A Distributed Self-Organization Approach to Minimize the Signaling and Delay Caused by Mobility Management Function in Cellular Networks

Merim Dzaferagic, Nicola Marchetti, Irene Macaluso, CONNECT, Trinity College Dublin, Ireland

I. INTRODUCTION

To manage mobility, RAN nodes in both 4G and 5G are grouped into a hierarchy of geographical areas. We demonstrate a 4G/5G compliant Network Level Mobility Management Optimization solution based on User Equipment (UE) Mobility to minimize signaling (i.e., handover signaling, paging and tracking area updates) and handover latency by dynamically reconfiguring the association between nodes in the Radio Access Network (RAN) and nodes (e.g. Mobility Management Entity), functions (e.g. Access and Mobility Management Function) and Location Regions (e.g. Tracking Area, Registration Area, Tracking Area List) in the core network.

Due to the increasing densification of the cellular networks and connected mobile devices (e.g. IoT devices for logistics and supply chain management), the number of handovers will significantly grow. Services like Augmented Reality, Virtual Reality, autonomous driving and smart buildings/cities are steering the development of the next generation of communication networks. These services are an important part of the Fifth Generation (5G) concept, imposing strict requirements in terms of latency, throughput, and mobility. Some of these requirements are conflicting, due to the network architecture and protocol procedures. For example, handover procedures are inevitable to ensure mobility, but at the same time they greatly affect latency. Different types of handovers (e.g., inter-Mobility Management Entity (MME), intra-MME) require distinct procedures that affect the latency differently. We designed a distributed and adaptive solution to reorganize nodes in the network so that the number of handovers requiring an MME/AMF re-allocation (inter-region handover) - the most inefficient type of handover in terms of delay and number of exchanged messages - is minimized. In fact, inter-region handover procedures require more parameter updates in the core network compared to intra-region handovers, which results in increased latency. According to [1], a handover with MME re-allocation requires on average 50% more signaling messages compared to other types. Hence the proposed solution significantly reduces the amount of signaling in the core network. The nodes reconfiguration is based on user movements and relies on existing protocol messages in the network [2], [3]. The mechanism also balances the load between the MME/AMF instances. Moreover, the designed solution is independent of the specific handover mechanisms implemented in the network (e.g. whether the handover is network-controlled or mobile-valuated, or one uses cell range expansion techniques, etc.). While the proposed mechanism can be applied to Tracking Area and Tracking Area List optimization, we will use the association between nodes in the RAN and the MME/Access and Mobility Function (AMF) as the main example in this text. Our solution successfully adjusts to both, the up and downscaling of resources in the core network (e.g. new Access and Mobility Function (AMF) instance or shutting down an AMF instance) and the changes in the RAN by reconfiguring the nodes accordingly. The solution is an online learning approach, which means that it is dynamic and adapts to the changes in the network. It takes advantage of the existing signaling messages, meaning that no additional signaling is introduced in the network, which reduces the probability of an escalation of the signaling traffic (i.e., signaling storm). The proposed solution is also architecture agnostic in the sense that we verify it with the 4G and 5G architecture, meaning that it can be implemented with the current Evolved Packet Core (EPC) as well as with the Software Defined Networking (SDN) and Network Function Virtualization (NFV) based Fifth Generation (5G) Service Based Architecture (SBA). Our solution is aligned with the recent trends in network optimization in that it leverages the power of self-organization and network configuration automation. Additionally, according to the recent 3GPP documents [4], [5] one of the key issues is the mobility management optimization based on the UE mobility pattern recognition, which aligns with the main objectives of our solution.

Self-organizing and autonomic network management and optimization will play a key role in 5G networks in that they will be essential to fully exploit in a cost efficient way the increased network flexibility and the dynamic on-demand deployment of virtual network functions in the NGC [6], [7]. A solution for autonomic distributed network-level optimization to reduce the handover latency can support the requirements of new services in 5G and the increased flexibility of the 5G architecture.

In the literature, a number of approaches have been proposed for autonomic network optimization. In [8] a mechanism to minimize the number of Serving Gateway (SGW) relocations is proposed. While MME relocation is also mentioned, the authors focus on the SGW. They propose the introduction
of a Service Area for idle users, which is a subset of the Service Area of each SGW. Although the results show a decrease of the number of SGW relocations for users in active mode, the mechanism to determine the idle-mode service areas is not provided and the implications in terms of existing standards are not discussed. In [9] a centralized heuristic mechanism to deploy network functions so as to minimize the number of SGW relocations is proposed. The minimization of MME relocations is only mentioned as a possible application.

However, the centralized nature of the proposed solution would require the introduction of new protocol messages and an increase in the signaling overhead. In contrast, the solution proposed by us is designed to exploit standardized protocols and mechanisms. This will facilitate the adoption of our solution. Moreover, since the proposed solution relies on a distributed approach based on local information, it is able to scale with the size of the infrastructure. Another approach in literature for inter-region handover optimization relies on a new architecture, proposed in [10]. The authors propose a recursive hierarchical algorithm to minimize the number of inter-region handovers in [11]. In contrast, the solution we propose is designed to work within existing (4G) and emerging (5G) mobile networks so as to maximize technology adoption. Specific solutions for handover latency reduction have been proposed in the case of femto cells [12], [13]. These two patents detail an optimized intra-HeNB GW (Home eNodeB Gateway) handover mechanism that reduces signaling to and from an MME. While our approach also aims at reducing the signaling to and from the MME, our network-wide mechanism is not restricted to the specific femto cell case and we take into account users mobility when optimizing the nodes configuration.

In [14], [15] the inventors detail a mechanism to re-assign eNodeBs to MMEs so as to balance the load of the MMEs. Our solution, while also taking into account the load, reconfigures the nodes based on the users movements.

Our solution is a generic approach to node to x-area (where x can be handover, tracking area or set of Tracking Areas) association. Therefore, closely related to the question addressed by our solution is the autonomous configuration of Tracking Areas (TAs) and TA Lists. Several solutions aiming at minimizing the signaling overhead caused by the periodic TA updates and paging have been developed in the research field [16] or patented [17], [18]. The approaches in [16]–[18] are centralized and therefore do not scale with the network size. Additionally, the authors do not provide a study of how to implement the proposed techniques in terms of data collection, and reconfiguration of the nodes. In contrast, we propose a distributed solution, that can be implemented within the existing 4G and the future 5G network architecture by leveraging the mobility information that are already been collected by the network. Our solution allows for multiple directions when it comes to implementation: (1) a completely distributed self-organizing network (SON) implementation: each node in the RAN runs the algorithm locally; (2) a decentralized SON: subsets of the network are being optimized by local centers; and (3) a centralized SON: the UE mobility data is collected for the whole network and the optimization is performed for the whole network. All three implementations leverage the power of the distributed decision making which results in high performance optimization (i.e. the complexity of the centralized and distributed solutions are O(N) and O(1) respectively compared for example to the complexity of the centralized approach in [16] which is O(LN) where L is the number of areas).

II. PROBLEM DESCRIPTION

The Policy control and Charging Rules Function (PCRF), Home Subscriber Server (HSS), MME, SGW and the Packet Gateway (PGW) are the main functional entities in the EPC architecture [19]. Figure 1 depicts these functional entities and it shows which interfaces are used for the communication between them. The PCRF encompasses policy control decision and flow based charging control functionalities. The HSS is the main database in the core network and it is responsible for the subscription management. The MME serves as the main control entity that handles all signaling events which are related to the core network (e.g., mobility management, paging, bearer setup, subscriber information). The SGW is the local mobility anchor point for the inter-eNodeB handover, and it is in charge for the forwarding and routing of user-data packets between the eNodeB and the PGW. The PGW is the interface between the EPC and other packet data networks. It is responsible for the policy enforcement, packet filtering, device IP address allocation etc.

The 5G core network is based on the SDN and NFV principles, which result in high flexibility. In order to support on-demand provisioning, a service based approach was the obvious way to go. Network functions, that were implemented as hardware nodes in the EPC (e.g. MME, HSS, SGW), are softwareized and decomposed into functional entities, which are accessible as services in the new 5G SBA. Figure 2 shows the mapping between the EPC nodes and the SBA functional entities. Some of the EPC nodes are mapped one-to-one into SBA functional entities (e.g., the PCRF is mapped into the Policy Control Function (PCF)), whereas other nodes have a one-to-many mapping (e.g., the MME is mapped into the AMF and Session Management Function (SMF)). Since we focus on the network mobility management, our node of interest in the EPC is the MME and in the SBA the function of interest is the AMF.

Two types of handover approaches exist, namely network controlled (the network forces the UE to move from one BS to
Intra-MME/SGW: this type of handover occurs when a UE moves between two eNodeBs that belong to the same MME/SGW pool. If an X2 interface exists between these two eNodeBs the handover is completed without EPC involvement, and we refer to this type of handover as X2-handover. If an X2 interface does not exist between the serving and target eNodeB, the EPC has to be involved in the handover, and since this signaling is carried out via S1 interfaces, we refer to it as S1-handover.

Inter-MME/SGW: this type of handover occurs when a UE moves between two eNodeBs that belong to different MME pools or SGW service areas. In order to perform this type of handover the involvement of the EPC is necessary and therefore we refer to this type of handover as S1-handover.

Inter-RAT: this type of handover occurs when a UE moves between two different radio technologies (e.g., a handover from LTE to WCDMA).

An MME Pool Area and an SGW service area are defined as areas within which a UE may be served without the need to change the serving MME and SGW respectively [1]. In this text, from now on, we will refer to the X2 handover and the S1 handover without MME re-selection as intra-region handover and to the S1 handover with MME re-selection as inter-region handover. Based on [4] the 5G network architecture includes similar concepts. The main difference is that the functionalities are virtualized, and the equivalent to the MME Pool Area is the AMF Region. An AMF region consists of one or multiple AMF Sets. An AMF Set consists of AMFs that serve a given geographical area [4]. The corresponding S1 and X2 interfaces are renamed and the interfaces of interest in the 5G architecture are the N2 and Xn interfaces respectively. Hence, the inter-region handovers are the N2 handovers with AMF re-selection and intra-region handovers are Xn and N2 handovers without AMF re-selection.

A Tracking Area is a logical collection of base stations (see Figure 3) used to perform tracking and reachability management functions to trace the geographical location of a UE in the idle state. The location of a UE in the idle state is known by the network on a Tracking Area List granularity [1]. An MME always manages whole Tracking Areas and Tracking Areas can also be managed by multiple MMEs. Traditionally, a node gets assigned to a Tracking Area based on its geographical location. Due to the static deployment of the nodes (the base stations are not moving), the Tracking Area assignment is static as well. The idea behind our algorithm is to dynamically rearrange the handover regions, based on the moving patterns of the UEs, in order to minimize the number of inter-region handovers. As a side effect the Tracking Areas are being rearranged as well. Similarly, according to [4] the 5G architecture relies on the concept of Tracking Areas and Tracking Area Lists. The AMF allocates registration areas, which represent a logical grouping of Tracking Areas within a network slice.

The inter-region handover procedures result in higher latency [20]. Our goal is to minimize the network latency and the amount of signaling between the RAN and the packet core by introducing a smart design of the handover regions. More precisely, we propose a distributed adaptive algorithm, that runs on the RAN nodes and the instances of MMEs/AMFs and takes advantage of local handover information available through the existing signaling messages, to form handover regions resulting in a minimum number of inter-region handovers.

III. DISTRIBUTED SELF-ORGANIZATION ALGORITHM

In this section, we present our proposed solution that allows us to optimize the handover regions in a way that minimizes the number of inter-region handovers. Additionally, we explain the benefits of a distributed approach compared to a centralized one.

Our solution consists of two main components. These two parts are supposed to be running on the RAN nodes and virtual instances of core nodes (e.g. MMEs/AMFs). The component that runs on the RAN nodes can be formalized with Algorithm
Algorithm 1 Cell resource allocation procedure

while Cell is operational do
    Wait for event \{handoff, reassign request\}
    if event == handoff then
        Update counters
        Calculate energy of attraction
        MakeAssignmentDecision()
    if event == reassign request then
        MakeReassignment(MME/AMF_id)

function MakeAssignmentDecision()
    M ← List of available MMEs/AMFs
    k ← number of MMEs/AMFs that are managing the cell
    A ← List of energies of attraction towards MMEs/AMFs
    Sort A in descending order
    Send assignment request to the first k MMEs/AMFs in A

function MakeReassignment(MME/AMF_id)
    A ← List of energies of attraction towards MMEs/AMFs
    Exclude the MME/AMF with MME/AMF_id from A
    Sort A in descending order
    while MME/AMF not assigned do
        Try to get assigned to an MME/AMF from A

1. As shown in Algorithm 1 when the node is turned on, it starts the initialization phase. There are multiple options when it comes to the initialization phase, for example: the node could be assigned to the MME/AMF that is in charge for its neighbors; the node could also use a random assignment to an available MME/AMF. After the completion of the initialization phase, the algorithm starts its optimization process. Since the solution relies on the handover counters available on the base stations, the optimization process is triggered whenever a handoff occurs or in case an MME/AMF sends a reassignment request.

In case of a handover, the base station updates its counters and the algorithm calculates the energy of attraction towards all available MMEs/AMFs. The energy of attraction of node \( n \) towards the \( m \)th MME/AMF is calculated as:

\[
A_m(n) = \sum_{i \in M} \frac{H_n(m)}{H_n(i)}
\]  

(1)

where \( H_n(m) \) is the number of handover requests that arrived at node \( n \) from nodes that are assigned to the \( m \)th MME/AMF. Therefore, the energy of attraction towards an MME/AMF is the ratio between the number of handover requests that came from this MME/AMF and the total number of handover requests that arrived on the observed node.

Once the counters are updated and the energy of attraction is calculated the base station, based on the attraction towards all available MMEs/AMFs, decides whether to stay assigned to the current MMEs/AMFs or to change its assignment.

In case an MME/AMF requested a reassignment of the base station to another MME/AMF, the base station starts a reassignment process. This implies sorting the list of available MMEs/AMFs based on the energy of attraction and excluding the MME/AMF that has sent the request for reassignment from this list. The next step is to get assigned to the next best (based on the energy of attraction) available MME/AMF. A flow chart describing Algorithm 1 is shown in Figure 4.

Algorithm 2 MME/AMF resource allocation procedure

\[ N \leftarrow \text{List of cells assigned to the MME/AMF} \]
\[ L \leftarrow \text{Current MME/AMF load} \]
\[ L_{\text{max}} \leftarrow \text{Load limit} \]
\[ A \leftarrow \text{List of energies of attraction of cells towards this MME/AMF} \]
Wait for assignment request from \( n \)
if \( L + L(n) < L_{\text{max}} \) then
    Assign cell to this MME/AMF
else
    if \( A(n) > \min(A) + \delta \) then
        Assign cell to this MME/AMF
        Inform the cell with \( \min(A) \) to assign to another MME/AMF
        Remove the cell with \( \min(A) \) from \( N \) and \( A \)
    else
        Reject the request

The second component runs on a virtual instance of the MME/AMF. This component can be formalized with Algorithm 2. As shown in Algorithm 2 the MME/AMF waits for a request from a node that wants to get assigned to it. If the sum of the current load of the MME/AMF and the load coming from the node that is requesting the assignment is lower than a threshold \( (L_{\text{max}}) \), the cell is going to be assigned to the MME/AMF. In case the total load is greater than the threshold,
the assignment can be accepted or rejected depending on the energy of attraction of the cell that is requesting the assignment. If the energy of attraction of the cell is lower than the attraction of all other cells that are currently attached to the MME/AMF of interest, the assignment request will be rejected. If the energy of attraction of the cell is greater than the attraction of any other cell that is currently attached to the MME/AMF of interest, the cell will be assigned to the MME/AMF and the cell with the lowest energy of attraction that was already assigned to the MME/AMF will be informed to change its assignment. Once the cell is reassigned it will be removed from the list of assigned cells on the current MME/AMF. A flow chart describing Algorithm 2 is shown in Figure 5.

An important fact, from the implementation point of view, is that all the information needed to run the algorithm is available through already existing signaling messages, removing the need for additional signaling and at the same time simplifying the integration with the existing architecture. The information of interest are the type of handover, the number of handovers and the source MME from which the handover originated from. To explain this let’s have a look at the procedures used to perform the two different types of handover (inter and intra-region). As shown in Figure 6, the intra-region handover procedure assumes direct communication between the involved nodes. The HandoverRequest message, among other information, contains a field that is reserved for the UE History Information, which contains information about the last visited cells by the UE [2]. This information is used to figure out where the handover is coming from. Considering that the request is sent through the X2 link between two nodes, the handover type is obviously intra-region and therefore the source MME is equal to the destination MME. This means that all counters can be updated appropriately (the total number of handovers is increased by 1, the number of intra-region handovers is increased by 1 as well).

Figure 7 shows the procedure that is used to perform a handover between nodes that are associated with different MMEs. The initial HandoverRequest message, contains a field that transfers the information from the source node to the target node (TargeteNB-ToSourceeNB-TransparentContainer), which contains the UE History Information similar to the previous case [3]. In case of an inter-region handover procedure, we have to figure out the source MME as well in order to update all counters needed for the energy of attraction (equation (1)). Considering that an MME covers whole tracking areas, the inter-region handover includes a Tracking Area Update (TAU) as well (the last part of the handover procedure shown in Figure 7). As shown in Figure 8, the first message that is sent from the UE to the base station is the TAU Request, which contains information like UE Core Network Capability, old GUTI, last visited TAI, etc. [1]. The old Globally Unique Temporary UE Identity (GUTI) is the identifier of interest to our algorithm, because it consists of two main components, namely the GUMMEI and the M-TMSI. Since the GUMMEI uniquely identifies the MME which has allocated the GUTI, we have all the information needed to update all counters for equation (1).

When applied to a network organization problem (in our example handover region organization), a distributed approach has several advantages compared to a centralized approach. The main benefits are reduced signaling overhead, scalability and on-demand resource scaling. A centralized approach assumes the existence of a node in the network which is responsible for running the optimization algorithm. The first step is to gather all the information at the centralized node, which
involves a vast amount of signaling between the RAN and the core network. The second step is to run the optimization algorithm, which is in our case an NP-hard problem (graph partitioning problem) and therefore does not scale with the size of the network. As the next generation of communication networks assumes a fast changing environment, which is supported by the on-demand resource scaling (e.g., network function placement), the organization of the RAN should be able to adjust to the dynamic core network. In case of a centralized implementation this would result in more signaling and due to the computational complexity of the optimization problem, the RAN organization adaptation would be delayed. A distributed implementation uses local information available at the nodes in the RAN, and therefore it does not require additional signaling. Additionally, a distributed adaptive solution enables dynamic reorganizations according to the on-demand resource scaling in the core network (i.e., each node decides based on its local information how to react to changes like new instances of core nodes, which does not involve additional computation and communication delays).

REFERENCES