Gateway Relocation Avoidance-Aware Network Function Placement in Carrier Cloud

Tarik Taleb
NEC Europe
Heidelberg, Germany
talebtarik@ieee.org

Adlen Ksentini
IRISA - INRIA - University of Rennes 1
Campus Universitaire de Beaulieu
Rennes, France
adlen.ksentini@irisa.fr

ABSTRACT

Building mobile networks, on demand and in an elastic manner, represents a vital solution for mobile operators to cope with the modest Average Revenues per User (ARPU), on one hand, and the ever-increasing mobile data traffic, on the other hand. An important research problem towards this vision of carrier cloud pertains to the development of adequate technologies and methods for the on-demand and dynamic provision of a decentralized and elastic mobile network as a cloud service over a distributed network of cloud-computing data centers, forming a federated cloud. An efficient mobile cloud cannot be built without efficient algorithms for the placement of network functions over this federated cloud. In this vein, this paper argues the need for avoiding or minimizing the frequency of mobility gateway relocations and discusses how this gateway relocation avoidance can be reflected in an efficient network function placement algorithm for the realization of mobile cloud. The proposed scheme is evaluated through computer simulations and encouraging results are obtained.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Distributed networks

Keywords

3GPP, mobile cloud, carrier cloud, network function virtualization, and network function placement

1. INTRODUCTION

There has been continuous need for higher data rates, shorter end-to-end communication delays and short latencies for connection setup. This has pushed for the development of diverse fast radio, backhaul, and mobile core network technologies. This, in return, has favored the launch of smart user equipment, even racing ahead of mobile networks, supporting diverse operating systems and offering both users and developers a wide plethora of tools to generate thousands, if not millions, of mobile applications [1][2]. This wide variety of mobile applications is not only changing the basic assumptions based on which mobile networks have been designed, but is also changing the users' behavior and having a strong impact on them. They further introduce important technological as well as economic challenges to mobile operators, particularly noting the modest average revenues per user (ARPU). Different protocol-level as well as architectural solutions have been proposed to cope with these challenges [3][4]. Most of these solutions present short-term "cosmetic" remedies to an ever-complicating problem. The concept of carrier cloud is perceived as an important long-term solution for mobile operators to cope with the tremendous increase in their mobile data traffic, the so-called mobile IP tsunami, and to get into the cloud computing area, seeking new business opportunities and defining new business models and strategies. Indeed, thanks to their interesting features such as pay-as-you-go, full elasticity support, and multiple tenancy support, cloud computing technologies will help mobile operators to decentralize their networks on demand, elastically, and in the most cost-efficient way, enabling operators to invest small in building virtualized mobile networks and growing on demand as per increase in the mobile data traffic.

As an important enabler of the carrier cloud concept, network function virtualization (NFV) is gaining great momentum among industries. NFV aims for decoupling the software part from the hardware part of a carrier network node, traditionally referring to a dedicated hardware, single service and single-tenant box, and that is using virtual hardware abstraction [5]. Network functions become thus a mere code, runnable on a particular, preferably any, operating system and on top of a dedicated hardware platform. The ultimate objective is to run network functions as software in standard virtual machines (VMs) on top of a virtualization platform in a general purpose multi-service multi-tenant node (e.g., Carrier Grade Blade Server). A suitable Software Defined Networking (SDN) technology can be used to interwork between the different virtualized network functions on the different VMs within the same data center or across multiple data centers, to ultimately realize a flexible, dynamic, rapidly deployable, and elastic mobile network on the cloud. To build an efficient mobile cloud that meets the general requirements of a mobile operator, the placement of network functions, namely the mobile Radio Access Network (RAN) functions, the mobile core network functions, and the caches or servers for the Packet Data Network, is of

MSWiM'13, November 3–8, 2013, Barcelona, Spain.

ACM 978-1-4503-2353-6/13/11 ...$15.00.
http://dx.doi.org/10.1145/2507924.2508000 .

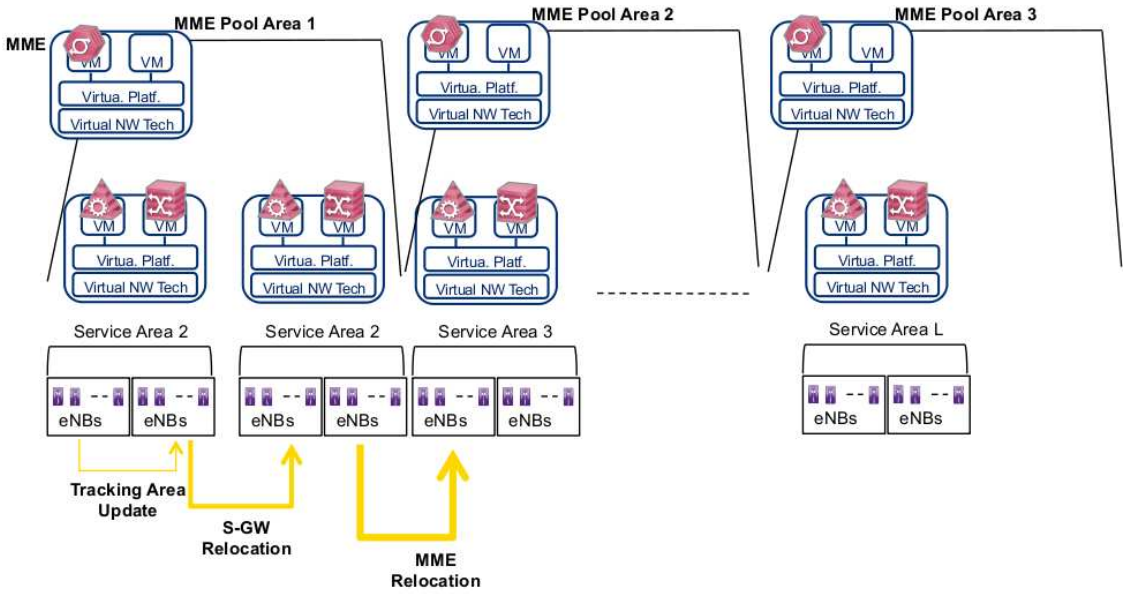

Figure 1: Need for avoiding mobility gateway relocation when placing network functions in carrier cloud.

utmost importance and shall be based on different metrics such as application type [6], data center location, data center load, end-to-end Quality of Service (QoS)/Quality of Experience (QoE), [3] that shall render the overall end-to-end communications optimal [7]. Another important metric for the network function placement consists in the avoidance of gateway relocation, due to the cost incurred with these processes and the impact they may have on the overall QoE [8]. Indeed, with the deployment of mobile network functions on multiple regional data centers, geographically distributed and interconnected, and forming a federated cloud network, the serving areas of mobility gateways (e.g., serving gateways in the context of the Evolved Packet System) and the pool areas of Mobility Management Entities (MMEs) are likely to get smaller and that is for the purpose of localizing mobility management. Frequents handovers with gateway and/or MME relocations may occur as shown in Fig. 1, resulting in additional delay to handoff, additional signaling overhead for bearer establishment and admission control [8][9]. Given these reasons and more, the specifications [9] indicate that gateway and MME relocation during handoff for user equipment in ECM (EPS Connection Management) connected mode (i.e., active mode) is to be avoided when possible. It is therefore the objective of this paper to reflect this recommendation in the planning of service areas and MME pool areas when deciding on where, when and at what scale to place virtualized network functions pertaining to these mobility gateways.

The remainder of this paper is organized in the following fashion. Section 2 presents some research work related to mobile network decentralization and gateway relocation avoidance. The proposed virtualized network function placement scheme, in terms of service areas planning and its modeling, is described in Section 3. Section 4 follows with an evaluation of the proposed scheme. The paper concludes in Section 5 with a summary recapping the main features discussed in this paper.

2. RELATED WORKS

With the introduction of the 3GPP Long Term Evolution (LTE), certain network decentralization features focusing, on both network architecture and network management were brought forward significantly improving the former UMTS (Universal Mobile Telecommunications System). Specifically, considering the network architecture, LTE merged the former Radio Network Controller (RNC) within NodeB introducing a new radio access element called eNB that flattens and simplifies the priori UMTS architecture [9].In terms of network management, LTE advances the prior UMTS-based configuration and optimization methods towards a distributed Self-Organized paradigm [11]. A further step towards an even more flatten UTMS architecture is in [12], where the GGSN (Gateway GPRS Support Node) and SGSN (Serving GPRS Support Node) are integrated in the NBs, while in [13] a similar approach is introduced in the context of EPS with the aim of bringing gateway functions to the edge of the network, merging in this way the radio access and core network split.

In principle, there is a fundamental technology and cost related trade-off behind the adoption of either centralized or distributed network architecture. Costly network equipment are usually shared, creating centralized architectures. This was the case in the initial phase of 3G deployment, wherein a centralized architecture was preferred to share core network utilities and processing resources, while keeping the base station cost low [12]. Nowadays the evolution of computer technology has significantly reduced equipment costs, advancing their deployment flexibility. However, the increased mobile-oriented data volumes still maintains the network utilization cost high, creating significant revenue problems for operators. For this reason operators are looking for data offloading solutions towards the network edge, introducing PDN-GW/S-GW (Packet Data Network Gateway and Serving Gateway) functionality close to eNBs (evolved Node B). Additionally, the provision of Content Distributed Networks (CDN) services towards the network edge may complement

data offloading, reducing further network costs as within the backhaul and via a more efficient resource usage.
A tutorial regarding the data offloading techniques including LIPA (Local IP Access), SIPTO (Selective IP Traffic Offload) and IFOM (IP Flow Mobility and Seamless Offload) focusing on 3GPP Rel-10 is available in [14], while further details on LIPA/SIPTO data offloading are provided in [4], which illustrates specific network architectures and service requirements that meet the current decentralized needs, enlightening also network management and deployment issues with emphasis on QoS and service continuity. A complementary analysis on the architecture and main benefits associated with the use of decentralized architectures from the IETF perspective is documented in [15], with the most important being route optimization and increased robustness. Currently the research efforts towards decentralized cellular networks focus on mobility and service continuity as well as on efficient resource management related to gateway selection and relocation. Ensuring service continuity and QoE for active user with reduced signaling, by avoiding the use of core network equipment, is the ultimate goal of decentralized mobile networks. Nevertheless, the means of achieving decentralization is distributed mobility management. A study that examines and compares centralized and distributed mobility in the context of 3GPP is presented in [16], elaborating also a dynamic and distributed mobility management solution, which merges the mobility anchor gateways and base stations. In particular, mobility is realized by the use of tunneling to forward traffic upon a handover, allowing also the user to establish flows via different mobility anchors for efficient resource usage. The fundamental concepts of gateway-based load balancing with respect to user performance are analyzed in [17], where an inter-GW load balancing protocol and policies are presented and evaluated considering a gateway pool in association with group of base stations. A more advanced solution to provide service continuity focusing on distributing the content of a centralized mobility anchor to a set of distributed mobility agents utilizing the concept of virtual routers and Distributed Hash Tables (DHT) is introduced in [18], with substantial benefits in load balancing and resiliency.
For highly mobile users, i.e. on board vehicles, decentralized mobile networks would also have a significant impact on the conventional Tracking Area Update (TAU) procedure and on the maintenance of PDN connections of UEs in idle mode. Although TAU approaches as in [10] aim to allocate to UEs a gateway that could potentially serve them for a large geographical area, considering the initially associated eNB position, this approach mostly holds for centralized architectures. In decentralized schemes, we envision smaller serving areas, thus additional mechanisms to compensate frequent serving area relocations are recommended. In [19], a self-organized method, which adopts the tracking areas according to the user mobility taking into account long term history data is introduced based on graph partitioning heuristics.

3. GATEWAY RELOCATION AVOIDANCE-AWARE NETWORK FUNCTION PLACEMENT SCHEME

3.1 Problem statement

As shown in Fig.1, each service area is managed by one virtual S-GW and consists of several Tracking Areas (TAs). Each TA consists, in turn, of several cells. In order to reduce the S-GW relocation it is important to ensure an optimal planning of the service areas considering the mobility patterns of users as well as their data traffic load. Furthermore, it is important to dimension the virtual function by defining the optimal number of virtual S-GWs to create. It shall be noted that throught the remainder of this paper, the focus is on service areas, but the same logic applies to pool areas of MMEs.
Let N denotes the number of Tracking Areas as planned by any underlying mobile network planning algorithm [21]. The set of TAs in the mobile network is denoted by $N_{TA} = \{1, \ldots, N\}$, and the set of SAs currently employed is $N_{SA} = \{1, \ldots, S\}$. The vector $SA = [s_1, \ldots, s_N]$ denotes the service area planning (i.e., mapping TA to SA), whereby s_i is the Service Area of TA i. This mapping can be represented by a binary symmetric matrix (NxN), denoted by $adj(SA)$. An element $adj(SA)$ represents the case whether or not two TAs are in the same SA, in other words

$$adj_{i,j}(SA) = \begin{cases} 1 & \text{if } s_i = s_j \\ 0 & \text{Otherwise} \end{cases}$$

Let $w(i)$ denotes the traffic load at a TA i, during a specific period of time. Intuitively, $w(i)$ equals to the aggregate load of traffic exchanged over cells belonging to TA i. Let $h_{i,j}$ denote the number of UEs moving from TA_i to TA_j during the same period of time. For the sake of simplicity, we assume that each time a UE moves from one TA into another TA, a S-GW relocation occurs incurring a cost denoted by $C_{relocation}$. Intuitively, in real life networks, a UE has to cross a number of TAs before a S-GW relocation takes place. Moreover, we assume that each S-GW has the capacity to handle SGW_{max} amount of traffic load generated by the covered service area.
At this point the problem of service area planning, i.e., placement of virtualized S-GW functions on a federated cloud or a distributed network of data centers, can be formulated as follows:

$$min \sum_{i \in N} \sum_{j \in N} Cost = C_{relocation} h_{i,j}(1 - adj_{i,j}(SA))$$

Subject to
$$for\ i = 1\ to\ N,\ \sum_{j \in N} w(j) adj_{i,j}(SA) \leq SGW_{max}$$
$$adj_{i,j}(SA) \in \{0, 1\}$$

In other words, the optimization problem consists in finding the optimal service area planning, and therefore the location of data centers where to place the virtualized S-GW functions, that minimizes the overall S-GW relocation cost, while ensuring that each SA traffic load does not exceed the capacity SGW_{max} of a SGW. Accordingly, the problem becomes an Integer Linear Program (ILP). This problem is similar to the TA planning algorithm, which is known as *NP* hard [21]. Solving optimality the service area planning problem may require excessive computational effort in view of its complexity. In the following section, we propose an effective solution based on a greedy algorithm, which finds an effective solution for this problem.

3.2 Greedy-based algorithm for virtualized SGW placement in carrier cloud

The proposed greedy algorithm (GA) successively builds up a solution for the service area planning. It selects step

Algorithm 1 Greedy algorithm for virtualized S-GW placement in carrier cloud

```
1: procedure CHECK_NEIGHBOR(k)
2:     Neighbor(k) ← ∃m ∈ SA : adj_{k,m}(SA) = 1;
3:     for all m ∈ SA : adj_{k,m}(SA) do
4:         if w(m) + load_s ≤ SGW_max then
5:             SA(m) ← S
6:             load_s ← w(m) + load_s;
7:         end if
8:     end for
9: end procedure
10: procedure MAIN
11:     SA ← {0, 0, 0, ..., 0}
12:     S ← 0
13:     while ∃i ∈ SA : adj_{i,j}(SA) do
14:         S ← S + 1
15:         Neighbor(j) ← ∃j ∈ SA : adj_{i,j}(SA) = 1;
16:         load_s ← w(i);
17:         Check_Neighbor(i);
18:         for all j ∈ Neighbor(i) do
19:             Check_Neighbor(j);
20:         end for
21:     end while
22: end procedure
```

by step the TAs assigned to a SA. This is repeated until the S-GW capacity is exceeded.

In the first run, the greedy algorithm selects a TA (i.e., starting TA noted TA_i) from the SA set. Then, it affects this TA to a SA. For each neighbor of TA_i (noted TA_j), it checks if it can be assigned to the same SA (S). A TA is assigned to S if the traffic load generated by this TA added to the current traffic load of (S) does not exceed the capacity SGW_{max}. If TA_j is assigned to the same SA, the GA continues to consider candidates TA from the neighbor set of TA_j. Since there is no more candidate in the neighbor of TA_j, the operation is repeated for another neighbor of the TA i. When there is no possibility to add a further TA to S, the greedy algorithm selects then another TA from the SA set (i.e., the selected TA is not yet assigned to any SA) and increments S. The algorithm ends when all TAs are assigned to a specific SA (S). Note that the proposed greedy algorithm is not efficient, as it does not take into consideration the cost. In order to address this issue, we propose another version of this algorithm, namely Repeated Greedy Algorithm (RGA), which improves the GA version by adding a loop on the initial TA for the GA. So, RGA respectively executes GA with different starting TA. The final Serving Area configuration or (SA_i) is the one that incurs the minimum cost, i.e. the minimum number of S-GW relocation.

It is worth noting that RGA gives a near optimal number of virtual S-GWs to be instantiated and deployed for a certain configuration of traffic load and mobility pattern, while reducing the number of S-GW relocations. The algorithm complexity is $O(N^3)$.

On the other hand, the RGA algorithm can be launched periodically, or if the mobile operator detects a noticeable change in the network traffic load or user mobility, so it can scale up or down the number of virtual S-GWs given some network load conditions.

Table 1: Simulation parameters.

	$w(i)/min$	$h_{i,j}/min$	SGW_{max}
Scenario1	Random [0, 25]	Random [0, 50]	200
Scenario 2	Random [0, 25]	Random [0, 25]	200
Scenario 3	Random [0, 50]	Random [0, 50]	200
Scenario 4	Random [0, 50]	Random [0, 25]	200

4. SIMULATION RESULTS

In this section, we present the results of our experiments conducted using MATLAB to evaluate the efficiency of RGA in gateway relocation avoidance. We compare the GRA results against those of GA (i.e., only one TA is used for initiating the greedy algorithm). We generated a random network topology, where each TA has an average of six neighbors. We considered four scenarios as described in Table 1. Scenario 1 represents the case of low traffic load and high mobility. Scenario 2 also shows the case of low traffic load, but considers users with low mobility features. Scenario 3 represents the case of high traffic load and high mobility features. Finally, scenario 4 illustrates the case of high traffic load and low mobility features. We varied the number of TAs for each scenario from 10, representing a small area, to 80, simulating a large area.

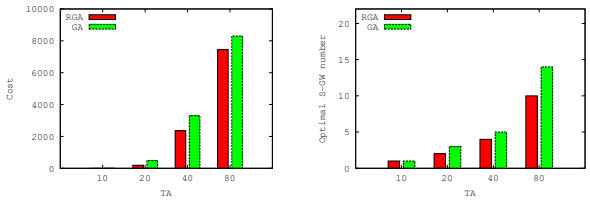

Figure 2: Scenario 1

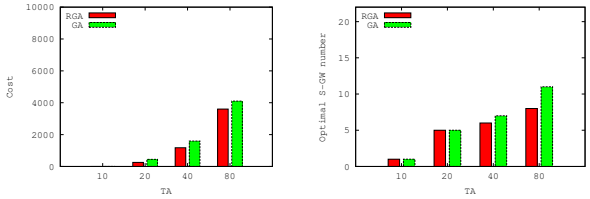

Figure 3: Scenario 2

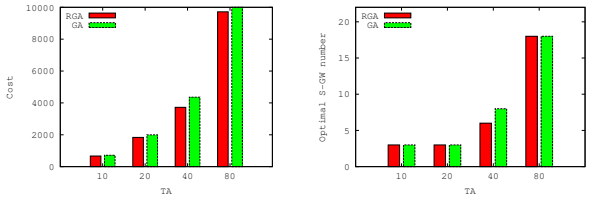

Figure 4: Scenario 3

Figs. 2 - 5 show the results obtained for the four simulated scenarios. The results are presented in terms of incurred cost and the required number of virtual S-GWs. From these results, we clearly observe that high traffic load implies the need for higher number of S-GWs, and when the mobility is high, the cost (S-GW relocation) incurred by both the GRA

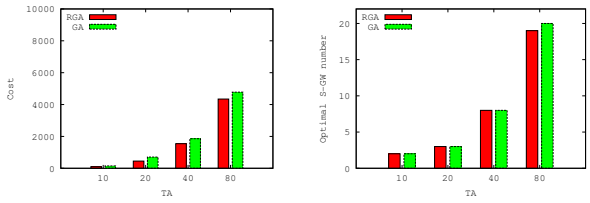

Figure 5: Scenario 4

and GA is high. For instance, in case of low traffic load (i.e., scenarios 1 and 2), the required number of S-GWs is one when TA=10, and this number exceeds 20 in case of scenario 4 (i.e., high traffic load) for TA=80. We also observe that the cost is highly skewed between scenarios 1 and 2, even they have the same traffic load. Scenario 1 incurs higher cost than scenario 2, which is intuitive as there are frequent handovers with TA change, which increases the probability of S-GW relocation. We notice the same behavior when comparing between scenarios 2 and 3. Finally, we remark that RGA always finds a better solution than GA, ensuring lower cost in terms of gateway relocations and involving less number of virtual S-GWs to instantiate.

Acknowledgment

The research work presented in this paper is conducted as part of the Mobile Cloud Networking project, funded from the European Union Seventh Framework Program under grant agreement number [318109].

5. CONCLUSION

In this paper, we presented a new algorithm for avoiding or minimizing the frequency of mobility gateway relocations in carrier cloud and that is via optimal placement of virtualized relevant network functions over federated clouds. We focused on the case of S-GW relocation in case of EPS, and formulated the network function placement as a service area planning optimization problem, where the aim is to reduce the cost of gateway relocation. This problem is NP-hard problem. We therefore introduced a heuristic based on a greedy algorithm. Simulation results indicate the efficiently of the proposed algorithm in reducing the gateway relocation cost.

6. REFERENCES

[1] T. Taleb and A. Ksentini, "Impact of Emerging Social Media Applications on Mobile Networks," in Proc. IEEE ICC 2013, Budapest, Hungary, Jun. 2013.

[2] T. Taleb and A. Kunz, "Machine Type Communications in 3GPP Networks: Potential, Challenges, and Solutions," in IEEE Commun. Mag., Vol. 50, No. 3, Mar. 2012.

[3] T. Taleb and A. Ksentini, "QoS/QoE Predictions-based Admission Control for Femto Communications," in Proc. IEEE ICC 2012, Ottawa, Canada, Jun. 2012.

[4] K. Samdanis, T. Taleb, and S. Schmid, "Traffic Offload Enhancements for eUTRAN", in IEEE Communications Surveys and Tutorials J., Vol. 11, No. 3, Aug. 2012, pp. 884-896.

[5] Authored by network operators, "Network Functions Virtualization: An Introduction, Benefits, Enablers, Challenges, & Call for Action," Oct. 2012

[6] T. Taleb and A. Ksentini, "On Efficient Data Anchor Point Selection in Distributed Mobile Networks," in Proc. IEEE ICC 2013, Budapest, Hungary, Jun. 2013.

[7] T. Taleb and A. Ksentini, "Follow Me Cloud: Interworking Federated Clouds & Distributed Mobile Networks," to appear in IEEE Network Magazine.

[8] T. Taleb, K. Samdanis, and F. Filali, "Towards Supporting Highly Mobile Nodes in Decentralized Mobile Operator Networks," in Proc. IEEE ICC 2012, Ottawa, Canada, Jun. 2012.

[9] 3GPP, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," TS 23.401, V10.2.1, Jan. 2011.

[10] A. Kunz, T. Taleb, and S. Schmid, "On Minimizing SGW/MME Relocations in LTE", in Proc. ACM IWCMC'10, Caen, Jun. 2010.

[11] 3GPP, "Technical Specification Group Radio Access Network, Evolved universal Terrestrial Radio Access Network (E-URAN), Self-configuring and self-optimizing network (SON) use cases and solutions", (Rel 9), TR 36.902, Apr. 2011.

[12] P. Bosch, L. Samuel, S. Mullender, P. Polakos, and G. Rittenhouse, "Flat Cellular (UMTS) Networks", IEEE WCNC Hong Kong, Mar 2007.

[13] Z. Yan, L. Lei, and M. Chen, "WIISE A Completely Flat and Distributed Architecture for Future Wireless Communication Systems", WWRF 21, Stockholm, Oct 2008.

[14] C.B. Sankaran, "Data Offloading Techniques in 3GPP Rel-10 Networks: A Tutorial", IEEE Communication Magazine, Vol.50, No.6, Jun 2012.

[15] R. Kuntz, D. Sudhakar, R. Wakikawa, and L. Zhang, "A Summary of Distributed Mobility Management", IETF Internet Draft, Feb 2011.

[16] P. Bertin, S. Bonjour, and J-M. Bonnin, "Distributed or Centralized Mobility?", IEEE GLOBECOM, Honolulu, Dec 2009.

[17] C. Xue, J. Luo, R. Halfmann, E. Schulz, and C. Hartmann, "Inter GW Load Balancing for Next Generation Mobile Networks with Flat Architecture", IEEE VTC-Spring, Barcelona, May 2009.

[18] M. Fischer, F-U. Andersen, A. Kopsel, G. Schafer, and M. Schlager, "A Distributed IP Mobility Approach for 3G SAE", IEEE 19th PIMRC, Cannes, Sep 2008.

[19] M. Toril, "Automatic Re-planning of Tracking Areas", SOCRATES Final Workshop on Self-Organisation in Mobile Networks, Karlsruhe, 22 Feb 2011.

[20] S. Pack, T. Kwon, and Y. Choi, "A Performance Comparison of Mobility Anchor Point Selection Schemes in Hierarchical Mobile IPv6 Networks", in Elsevier J. Computer Networks, Vol. 51, No. 6, Apr. 2007.

[21] Y. Bejerano, M-A. Smith, J. Naor, and N. Immorlica. "Efficient Location Area Planning for Personal Communication Systems", in IEEE/ACM Transactions on Networking, Vol. 14, No. 2, Apr. 2006.