Performance Evaluation of a SDN/OpenFlow-Based Distributed Mobility Management (DMM) Approach in Virtualized LTE Systems

Luca Valtulina1,2, Morteza Karimzadeh1, Georgios Karagiannis1,3, Geert Heijenk1, Aiko Pras1
1Design and Analysis of Communication Systems (DACS), University of Twente, Enschede, The Netherlands
2SpeakUp, Enschede, The Netherlands
3Huawei Technologies, Düsseldorf, Germany

Abstract— Currently most of the mobility management solutions rely on a centralized mobility anchor entity, which is in charge of both mobility-related control plane and user data forwarding. This makes mobility management prone to several performance limitations such as suboptimal routing, low scalability, potential single point of failure and the lack of granularity for the mobility management service. Distributed Mobility Management (DMM) is a mobility management solution that can be applied to overcome these limitations. In this paper we introduce a novel Software Defined Networking (SDN)/OpenFlow based DMM approach that can be applied in virtualized LTE systems. Using NS-3 simulation experiments we show that the introduced approach meets the performance related mobility management requirements.

Keywords—SDN, NFV, Virtualized LTE, DMM, IP address continuity, OpenFlow switches

I. INTRODUCTION

Mobile operators are currently focusing on providing technological solutions for the issues arising from the significant growth of data traffic due to the continuous increase of mobile users, devices and applications. Long Term Evolution (LTE) as the fourth generation (4G) cellular system, standardized by the 3rd Generation Partnership Project (3GPP) [1], is the most promising approach to cope with this challenge. The main goals of LTE are to substantially improve end-user throughput, e.g., up to 100 Mbps in the downlink and up to 50 Mbps in the uplink, and reduce user plane latency, e.g., under the 5 ms, while significantly improve user experience with full mobility.

Two main network parts in the LTE system are the Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) and the Evolved Packet Core (EPC). The E-UTRAN consists of base stations, denoted as Evolved Node-Bs (eNodeBs), which control different cells and provide radio coverage and connectivity between the User Equipment (UE) and the EPC. The eNodeBs are interconnected with each other by means of the X2 or S1 interfaces. The EPC is composed of several network elements. The key entities are the Serving Gateway (S-GW), the Packet Data Network Gateway (P-GW) and the Mobility Management Entity (MME). The P-GW, as the main EPC mobility anchor point, connects the EPC to external networks and is in charge of mobility related user data forwarding. Furthermore, it also performs various functions such as IP address/IP prefix allocation. The S-GW supports the transport of the user data between the E-UTRAN and the EPC. The MME, as a control node, is in charge of mobility management signaling (i.e., handover) between the UE/E-UTRAN and the EPC. [1].

Mobility management refers to a set of mechanisms to keep ongoing session continuity while a mobile user changes its mobility anchor point in the network. Real-time IP applications such as video conferencing, Voice over IP (VoIP), online gaming, download/upload of large size files (particularly in cloud computing environment) are examples of notably demanding applications in mobile network environments in which supporting IP mobility and seamless session continuity is a necessity for the users changing their mobility anchor points during inter-operator and intra-operator/technology handovers. Most of the existing IP mobility solutions standardized by both IETF (Internet Engineering Task Force), e.g., MIP, PMIP [3], and 3GPP [4], rely on a centralized mobility anchor entity which is in charge of both the mobility-related control plane and the data plane. The presence of a centralized network node makes mobility management prone to several restrictions such as centralized management of one/several hierarchical tunnels for each Mobile Node (MN), processing overhead due to encapsulation/de-encapsulation during tunneling updates, a potential single point of failure and resulting reliability issues, network bottleneck, and non-optimal routing (see e.g., [5],[6]). Over the last few years, operators and research communities are investigating alternative mobility solutions that exploit a flatter system in which the mobility anchors are distributed and located at the edge of the access network. In this approach, denoted as Distributed Mobility Management (DMM), only the data plane (partially DMM) or both the data and control planes (fully DMM) may be handled by the distributed anchor points. Double NAT (D-NAT) [7], Distributed Mobility Anchoring (DMA)[8], Inter-domain DMM, Local IP Access (LIPA) and Selected IP Traffic Offload (SIPTO) [9] are examples of such approaches. As we argued in [10], an OpenFlow-based SDN architecture is a promising candidate approach, supporting DMM concepts and outperforming existing solutions. In this paper we aim to demonstrate that the proposed DMM solution is able to provide the expected level of Quality of Service (QoS) for the end-user both in terms of latency and throughput of the mobility redirected traffic.
The EU FP7 Mobile Cloud Networking (MCN) project [11] integrates the use of cloud computing concepts in LTE system in order to increase its performance. This is accomplished by building a shared distributed LTE mobile network that can optimize the utilization of virtualized computing, storage and network resources and minimize communication delays. The integration of cloud computing concepts in a LTE system, as shown in Fig. 1, can be realized by; (1) extending the cloud computing concept beyond the typical (macro) data centers towards new smaller (micro) data centers that are distributed within the E-UTRAN and the EPC, and (2) deploying and running cloud-based E-UTRAN, and EPC denoted as RAN as a Service (RANaaS) and EPC as a Service (EPCaaS), respectively. This approach could support on-demand provisioning and elasticity of the LTE's cloud based components, allowing to automatically scale in/out, based on the demanded data traffic load. This trend is also in line with the emerging European Telecommunications Standards Institute (ETSI) activities in Network Functions Virtualization (NFV).

A DMM approach that relies on the OpenFlow-based SDN architecture can be applied in the cloudified LTE environments to support seamless ongoing session continuity while a mobile user is roaming/moving and its original mobility anchor (e.g., P-GW) is relocated. Furthermore, this approach could support the user's traffic redirection when the virtualized mobility anchor entity, running on a virtualization platform (e.g., originating data center), is migrated to another virtualization platform (e.g., destination data center) and ongoing sessions supported by this entity need to be maintained. In this paper we introduce and evaluate a novel SDN-based DMM approach. In particular, the research questions that are answered by this paper are:

1. How can the OpenFlow-based SDN architecture be applied for DMM support in virtualized LTE systems?
2. Is the provided solution able to support ongoing session continuity seamlessly?

This paper is organized as follows. The DMM architecture is introduced in Section II. Section III introduces the proposed DMM approach and answers the first research question. Section IV presents the performance evaluation experiments and analyzes the obtained results answering the second research question. Finally, Section V concludes the paper and provides recommendations for future work.

II. DMM FUNCTIONAL FRAMEWORK AND ARCHITECTURE

In the context of the IETF DMM working group, [12] defines a functional framework for DMM and introduces a set of Functional Entities (FEs) which are required to support IP address continuity in a network with distributed mobility anchors. IP address continuity is a property of a network to seamlessly hand over an ongoing session from a source mobility anchor to a destination one during the lifetime of the session. In this paper the DMM functional framework specified in [12], is used and applied for cloud based LTE systems as follows.

A. Proposed DMM Architecture in a virtualized LTE system

Fig. 2, depicts our proposed DMM architecture. The FE_MCTX (Functional Entity Mobility Context Transfer) provides mobility context information to the control function (FE_ICE) to allow forwarding of packets to the UE's currently used mobility anchor. This function could be co-located with the MME entity which provides mobility context information to other EPC elements (e.g., P-GW). FE_I (Functional Entity Ingress) establishes the forwarding of the UE's data packets to the appropriate Egress function (FE_E). FE_E (Functional Entity Egress) is a function that receives downlink packets forwarded by the DMM Ingress function (FE_I) and terminates the data forwarding plane. FE_ICE (Functional Entity for Ingress/Egress Control) is a function to establish, update and remove policies in the DMM Ingress (FE_I) and Egress (FE_E) functions to allow forwarding of UE's data packets towards the currently used mobility anchor (target P-GW). UE's mobility context information is delivered to this function by the FE_MCTX function (MME) upon triggering of the specific handover procedure.
tunnel to the Egress function, address translation to a routable IP address or other means. In order to avoid encapsulations/de-capsulation processing overhead during the commonly deployed IP tunnelling (and tunnel updating) to deliver the UE's downlink traffic to the currently used mobility anchor, alternative forwarding techniques are introduced. The source address carried by the uplink packets is also topologically incorrect. For the uplink packets to be assumed as routable, IP routers of the mobility domain must not apply filtering according to the source addresses. Since IP address continuity is a problem concerning downlink traffic, this paper focuses mainly on this particular case.

In the current 3GPP's LTE standard, IP address continuity for users that change their EPC mobility anchor point is not supported. This means that an ongoing flow cannot survive the change of the user mobility anchor in the current LTE network system. Certain modifications are therefore required in the protocol, particularly on both X2 and S1 handover procedures. We have described these modifications in detail in [13].

### III. SDN/OpenFlow Based DMM Solution

Based on our previous study [10], OpenFlow-based SDN is the leading candidate approach for efficient mobility management in virtualized LTE systems. The SDN architecture offers a logically centralized control model, which decouples control and data planes, and enables the underlying infrastructure to be abstracted for applications and network services. SDN makes networks programmable, manageable, and adaptable to a wider extent that is ideally suited for highly scalable mobile wireless networks. Capabilities offered by OpenFlow, as the most common communication protocol used in SDN approach [14], would be as an enabler to access the forwarding plane of OpenFlow-enabled switches over the network and reconfigure it according to the needs of applications and network services. The flow-tables in OpenFlow-capable switches can be remotely programmed via the OpenFlow Controller (OC), such that each traffic path could be traced from an Ingress node to an Egress node as a separate flow and traffic could be redirected to a new mobility anchor point without any IP address translation or modification. OpenFlow-enabled switches contain a set of actions that could be applied to flow-specific packets providing further capabilities. The Set-Field action is the most relevant one for our purpose, giving to OpenFlow switches the possibility to modify packets' and frames' headers. A combination of Set-Field and Output actions (specifying the output interface) could be utilized to support dynamic per-flow redirection. Both flow-tables and action lists modifications are accomplished by the OC via a dedicated secure connection with each switch [14].

Two main different DMM solutions can be deployed using the set of features provided by OpenFlow as Full OpenFlow and Partial OpenFlow. In the first approach, all switches in the operator's transport network are OpenFlow-enabled and no modifications to incoming packets are needed for traffic redirection. For the Partial OpenFlow solution, a hybrid network model where OpenFlow switches are placed at the downlink Ingress (FE_I) and Egress (FE_E) points while IP routers are used in the core network, is introduced. In this approach, traffic redirection on the core part is based on currently deployed layer 3 routing mechanisms. The choice of which design is more suitable depends on the operator's investment roadmap, given the fact that switching to a full OpenFlow-enabled network can be a large economic investment. In [13], both solutions have been addressed. However, due to the paper length limitations, only the Partial OpenFlow based approach is considered in the following.

#### A. Partial OpenFlow based DMM

Fig. 3, shows the Partial OpenFlow based DMM approach topology based on the DMM architecture presented in Section II.A. In this topology the OpenFlow-enabled Ingress (FE_I) and Egress (FE_E) switches are managed by the same OC (acting as the FE_IEC). When a UE attaches to the access network for the first time, it gets an IP address allocated from its serving P-GW, which is the mobility anchor for flow(s) initiated by the UE when attached to it. This IP address is used to forward traffic only outside of the operator's transport network (e.g., in the EPC, in the Internet, etc.). To provide IP address continuity when a UE changes its mobility anchor point, while keeping some flows active, inside the operator's transport network a different IP address will be used to route traffic to the current UE's P-GW. This address can be the SGI address of the current UE's mobility anchor point. Ingress and Egress OpenFlow switches will take care of replacing the original destination IP address with the SGI address using the Set-Field action. The Output action is needed in both Ingress and Egress OpenFlow switches to forward incoming downlink traffic to the transport network and EPC network, respectively.

![Partial OpenFlow based DMM](image)

Fig. 3. Partial OpenFlow based DMM (before and after handover)

Fig. 3, represents the downlink path for the flow before and after that UE has been handed over from the source eNodeB to the target eNodeB (X2 capabilities between eNodeBs can be used to forward user's downlink traffic during the handover procedure, if available). Assume that the UE was previously attached to the source P-GW and had initiated a flow using an IP address allocated from the P-GW's address pool (for instance 10.0.0.1). When it is handed over to the target P-GW and wishes to keep the previously initiated flow active, Set-Field instructions need to be present into the action set of both Ingress OF and Egress OF-2 switches (to replace the destination IP address (10.0.0.1) with the target P-GW's SGI address in the Ingress OF and to do the contrariwise
IV. PERFORMANCE EVALUATION AND EXPERIMENTS

This section describes the simulation experiments that have been accomplished in order to evaluate the proposed DMM solution. The impact of various positions for both OC and Ingress OpenFlow switches has been studied. In the experiments, a session is considered to be handed seamlessly if its mobility management procedure is accomplished in less than 150 ms. This threshold is the maximum one-way latency for real time applications (e.g., VoIP) specified by ITU-T G.114 [15]. Furthermore, we also studied the impact of migrating a virtualized mobility anchor (P-GW) or part of its functionalities during a handover procedure. These experiments have been performed using the NS-3-LENA simulation environment [16]. The logical topology used in the simulation experiments is shown in Fig. 4. UEs are handed over from the source eNodeB to the target eNodeB being served by two different P-GW namely the source P-GW and the target P-GW. X2 handover capabilities are assumed to be available between the eNodeBs. IP flows are initiated by UEs when attached to the source eNodeB using an IP address allocated from the source P-GW address pool. After the X2 handover procedure is completed, the Partial OpenFlow based DMM solution is used to redirect the traffic to the current UEs' mobility anchor point (target P-GW). It is assumed that, all the data centers, which are running the virtualized LTE's components (e.g. P-GW), are connected with each other via the operator's IP transport network. The topology used for the operator's IP transport network is a small part of the real network topology of one of the biggest European ISP (EBONE). In the simulation scenario the backbone part of the network uses 10-gigabit Ethernet links while the rest of the operator's network and the Internet network use one-gigabit Ethernet links. In order to emulate realistic situations, it is considered that all wired and wireless links are 80% utilized using background traffic during the simulation experiments. The generated background traffic on the LTE wireless links is a traffic mix based on [17]. VoIP traffic represents the 30% of the total background traffic. The background traffic on all the wired links is generated using the PPBB (Poisson Pareto Burst Process) model specified in [18] to match the statistical properties of real-life IP network. The source code used for the implementation of the traffic generators, simulation topology, DMM solutions, LTE, network, OpenFlow structures and handover trigger model is available via [13].

A. Performance metrics

The following metrics have been used to evaluate the seamlessness of the introduced DMM solution.

• Average latency of data packet delivery via X2 tunnel (ms). It represents the downlink path average latency of data packet(s), pushed from the source to the target eNodeB via the X2 path during inter P-GW handover procedure.

• Average latency of first received DMM-redirected data packets (ms). This metric represents the downlink path average delay of the first data packet received by the moving UEs after the completion of the handover.

• Cumulative Distribution Function (CDF) of the latency of the first received DMM-redirected data. It is used in order to observe the maximum latency of the first DMM-redirected data packet after the completion of the handover procedure. Since VoIP is the studied UEs' traffic type, the CDF will be used to demonstrate how the measured DMM-redirected data packets latencies fit with respect to the previously introduced 150 ms threshold.

• X2 path throughput (packet/s). It shows how many downlink data packets have been successfully received by the UE via the X2 path over the simulation time. The X2 path is used upon handover as long as the DMM traffic redirection is not ready yet.

B. Simulation experiments

In this research two sets of experiments have been defined:

• X2 handover with P-GW relocation: Evaluation of DMM performance in mobility scenarios, which require P-GW relocation.

• X2 handover with virtual P-GW migration: Evaluation of the impact of migrating the virtualized target P-GW entity or part of its functionalities during a handover procedure.

In these experiments 150 UEs are initially attached to the source eNodeB and are generating VoIP traffic following the traffic models mix described in Section IV. The reason for choosing this amount of UEs is given by the fact that an 80% level of cell utilization is used in the simulations. Furthermore, 80% of the source eNodeB's active UEs (120) are considered static and remain attached to the source eNodeB for the entire simulation time. Their positions are uniformly distributed around the eNodeB in a disc with radius equal to 2 km. The
remaining 30 UEs (running a VoIP application with a peer remote host) will be handed over to the target eNodeB during the execution of the simulation. With the same setup as the source eNodeB's, 120 UEs are initially attached to the target eNodeB and are generating traffic using the same mix. This is done in order to reach also a maximum level of cell utilization when all the UEs have been handed over from the source eNodeB. In all the performed simulation experiments 95% confidence intervals are computed. The following parameters are varied during the simulation experiments:

- **Average distance (hops) between Ingress and Egress points.** Three different positions for the Ingress (FE_I) and Egress (FE_E) switches, with the average hop counts (distance) of 1, 4 and 7 between them, have been defined.

- **DMM Controller position.** Placement of the DMM controller (OC) in the operator's network can be critical. In these experiments three positions have been chosen; (1) co-located with the MME, (2) in the middle of the core network (CN), at same distance from both Ingress and Egress points, and (3) on the outside of CN, closer to the operator's Internet PoPs, where the distance from Egress functions reaches the 9 hops with a mixed link speed (1 Gbps or 10 Gbps).

- **Delete Session Timer:** It reflects the functionality of the timer used in the 3GPP's standard by the MME to delete the moving UE's session(s) in the source P-GW. If it is set to 0 ms, the moving UE's session(s) will be removed in the source P-GW upon the path switch procedure has been completed.

### C. Experiment results

- **X2 handover with P-GW relocation**

  After the completion of the X2 handover procedure, the Partial OpenFlow based DMM solution is used to redirect the traffic to the current UEs' mobility anchor point (target P-GW). Fig. 5, shows the average latencies of downlink data packets delivered via the X2 path and via DMM traffic redirection to the moving UEs. The graph shows that independently from the distance between Ingress and Egress OpenFlow switches and the location of the OC, a seamless handover is experienced by all the moving UEs.

Fig. 6, shows the CDF of latency of the first DMM-directed data packet after the completion of the Partial OpenFlow traffic redirection procedure.

The CDF clearly shows that in all studied scenarios there is a 100% probability of having DMM-directed redirected data packets which do not exceed the maximum 150 ms latency threshold of a VoIP session. Furthermore more than 95% of the observed packets have latency lower than 50 ms.

The impact of placing the Ingress and Egress OpenFlow switches at different distances (hops) with respect to the positions of the OC is studied by analyzing the load and throughput of data traffic forwarded via the X2 path. Fig. 7, shows that, independently from the distance between Ingress and Egress switches, the X2 tunnel is equally utilized when the OC is co-located with the MME. Only few packets are lost when the Delete Session Timer is set to 0.

As Fig. 8, illustrates a similar behavior has been observed for both cases when the OC is located in the middle of the operator's transport network, and when the OC is located outside of the operator's transport network.
To study the impact of the X2 data forwarding capability, the total load and throughput of the system have been collected both for the case; (1) when X2 data forwarding is available and the UE's session is kept active in the source P-GW also during handover execution phase, and (2) for the case when X2 data forwarding is not used during handover (see [13] for more detail). The results show that independent of the OC's positioning, packets are lost when they are not forwarded via the X2 path. The gap between the throughput of the system with X2 data forwarding and the throughput of the system without X2 data forwarding increases when the OC is placed further away from the EPC.

• X2 handover with virtual P-GW migration

The impact of VM (e.g., P-GW) migration is on the latency of the first packet received after the completion of the migration procedure. The average latency is in the order of seconds. Obviously, none of the observed latencies were below the standard one-way delay of a VoIP session, making as expected, the implemented mobility scenario non seamless. As a result, a mobility prediction function is required in the network to cope with the high latencies introduced by the migration of a VM entity or function which tasks cannot be temporarily replaced. More details can be found in [13].

CONCLUSIONS AND FUTURE WORK

This paper introduced and evaluated a novel SDN/OpenFlow based DMM approach, denoted as Partial OpenFlow based DMM that can be applied in virtualized LTE systems. In particular, the paper mainly studied whether this novel DMM solution is able to support the continuity of a session that is handed over from a source P-GW to a target P-GW in a seamless way. The results show that when the X2 path is used for support of the inter P-GW handover, the sessions can be handed over seamlessly to the target P-GW, by using the proposed approach, for a duration less than 150 ms. The best results in terms of latency in configuring traffic redirection in the DMM transport network were obtained when the OC has been positioned in the middle of the core network and the distance between Ingress and Egress OpenFlow switches was lower than or equal to 4 hops. In all the studied network topologies and independently of the positioning of the OC, it is concluded that X2 data forwarding is required also for a period of time following the conclusion of the path switch request procedure. The experiments also show that, in the case of the migration of the virtual P-GW from one source data center to a target data center, a function to predict the mobility of UEs need to be present in the network in order to setup DMM traffic redirection prior to trigger of the handover procedure.

For the future work, bigger network topology scenarios needs to be investigated. Furthermore a scenario where also the MME entity is relocated may be studied and evaluated. The impact of utilizing multiple DMM controllers during ongoing sessions and within the same DMM transport network requires further research since it may impact the latency in configuring the traffic redirection and the design of the signaling protocol between EPC and DMM-plane. Implementation a prototype of the proposed architecture in OpenStack [18] virtualization test bed as a supplementary validation, could be as the next research, as well. Last but not the least, the use and integration of a mobility prediction function, when the virtualized anchor point has to be migrated, in the DMM solution needs to be evaluated.

ACKNOWLEDGEMENTS

This work has been funded by the EU FP7 Mobile Cloud Networking (#318109). We would like to thank the MCN project partners for their contributions and comments.

REFERENCES