On V2X Network Slicing: Using Context Information to Improve Mobility Management

Panagiotis Spapis, Chan Zhou
Huawei German Research Center
Munich, Germany
email: panagiotis.spapis@huawei.com
chan.zhou@huawei.com

Alexandros Kaloxylos
Department of Informatics and Telecommunications
University of Peloponnese
Tripoli, Greece
email: kaloxyl@uop.gr

Abstract— **Network slicing in for 5th Generation (5G) networks enables the support of multiple logical networks, tailor-cut to the requirements of specific services. Initial specifications have already been produced by the 3rd Generation Partnership Project (3GPP) that describe the operation of slicing. However, the existing specifications lack specific details on how the network functions can be fine-tuned to fully optimize the network performance for specific use cases. This paper provides a comprehensive overview related to the latest status of the 3GPP standardization process related to slicing. Also, it proposes a new mobility management scheme, called Context Enhanced MOBility management (CEMOB), that is tailor-cut for communicating vehicles. The point we make is that by taking advantage of contextual information, in this case for vehicle-to-everything (V2X) communications, the performance of network control functions such as mobility management can be significantly improved. Slicing and the overall 5G architecture, support a simple introduction of contextual information into the network functions of a slice.**

Keywords-network slicing, mobility management, V2X communications.

I. INTRODUCTION

5G networks target, apart from the support of the telecommunications sector, also the "vertical industries" like autonomous driving, smart factories, new health services, etc. An extensive list of 5G use cases can be found in [1] [2]. A thorough examination of the verticals has identified that these sectors have diverse requirements. These requirements are mapped to different network Key Performance Indicators (KPIs). These KPIs indicatively include throughput, transmission reliability, latency, energy consumption, blocking probability, etc. Since every vertical has a different operation environment in terms of the density and mobility of the users, the arrival rate and the duration of different application and services, it is evident that no single network can support efficiently all these different use cases.

Thus, it appears that the deployment of parallel logical networks over the same network infrastructure is a necessity. These logical networks may have network functions (NFs) configured differently or even introduce new network functions both in the Radio Access Network (RAN) [3] as well as the Core Network (CN) [4].

The 3rd Generation Partnership Project (3GPP) has defined a network slice to be " A logical network that provides specific network capabilities and network characteristics" [5]. A "Network Slice" is implemented by a "slice instance" that in its turn is created by a "network slice template". The latter is a template that defines a complete logical network including the NFs, their interfaces and their corresponding resources.

Network slicing has been intensively investigated during the past years both by industry and academia. There are several research proposals that target full flexibility in terms of selecting, organizing and deploying NFs [6]. At the same time, 3GPP is currently working on the phase one specifications for 5G networks that include also the support for slicing. The standardization activities follow a more cautious path and attempt to re-use existing NFs or share NFs across different slices as much as possible. Note that although the use cases to be supported, as well as their requirements, have been thoroughly studied [7], current specifications do not provide fully tailor-cut solutions for them. In order to do this, it is needed to work really close with the representatives of the so called "vertical industries" (e.g., transportation, health, factories, energy). This is needed to understand not only the requirements and the operational environment, but also the contextual information produced and how these can be used to optimize network functions. For example, the newly founded 5G Automotive Association 5GAA [8] is working towards such a direction. Still, the activities towards the proposal of mechanisms driven by these organizations in the standardization are in primitive steps.

In the current paper, we present the latest status of the standardization activities related to network slicing. We also provide a new mobility management mechanism for autonomous driven vehicles that takes advantage of contextual information and we demonstrate how this information can be used by the standardized 5G NFs to bring significant benefits in a control operation such mobility management. This is an exemplary scheme to highlight that different requirements need very different solutions and the network shall support these solutions.

The rest of the section is organized as follows. In Section II we provide the latest status of 3GPP in relation to slicing. Section III discusses how mobility management is planned to be supported in the technical specifications and why we consider this not to be applicable for moving vehicles. In Section IV, we provide the details of our scheme design to use the 5G network functions. In Section V we present

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-596-8

quantitative results that illustrate the benefits of our scheme. Finally, Section VI concludes the paper and describes future directions.

II. SLICE SUPPORT IN 3GPP

3GPP has decided to treat 5G specifications in two phases. The first one is to be completed by September 2018 (Release 15). This phase addresses a more urgent subset of the commercial needs. Phase 2 is to be completed by March 2020 (Release 16) for the IMT 2020 submission, having addressed all identified use cases & requirements. In relation to slicing, several working groups are currently progressing on the key elements and procedures that have to be specified.

In [5] and [9], the 5G network architecture is presented. There, a list of technical key issues, as well as potential solutions for slicing is presented. For example, in these documents the issues of slice selection, slice isolation, sharing of NFs, multi-slice connectivity, management of slices, etc. are being addressed.

Although several issues remain open, it seems that there is convergence in several principles. The first principle is that NFs, previously incorporated in monolithic network components, are now decomposed to smaller modules. The target is to allow a synthesis and configuration of the NFs on a per slice type basis. A second principle is the further splitting of user and control plane functions to facilitate a more flexible evolution of NFs. A third key principle is the exposure of NFs to service through appropriate APIs. This is expected to allow a better collaboration among network operators and service providers.

Figure 1 presents a summary of the supported NFs. The CP function in the CN are considered to be the following:

- **Unified Data Management (UDM):** supports the Authentication Credential Repository and Processing Function (ARPF).
- **Authentication Server Function (AUSF):** supports the Authentication Server Function (AUSF)
- **Policy Control function (PCF):** supports unified policy framework to govern network behaviour, provides policy rules to control plane functions
- **Core Access and Mobility Management Function (AMF):** supports mobility management, access authentication and authorization, security anchor functions and context management
- **Session Management Function (SMF).** supports session management, selection and control of UP functions, downlink data notification and roaming
- **User Plane Function (UPF):** is the anchor point for inter/intra RAT mobility and the external PDU session point of interconnection, supports packet routing and forwarding, QoS handling for user plane, packet inspection and policy rule enforcement
- **Network Exposure Function (NEF):** provides a means to securely exchange information between services and 3GPP NFs.
- **NF Repository Function (NRF):** maintains the deployed NF Instance information when deploying/updating/removing NF instances
- **Network Slice Selection Function (NSSF):** supports the functionality to bind a UE with a specific slice

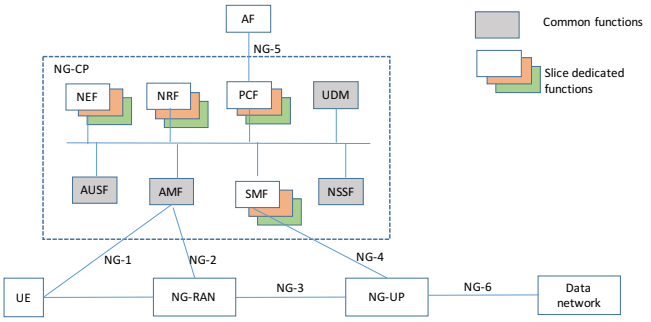

Figure 1: 5G service based architecture (adapted from [5])

Note that some of these functions are common for all slices while others can be dedicated for different slices. A UE may access multiple slices concurrently via a single RAN. For such cases, it is assumed that the involved slices should share some control plane functions, like the AMF. The abovementioned logical network allows the support of Application Functions (AF) and provides connectivity to typical external data networks.

Interestingly enough the question whether RAN is sliced or not still remains open. However, it has been agreed that RAN will be slice-aware so as to treat slice traffic according to the customer needs. Also, RAN shall support resource isolation among slices so as to avoid shortage of shared resources in one slice to break the service level agreement on another [6].

However, detailed alternative solutions have been proposed on how RAN is involved in slice selection by passing an appropriate identifier to the core network elements. Currently slicing for RAN essentially focuses on different scheduling schemes for different slices and also by providing different L1/L2 configurations. Moreover, it is considered that even if a User Equipment (UE) is connected to multiple slices a single Radio Resource and Control (RRC) entity will be used were as different protocols (i.e., Packet Data Convergence Protocol – PDCP and Radio Link Control - RLC) can be used.

Every slice is identified by a Single Network Slice Selection Assistance information (S-NSSAI) identifier. This identifier consists of a Slice/Service type (SST) and a slice Differentiator (SD) which optional information used to differentiate among different slices of the same type. Currently only 3 SST values have been agreed to be supported. These are a) enhanced Mobile Broadband (eMBB), b) Massive Internet of Things (MIoT), and c) Ultra Reliable Low Latency Communications (URLLC) [5]. This information is exchanged as part of Non-access stratum signalling through the RAN.

In [11], the lifecycle of a network slice is described by the following phases: a) Preparation phase, b) Instantiation,

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-596-8

Configuration and Activation phase, c) Run-time phase and d) Decommissioning phase.

III. CURRENT STATUS FOR MOBILITY MANAGEMENT IN 5G NETWORKS

Mobility management for legacy systems was performed as follows. The network was divided into non-overlapping regions called Tracking Areas (TAs). Idle UEs would have to inform the network each time they cross the border of such areas or when a timer, typically set at 54 minutes) expires. However, this design was initially static and the cost for re-arranging the coverage areas of TAs was quite high. Moreover, a problem appeared from excessive Tracking Area Update (TAU) messages due to the movement of the users near the Tracking Area boarders. That's why the notion of Tracking Area Lists (TAL) was introduced. TALs were assigned per UE and allowed the overlapping of TAs. The algorithm to define the TAL is proprietary and the operator according to his strategy decides whether to allocate large or short TALs. Whenever a UE has to be discovered for delivering data to it or in case of an incoming call, paging is executed in a subset or all the cells in the TAL according to the operator strategy [12]. If a subset of the cells of the TAL is paged there is a risk of increased delay due to page misses or but if all the cells are pages there is increased signalling cost. Also the size of the TAL relates to a signalling tradeoff since small TALs have reduced paging signalling cost but require frequent TAU and large TALs vice versa.

Even with these improvements, it has been noticed that for idle UEs that had to switch to connected mode, signalling had again to be exchanged up to the core network and more specifically the Mobility Management Entity (MME) where contextual information (such as security credentials) were kept. Considering that smartphones have a number of applications (e.g., facebook, skype, instant messaging) that have to wake up asynchronously and exchange small amount of information, this created in practise significant signalling load.

This is why for the 5G systems, mobility for idle terminals had to be redefined [10]. In the latest specifications, the RAN-based Notification Area (RNA) has been defined. This can be considered as a smaller subset of a TAL where a UE can move within without informing the network about its exact location. Also, a new state called RRC_INACTIVE is introduced where the context information of a UE is kept locally so as to avoid contacting the CN entities (i.e., AMF) when the UE switches again to the connected mode. This addresses the problem of frequent waking-up of devices (e.g., smartphones) and minimizes that signalling load towards the CN. In terms of mobility management, the UE context is kept in the last serving base station, called gNB in 5G systems. If the UE wakes up and becomes connected under a new gNB inside the same RNA then it uses the *RRCConnectionResume* messages to force the new gNB retrieve its context from the last serving gNB. The new gNB may also trigger a path switch by communicating with the AMF. Paging a UE takes place from the last serving gNB to all gNBs that are member of the RNA. These procedures are illustrated in Figure 2. On top of these messages, note that whenever a UE crosses the border between RNAs it needs to receive the gNBs identifiers that are members of the new RNA.

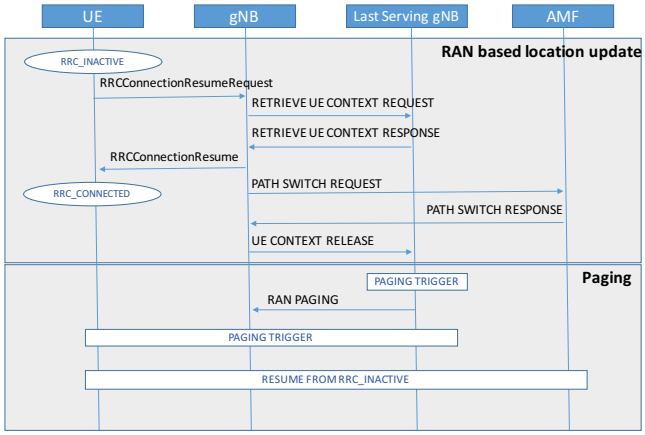

Figure 2: RAN based mobility management (adapted from [10])

This mechanism treats indeed several of the inefficiencies present in existing cellular systems. However, as explained in [13], the RAN based mobility management scheme suffers from excessive load for high moving UEs. This is why Hailu and Säily suggest a hybrid scheme where a typical CN mobility management takes places for high moving UEs, while RAN based mobility management is executed for UEs of lower mobility. To do this the UEs have to report the mobility to the CN at some intervals (e.g., during location update). Moreover, the authors also indicate potential delay issues that may arise if there is no direct interface between the last serving gNB and the new one. In such a case signalling has to travel essentially through the CN. The lack of a direct link between base stations is not uncommon in deployed and operating mobile networks. Note that in the current standard specification both the typical CN mobility management as well as the RAN based are supported.

The adoption of RAN based mobility management scheme will be beneficial for some of the 5G use cases but totally inefficient for others. A not applicable use case is the one of the autonomously driving vehicles because of the high mobility. To optimize a control procedure like mobility management, one has to take advantage of contextual information that will be available to operator as we discuss in the next section.

IV. CEMOB: CONTEXT ENHANCED MOBILITY MANAGEMENT

Autonomous driving is one of the key targets of the industry for the next decade. 3GPP has already specified an architecture and mechanisms to support inter-vehicle communication and access to service specific servers (i.e., V2X application server - [14]). The support of such services introduces additional contextual information that if used can greatly improve even control operations for a mobile network. More specifically, it is expected that in order to form a route, a vehicle will communicate with a server to receive the path to be followed. These servers can also estimate the time a vehicle will need to be at a certain position in the path. Such functionality exists even today with well-established

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-596-8

applications like Google maps or any other GPS navigators. Obviously, these applications do not and they should not know any information about the deployed base stations of an operator. However, by passing the information of a route to an operator, it is an easy task to perform a translation of path coordinates to predicted serving gNBs. Furthermore, the specific geography of the roads can significantly assist in determining the exact cells a vehicle is going to pass through. Such information can be used to really optimize mobility management operation by optimizing the TALs allocation, and at the same time optimizing the paging strategy. Additionally, the functions modularization in 5G facilitates the optimum functionality placement in the network which for the mobility management functionality in certain use cases (like the V2X ones) would make sense to be split between the RAN and the Core. Moreover, 5G networks will allow, through secure APIs, for services to communicate with network components and exchange information.

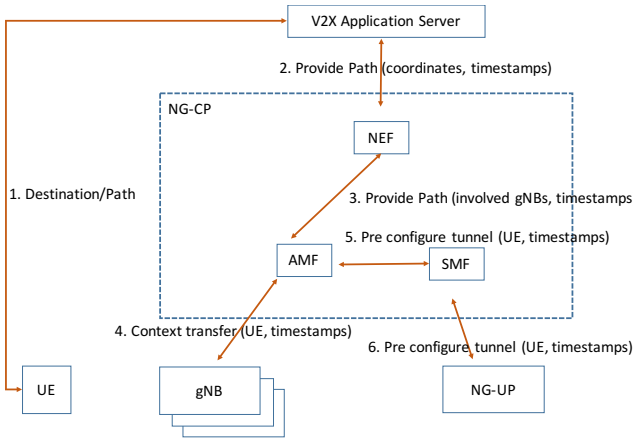

Figure 3: Mobility management for vehicles in 5G networks

In Figure 3, we present how a new mobility management scheme for vehicles operating in 5G networks can be implemented. Whenever a UE/vehicle wants to reach a specific destination, it will communicate with a V2X application server and it will receive the path so as the computer inside the car to start the autonomous driving functions. Upon calculation of such a path by the server the information in terms of coordinates and timestamps (time when the vehicle will be at a specific point) can be communicated to the mobile operator. This will take place through message exchange with the NEF. The NEF can also translate the coordinates into specific gNBs and forward further this information to the involved AMFs. These entities on their turn can transfer the UE context to the involved gNBs. Moreover, they will communicate with the corresponding SMFs so as to pre-configure the data path for the vehicles. Note that this pre-configuration does not imply that resources will be allocated for large period of times but rather only for a short time for which a vehicle is expected to be in a certain area. Obviously, a vehicle (or the respective server) may need to re-calculate a path, but this again will take place through the same communication with the V2X application server, so the same process will be repeated. The communication of a UE with an application server, especially if located inside the domain of the mobile operator can be in terms of a few tenths of millisecond [15], thus any updating of network components is not expected to affect the location management process, since even high moving vehicles will not have change their position more than a few meters.

When a UE wants to communicate with a neighbouring one, the request will stop in the gNB and the gNB mobility management function will perform the paging to this cell and the neighbouring ones, since there is no need to communicate with the core for transferring the UE context in the RAN because it already resides there and the actual location of the UE is well known with quite good accuracy.

The benefits of such scheme are manifold. Firstly, the mechanism is fully optimized for moving UEs no matter their speed. Thus, it is not necessary to revert to the typical CN mobility management scheme if their speed is high and reach high paging load. Secondly, there is no need to exchange control messages for UE location updates over the wireless interface which is the bottleneck for any wireless system. Furthermore, delay for transferring the context information of a UE from a serving to a new gNB is zero, since this information is in place before hand. This delay in the RAN based scheme can be significant as we have already explained in the cases where the gNBs have no direct interface and their communication takes place through the CN. Finally, the paging cost is significantly lower than the CN based scheme and as well better than the RAN based one since the known geography of the streets can minimize the number of involved cells only to very few ones. All these benefits are possible because the proposed scheme takes advantage of contextual information that can be available to the NFs of the mobile operator through the modularized architecture that allows different NFs to be used for different logical networks (i.e., slices).

In the next section, we quantify the aforementioned gains of the proposed scheme.

V. PERFORMANCE ANALYSIS

To evaluate the performance of CEMOB we compare it with the CN and RAN based mobility management schemes. Firstly, in order to calculate the signalling cost during paging we follow the analysis in [13]. Let M be the number of cells and N the number of gNBs. As an exemplary analysis we also consider 3 cells are supported by a single gNB. The RAN based scheme requires M messages over the radio plus N-1 messages (from the last serving gNB to the neighbouring gNBs inside the RNA). As for the CN based mobility management scheme, M messages over the radio plus N messages from the CN to the gNBs of an area (considered in this analysis of having the same size like the RAN based scheme), plus 6 additional messages that are needed to inform the CN NFs that the UE is in RRC_INACTIVE state. Concerning CEMOB, the knowledge of the position of a UE with a high accuracy even under some time coarse time period

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-596-8

require to page only the gNB where the vehicle is camped under. Also knowing the topology of the streets and the direction of the vehicle, it is easy to make sure that there will be no page miss by also paging the previous and the following gNBs. Considering an inter site distance (ISD) of even 500m, the vehicle is paged in an area of 1,5 Km that makes the probability of success rather high. As shown in Figure 4, as long as the number of gNBs increases the benefits of the RAN-based scheme, compared to CN based is rather low. On the other hand, CEMOB outperforms these two schemes considerably since we take advantage of the accurate information about the location of the UE/vehicle.

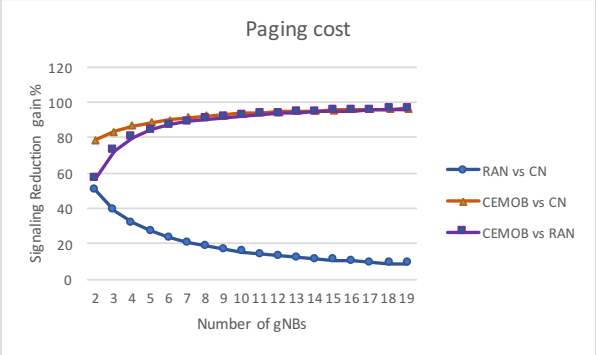

Figure 4: Paging cost CEMOB vs. RAN baseb vs CN based

To estimate the number of messages to be exchanged during a location update we perform the following analysis. As shown in Figure 2, for the RAN based scheme 7 messages need to be exchanged every time a UE crosses the border of an RNA or when it resumes an RRC connection in a gNB different from the last serving gNB. A similar number of messages is needed for the CN based scheme, but this time the communication takes places between a gNB and AMF instead of the last serving gNB. For the CEMOB case, context needs to be transferred to all gNBs of an area before the UE enters into it. Also, in case a UE selects with a probability p, a different path for any reason, then it will communicate again with the V2X application server and the context will have to be updated to all the gNBs of an area.

To perform an evaluation of CEMOB for the signalling load we consider an area of 15 gNBs divided into 3 RNAs. We also consider that a street has two lanes. According to [16], vehicle traffic flow with measurement at a point is "the number of vehicles that pass a point on a highway or a given lane or direction of a highway during a specific time interval". Traffic flow q is expressed in vehicles/hour is given by:

$$q = \frac{n_t}{t} \quad (1)$$

Related to the flow of vehicles the space headway parameter can also be used. It is defined as the distance measured between the front ends of two successive vehicles (as the sum of the vehicles' in-between space and a vehicle's length). Based on this parameter the traffic flow can be calculated as:

$$q = \frac{\bar{v}}{\bar{hs}} \quad (2)$$

where the flow q is calculated as the average speed of the vehicles divided by their average space headway. Based on this we are able to calculate the traffic flow of vehicles passing through the 3 RNAs border areas per hour. Our assumption is also that for the baseline, a UE will resume its connection once every 5 cells. Having also a fixed road topology and assuming a uniform distribution of vehicles with fixed space headway distance among them, it is easy to calculate the number of vehicles in this area. Using this number, we can select a probability that some of the vehicles will change their path, so CEMOB will have to update all the gNBs of an RNA. Figure 5 presents the results for different vehicle speeds (from 20 to 60km/h) and different space headways (from 4.5 to 22.5 meters). For this experiment , we consider that every 30 sec the 20% of the vehicles will request a path update.

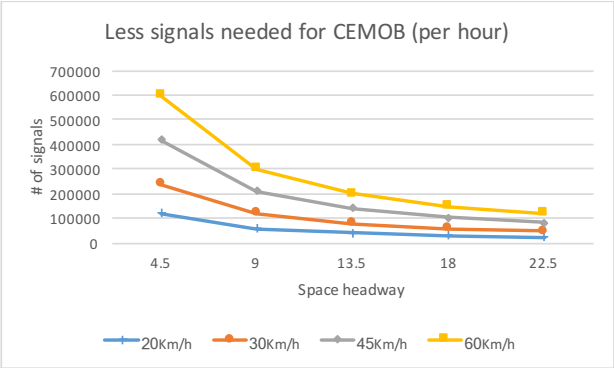

Figure 5: Signaling comparison between CEMOB and baseline scheme

As seen from the figure, CEMOB significantly outperforms the baseline scheme. The reason is that the on demand context transfer requires a lot of signalling even if this requested from one gNB to another. In this case the CN has to be notified so that path switching is performed. On the other hand, CEMOB has to notify the gNBs once and pre-configure the RAN-CN path at the same time. For a small number of cells like this discussed example topology, this means a considerable reduction. Also note that although CEMOB needs to update the gNBs every time a UE changes its path this is cost is at the end related only to the number of gNBs. In the case of the baseline, the cost is heavily affected by a complex process that may take place every time a UE is paged or resumes a connection to transmit data.

Obviously the penalty for CEMOB is the transfer of context information much more gNBs (all the gNBs inside an RNA) compared to the baseline where this is transferred only from one gNB to another. According to [17], the security information that needs to be transferred consists of a) K-ASME key (256 bits), b) K-eNB key (256 bits) and c) NONCE (32 bits). Also the Globally Unique Temporary UE Identity GUTI (80 bits) needs to be transferred to be associated with the abovementioned values.

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-596-8

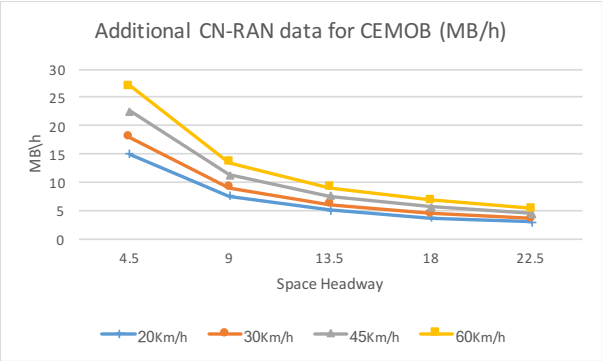

Figure 6: Additional data transfer needed for CEMOB

In Figure 6, we present the additional information needed to be transferred for CEMOB when compared to the baseline in terms of MB/h. The settings of this experiment were the same with the previous one (e.g., number of gNBs, size of an RNA, probability of changing path, etc.). As expected CEMOB always underperforms compared to the baseline, although the amount of information over wireline CN-RAN link seems to be rather manageable from today's networks. This would be the case additional context information is needed. For example, for the worse case of our experiment where vehicles are moving with 60Km/h and the 5000 bits need to be transfered per context transfer, then the overall traffic between CN-RAN would be 225MB/h.

VI. CONCLUSIONS

This paper makes the case that although the specification of 5G networks is well underway and slicing is gradually reaching a mature status several inefficiencies still exist. Standardization activities have sensibly focused on introducing new principles like NF modularization and the support of different numerologies in RAN and ported existing functionalities into the new principles.

What is still missing though are further optimizations, that can be realized if use case specific context information is taken into account. In this paper we have presented a new mobility management scheme that outperforms the baseline for the case of high moving UEs, like the autonomous driven vehicles. By taking advantage of the knowledge of the path that a vehicle will follow and by tailoring cut the involved network functions (e.g., AMF, NEF) appropriately, then significant benefits can be achieved in terms of signalling load with a manageable penalty of additional information being moved inside the network. As a next step of the current work, we will evaluate the proposed scheme using event driven simulations.

REFERENCES

[1] NGMN Alliance, 5G white paper, v 1.0, 2016 available from: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf, access date 12-09-2017.

[2] SE Elayoubi, M. Fallgren, P. Spapis et al., "5G service requirements and operational use cases", European Conference on Networks and Communications - EuCNC 2016, DOI: 10.1109/EuCNC.2016.7561024.

[3] P. Marsch et. al., "5G Radio Access Network Architecture: Design Guidelines and Key Considerations", IEEE Communications Magazine, vol. 54, issue 11, pp 24-32, November 2016.

[4] X. An, et. al., "Architecture Modularisation for Next Generation Mobile Networks", European Conference on Networks and Communications - EuCNC 2017, DOI: 10.1109/EuCNC.2017.7980664.

[5] 3GPP, TS 23.501 "System Architecture for the 5G System; Stage 2 (Release 15)", Version 1.2.0, July 2017.

[6] 5G-PPP Architecture Working Group, "View on 5G Architecture (Version 2.0), July 2017, available at: https://5g-ppp.eu/5g-ppp-revised-architecture-paper-for-public-consultation/, access date 12-09-2017.

[7] 3GPP, TS 22.261, "Service Requirements for the 5G System", V16.0.0, June 2017.

[8] 5G Automotive Association -5GAA, "The case for Cellular V2X for Safety and Cooperative Driving", available at: http://5gaa.org/pdfs/5GAA-whitepaper-23-Nov-2016.pdf, access date 12-09-2017.

[9] 3GPP, TS 23.502, "Procedures for the 5G System", Stage 2 (Release 15), Version 0.6.0, August 2017.

[10] 3GPP, TS 38.300, "NR and NG-RAN Overall Description", Stage 2 (Release 15), September 2017.

[11] 3GPP TR 28.801, "Study on management and orchestration of network slicing for next generation networks", Release 15, Version 1.2.0, May 2017.

[12] K. Chatzikokolakis, A. Kaloxylos, P. Spapis, N. Alonistioti, and C. Zhou, "A survey of location management mechanisms and an evaluation of their applicability for 5G cellular networks", Recent advances in Communications and Networking Technologies, vol. 3, no. 2, 2014.

[13] S. Hailu and M. Säily, "Hybrid paging and location tracking scheme for inactive 5G UEs", European Conference on Networks and Communications - EuCNC 2017, DOI: 10.1109/EuCNC.2017.7980730.

[14] 3GPP TS 23.285, "Architecture enhancements for V2X services", Release 14, March 2017.

[15] R. Trivisonno, R. Guerzoni, I. Vaishnavi, and D. Soldani, "Towards zero latency software defined 5G networks," in IEEE International Conference on Communication Workshop (ICCW), June 2015, pp. 2566–2571.

[16] T. V. Mathew and K. V. Krishna Rao, "Introduction to Tranportation Engineering", Chapter 30, Fundemental parameters of traffic flow, May 2007, available at: http://nptel.ac.in/courses/105101087/downloads/Lec-30.pdf, access date 12-09-2017.

[17] 3GPP, TS 33.401, "Security Architecture", Release 15, Version 15.0.0 June 2017.

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-596-8