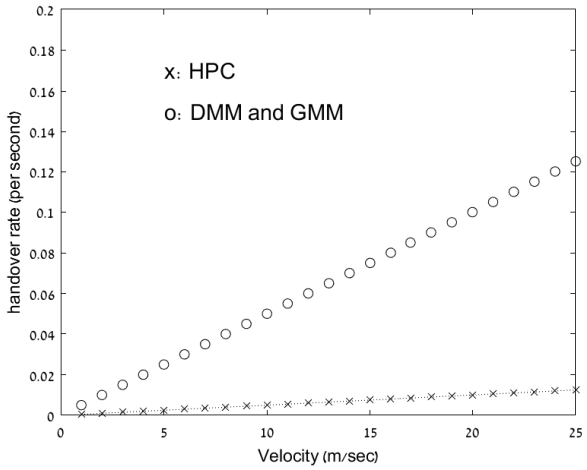

Fig. 2. The handover rate per second, as a function of the device velocity, for DMM and GMM versus HPC, for a directional motion.

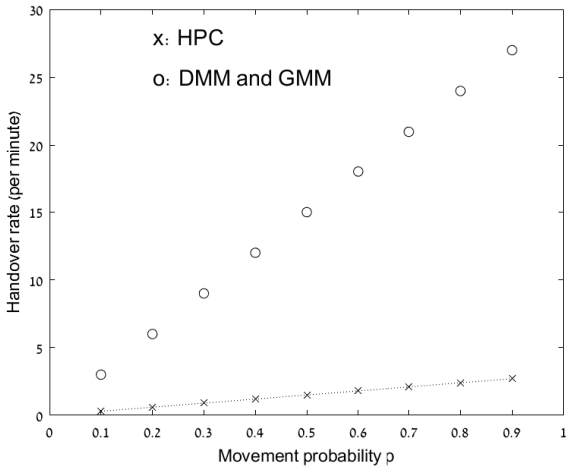

Fig. 3. The rate of handover events per minute, as a function of the movement probability p during the session, for DMM and GMM versus HPC, for a random walk motion.

bandwidth must be allocated to each MN as an individual. Hence, the rate of handover events that must be handled, expected for real-time video applications is the same for DMM and GMM. In contrast to these schemes, HPC uses a vehicle-installed dedicated device, the CH, which can serve its cluster members as a "mini" mobile BS. The CH is capable to support its cluster members and provides the bandwidth they need, up to a pre-defined value, which depends on the cluster (e.g., a bus, train $\cdots$). The network has to deal only with the CH, and the aggregated bandwidth consumed by its cluster. Two mobility patterns are considered: A directional motion and a random walk motion.

Fig. 2 depicts the handover rate of a CH moving in a crowded city, that communicates as a nomadic cell with AANs, in comparison with the handover rate of a single device that uses either GMM or DMM, that communicates with terrestrial BSs. The handover rate is depicted as a function of the CH velocity. The system under consideration is based on the two-dimensional infinite grid topology described in Section V, where each tile has sizes of 200 m $\times$ 200 m for a terrestrial BS, and 2 km X 2 km for the cells handled by an AAN. A directional motion model is used to describe the motion of the CH/MN (consider, for instance, a tram in a metropolitan area). The CH/MN velocity ranges from 1 m per second (3.6 km per hour, a pedestrian) to 25 m per second (90 km per hour, a vehicle). The superiority of HPC over both DMM and GMM is very significant. It should be noted that since we consider a single device the handover rate of DMM and GMM is the same.

Fig. 3 depicts the handover rate of a cluster moving in a crowded city. The system under consideration is based on the two-dimensional infinite grid topology described in Section V, where each tile has sizes of 200 m $\times$ 200 m. A random walk model is used to describe the motion of a bus in a metropolitan area. Therefore, as in the analysis described in Section V, the metric used is the Manhattan metric, which is commonly used for a metropolitan area. The bus average velocity is 36 km per hour, and at any given moment there are 10 active video sessions used by the passengers in the bus. It is assumed that the CH installed on the bus is capable to support up to 20 video sessions simultaneously. Using the analysis in Section V, Fig. 3 depicts the rate of handover events per minute, as a function of the movement probability p, defined in Section V, for DMM and GMM versus HPC. Even for the random walk model, for which the expected rate of handover events is relatively low, a significant superiority of HPC over DMM and GMM is demonstrated. While DMM and GMM must handle 10 handover events simultaneously for each and every cell switch, the CH alone handles handover events, yet with the aggregated bandwidth for 10 users.

The second mobility pattern to be considered is directional motion. We consider a train moving at a velocity of 360 km per hour on an infinite one-dimensional system. BSs are located at every 3 km intervals along the track. Location information of the train is used for handover between BSs, as described in [33], and reported recently in [34]. Fig. 4 depicts the rate of handover events per minute, as a function of the number of the active video sessions on the train, for DMM and GMM versus HPC. While both DMM and GMM need to handle each MN separately, for HPC it is sufficient to allocate the accumulated bandwidth, as reported by the CH, in advance. Consequently, as it is shown in Fig. 4, the rate of handover events can be significantly reduced.

To compare the load on the PRACH for GMM versus HPC, we examine the number of network binding events for both methods. Note that as explained in detail in Section V, the load on the PRACH expected for HPC should be significantly reduced, in comparison with DMM. We consider a bus in an urban area, carrying 20 passengers. We consider a random walk motion in a two-dimensional grid topology system, as described above for Fig. 3. It is assumed that the service zone of each BS is the tile in which this BS is located and that upon moving to another tile, a network binding process must

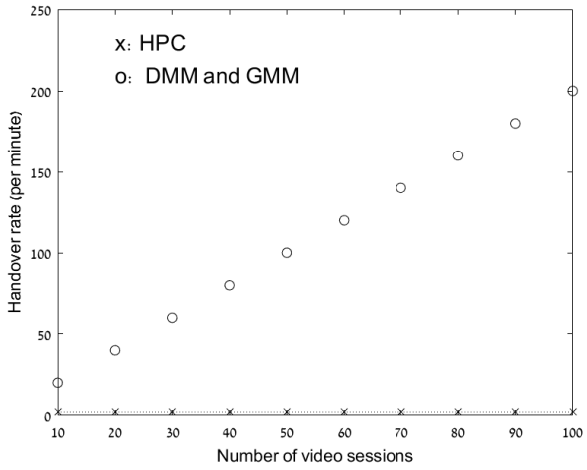

Fig. 4. The rate of handover events per minute, as a function of the number of video sessions in the cluster, for DMM and GMM versus HPC, for a directional motion.

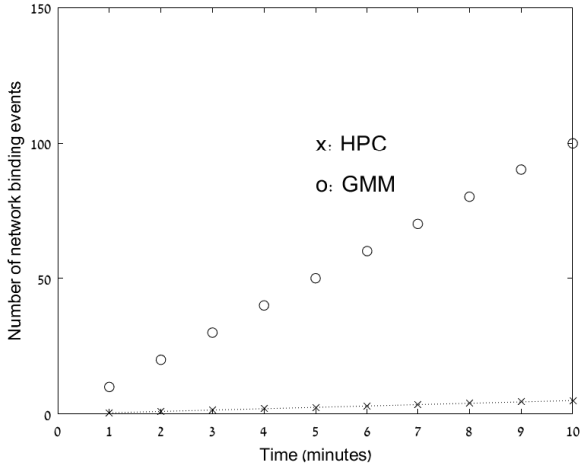

Fig. 5. The number of network binding events, as a function of the time duration of creating the GMM mobility group, for GMM versus HPC, for a random walk motion.

be initiated. That is, each BS is used as a MAG. Therefore, applying the analysis in Section V, the condition under which a location update event must be initiated is $D = 1$. Fig. 5 depicts the number of network binding events as a function of the time duration (measured in minutes) required to create the GMM mobility group, for GMM versus HPC, during the initial network binding procedure. While HPC exploits a bus-installed device as a nomadic cell, GMM must handle each passenger individually, until the mobility group is created. It is demonstrated that the number of network binding events needed to be handled is significantly larger for GMM. Thus, the load on the PRACH is significantly larger for GMM, in comparison with HPC.

TABLE I
LIST OF ACRONYMS.

Acronym	Description
5G	Fifth-generation
AAN	Aerial access networks
QoS	Quality of service
UAV	Unnamed aerial vehicle
MIPv4	Mobile IP version 4
MIPv6	Mobile IP version 6
PMIPv6	Proxy mobile IP version 6
DMM	Distributed mobility management
SDN	Software defined network
NEMO	Network mobility
GMM	Group mobility management
MN	Mobile node
BS	Base station
PRACH	Physical random access channel
RAN	Radio access network
HA	Home agent
CH	Cluster head
LMA	Local mobility anchor
MAG	Mobility access gateway
HPC	Hierarchical proximity-based consolidation

VII. SUMMARY AND CONCLUDING REMARKS

This study suggests reducing the burden caused by the mobility of a massive number of highly mobile devices by partitioning the devices into mobility clusters, such that the network has to handle only one representative for each cluster. Using AANs to support highly mobile devices, each cluster moves between large macro cells that are significantly larger than typical cells expected for beyond 5G networks. Consequently, the rate of handover requests can be significantly reduced. The major difference between HPC and previous group mobility schemes is that HPC is a scalable user-based distributed scheme, formed hierarchically by two-levels proximity-based clusters, while previous group mobility schemes are centralized network-based schemes, based on consolidating the users' mobility patterns, that are not scalable, and cannot support real-time applications.

REFERENCES

[1] M. R. Palattella *et al.,* "Internet of things in the 5G era: Enablers, architecture, and business models," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510–527, 2016.
[2] "5G latency - Reality checks," SENKI. Dec. 2018. Retrieved Oct. 2019.
[3] 5G Americas, "New services and applications with 5G ultra-reliable low latency communications," [Online]. Available: https://www.5gamericas.org/new-services-applications-with-5g-ultra-reliable-low-latency-communications/
[4] Sabine Dahmen-Lhuissier. "ETSI - Mobile," ETSI.
[5] "Customers in Chicago and Minneapolis are first in the world to get 5G-enabled smartphones connected to a 5G network," verizon.com. Apr. 2019. Retrieved May 2019.
[6] G. A. Akpakwu , B. J. Silva, G. P. Hancke, and M. A. Abu-Mahfouz, "A survey on 5G networks for the Internet of things: Communication technologies and challenges," *IEEE ACCESS*, vol. 6, pp. 3619–3647, Dec. 2017.
[7] C. Perkins, "IP mobility support for IPv4 (2002)," [Online]. Available: http://www/rfc-editor.org/info/rfc3344
[8] C. Perkins, IP mobility support for IPv4, revised (2010). http://www/rfc-editor.org/info/rfc5944.
[9] V. Devarapalli, K. Chowdhury, S. Gundavelli, B. Patil, and K. Leung, "Proxy mobile IPv6," IETF, RFC 5213 (2008). RFC 5213, DOI 10.17487/RFC5213, http://www.rfc-editor.org/info/rfc5213.

[10] H. Yokota, K. Chowdhry, R. Koodi, B. Patil, and F. Xia, "Fast handover for PMIPv6," IETF 2010, RFC 5949, DOI 10.17487/RFGC5949. http://www.rfc-editor.org/info/rfc5949.
[11] M-C. Chuang and J-F. Lee, "FH-PMIPv6: A fast handoff scheme in proxy-mobile IPv6 networks," in *Proc. CECNet*, 2011.
[12] F. Giust, C. J. Bernardos, and A. de LA Oliva, "Analytic evaluation and experimental validation of network-based IPv6 distributed mobility management solution," *IEEE Trans. Mobile Comput.*, vol. 13, no. 11, pp. 2484–2497, 2014.
[13] J. H. Lee, J. M. Bonnin, I. You, and T. M. Chung, "Comparative handover performance analysis of IPv6 mobility management protocols," *IEEE Trans. Ind. Electron.,* vol.60, no. 3, pp. 1077–1088, 2013.
[14] Y. Li *et al.,* "Beyond 5G: Reliable extreme mobility management," in *Proc. ACM SIGCOMM*, 2020.
[15] A. Bakshi, Y. Mao, K. Srinivasan, and S. Parthasarathy, "Fast and efficient cross band channel prediction using machine learning," in *Proc. ACM MobiCom*, 2019.
[16] H. Chan *et al.,* "Requirement for distributed mobility management," IETF RFC 7333,2014.
[17] F. Giust, L. Cominardi, and C. J. Bernardos, "Distributed Mobility Management for future 5G networks: Overview and analysis of existing approaches," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 142–149, 2015.
[18] T. Nguyen, C. Bonnet, and J. Harri, "SDN-based distributed mobility management for 5G networks," in *Proc. IEEE WCNC*, 2016.
[19] V. Devarapalli, R. Wakikawa, A. Petrescu, and P. Thubert, "Network Mobility (NEMO) basic support protocol," IETF RFC 3963, 2005.
[20] J. Guan, I. You, C. Xu, and H. Zhang, "The PMIPv6-based group binding update for IoT devices," *Hindawi J. Mobile Inf. Syst.*, 2016.
[21] S. Jeon S and Y. Kim, "Cost-efficient network mobility scheme over proxy mobile IPv6 network," *IET Commun.,* vol. 5, no. 18, pp. 2656–2661, 2011.
[22] M. S. Kim and S. Lee, "Group-based fast handover for PMIPv6-based network mobility in vehicular networks," in *Proc. IEEE INFOCOM WORKSHOPS,* 2015.
[23] J. H. Lee, T. Ernst, and N. Chilamkurti, "Performance analysis of PMIPv6-based network mobility for intelligent transportation systems," *IEEE Trans. Veh. Technol.*, vol. 61, no. 1, pp. 74–85, 2012.
[24] I. Soto, C. J. Bernardos, M. Calderon, A. Banchs, and A. Azcorra, "Nemo-enabled localized mobility support for Internet access in automative scenarios," *IEEE Commun. Mag.*, vol. 47, no. 5, pp. 152–159, 2009.
[25] H. L. Fu, P. Lin, H. Yue, G. M. Huang, and C. P. Lee, "Group mobility management for large-scale machine-to-machine mobile networking," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1296–1305, 2014.
[26] Z. Naor, "Efficient mobility support services for highly mobile devices in 5G networks," in *Proc. IEEE SMDS*, 2021.
[27] System architecture for the 5G system, ETSI 3GPP TS 23.501 version 15.2.0 Release 15, 2018-06.
[28] 3GPP TR 23.793 V16.0.0, Technical Report 3rd generation partnership project, technical specification group service and system aspects; study on access traffic steering, switch and splitting support in the 5G system architecture, Release 16, 2018-12. 2014.
[29] O. Semiari, W. Saad, M. Bennis, and B. Maham, "Caching meets millimeter wave communications for enhanced mobility management in 5G networks," *IEEE Trans. Wireless Comm.,* vol. 17, no. 9, 2018.
[30] J. Zhang, Z. Wang, Z. Yang, and Q. Zhang, "Proximity based IoT device authentication," in *Proc. IEEE INFOCOM*, 2017.
[31] M. R. Garey and D. S. Johnson, "Computers and intractability: A guide to the theory of NP-completeness," New York: W. H. Freeman, 1979.
[32] L. Kleinrock, Queueing Systems, Vol. 1,2, Wiley, 1976.
[33] Report ITU-R M.2395-0, "Introduction to railway communication systems," Nov. 2016.
[34] [Online]. Available: https://www.everythingrf.com/news/details/7622-NEC-5G-Base-Stations-Used_Transmit-HD-Video-to-a-Moving-Train Reported: February 8, 2019.
[35] 3GPP TS 29.274, "Evolved general packet radio service (GPRS) tunneling protocol for control plane (GTPv2-C)," Sept. 2011.
[36] A. Orsino *et al.,* "Effects of heterogeneous mobility on D2D- and drone assisted mission-critical MTC in 5G," *IEEE Commum. Mag.*, vol. 55 no. 2, pp. 79–87, 2017.
[37] [Online]. Available: https://www.gartner.com/id/3165317

Zohar Naor received the Ph.D. degree in Computer Science from Tel Aviv University, Tel Aviv, Israel, in 2000. Since 2003 he is with the University of Haifa, Israel. His areas of interests include video streaming, P2P networks, wireless networks, resource management of computer networks, mobility management, search strategies, and multiple access protocols.