On Minimizing Serving GW/MME Relocations in LTE

Andreas Kunz, Tarik Taleb, Stefan Schmid

NEC Europe, Ltd.

Heidelberg, Germany

{kunz, taleb, schmid}@neclab.eu

ABSTRACT

In the System Architecture Evolution (SAE) study of the next-generation mobile network in 3GPP, Serving Gateways (SGWs) and Mobility Management Entities (MMEs) are grouped to form a number of service and pool areas, respectively. While this concept of SGW service areas in the Evolved Packet Core is interesting to limit the administrative scope of SGWs and also provides a means to optimize the routing, the use of fixed/hard area boundaries can result in frequent unnecessary SGW relocations and can severely impact the Quality of Experience (QoE) of users. To avoid the drawback of fix/hard service (or pool) area boundaries, this paper proposes a scheme whereby every SGW can have a flexibly configurable service area, which is defined by a set of LTE (Long Term Evolution) cells or Tracking Areas (TAs). The service area of a SGW defines the LTE area (e.g., cells or TAs) that the SGW can serve. The working of the proposed mechanism is validated via computer simulations and encouraging results are obtained.

Categories and Subject Descriptors

C.2.1 [**Computer Communication Networks**]: Network Architecture and Design, Internetworking, Mobile Networking.

General Terms

Management, Performance, and Standardization.

Keywords

MME Pool, 3GPP networks, LTE, Serving gateway relocation, MME relocation, handover, EPC, EPS.

1. INTRODUCTION

3GPP's Evolved UMTS Terrestrial Radio Access Network (EUTRAN), also known as LTE, along with the Evolved Packet Core (EPC), form the basis for the Evolved Packet System (EPS). The main benefits of LTE networks, compared to their 3G counterparts, consist in data throughput maximization (i.e., via usage of cutting-edge radio access and antenna technologies such as OFDMA and MIMO) and reduced latency (i.e., via a reduction in the number of network nodes involved in data processing and transport, and separation between user data and control planes) [1][2]. Indeed, concerning the latter, in contrast to the deployment of four nodes in the user plane (i.e., GGSN, SGSN, RNC, and NodeB) in WCDMA/HSPA networks, EPS defines a flat network architecture consisting of LTE base stations (i.e., eNodeBs), Mobility Management Entities (MMEs) as well as Serving Gateways (SGW) and Packet Data Network Gateways (PDN GWs) in the user plane. In the EPS architecture, an eNodeB is connected to more than one SGWs, following a many-to-many mapping style as shown in Figure 1, in order to enhance system scalability.

MMEs and serving gateways are grouped in a number of pool or service areas, respectively. Each pool or service area is served by one or more MMEs/SGWs in parallel. These pool areas are typically a collection of complete Tracking Areas (TAs) and may overlap each other [3][4]. With the concept of pool areas, two handover types can be envisioned:

- Cell handover: it takes place when a UE moves from a source base station (BS) to a target BS that is controlled by the same pool of gateways as the source BS.
- SGW relocation: it occurs when the target BS is controlled by a pool of gateways different than that of the source BS.

The pool of gateway concept primarily aims for reducing the frequency of SGW relocation as delays associated with such handovers are typically longer than those of normal cell handovers. Such long handover latencies may lead to significant drops of in-flight packets; a fact that may ultimately impact the quality of experience (QoE). Another important objective of pool of gateways is to introduce redundancy that shall enhance system reliability in case of overload or hardware failure of individual gateways. The main objective of this paper is to further reduce the frequency of SGW relocation.

Whilst the same explanation applies to MME and MME pool areas, the focus in this paper will be on SGWs and their service areas. The functions of the SGW include:

- local Mobility Anchor point for inter-eNodeB handover;
- mobility anchoring for inter-3GPP mobility;
- downlink packet buffering and initiation of network triggered service request procedure;
- lawful interception;
- packet routing and forwarding;
- transport level packet marking in the uplink and the downlink, e.g., setting the DiffServ Code Point based on the QCI of the associated EPS bearer; and
- accounting for inter-operator charging.

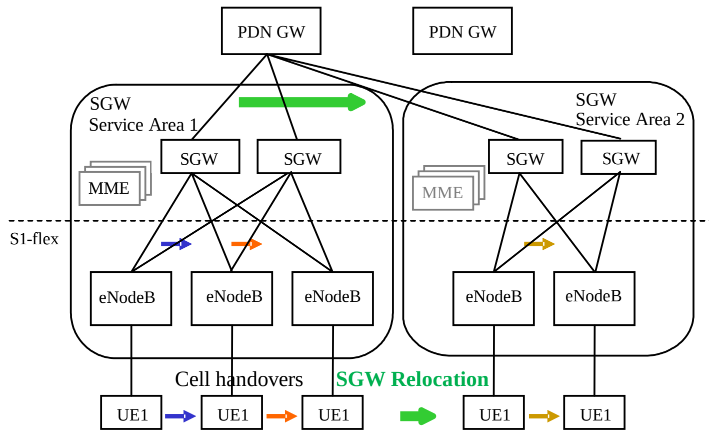

Figure 1. SGW service or pool areas – with fix/hard SGW relocation boarders at service or pool area boundaries.

The S1-flex interface allows eNBs to connect "flexibly" to any SGW as long as the eNBs belong to the administrative scope of the SGW (Figure 1). A UE, camping in LTE, has only one SGW at any point in time and SGW relocation typically takes place while UEs are in idle mode (i.e. ECM-IDLE). SGW relocation for UEs in active mode (i.e. ECM-CONNECTED) must be supported (e.g., due to inter-PLMN), but should be avoided as much as possible to circumvent the need for a complex and high-performance relocation procedure. The following questions regarding SGW selection and relocation are still open:

- How to organize SGWs and their administrative scope in a scalable way (i.e., in a way that the administrative scope of the SGW can be smaller than the whole PLMN)?
- How to reduce the number of SGW relocations (i.e., despite limiting the administrative scope of SGWs)?

As illustrated in Figure 1, the problem of SGW relocation arises when UEs move across service areas – i.e., when a UE hands off to an eNB that is outside the administrative scope of the current serving SGW; relocation to one of the SGWs of the target Service Area is required. While the concept of service areas aims for limiting the administrative scope of SGWs and also provides a means to optimize routing, the use of fixed area boundaries can constitute a drawback. For example, users that frequently cross the service area boundary (e.g., because of their living space) will experience SGW relocation every time they cross. For idle mode UEs, the drawback is mainly that extra signaling load is introduced every time the UE crosses between service areas, which could also be avoided if the service area boundaries are not rigid. However, for active mode UEs, the drawback is likely to be more severe as service degradation might be experienced during active communications. As such, frequent SGW relocation for UEs in active mode should be avoided by all means. To avoid the drawback of fix/hard service area boundaries, this paper proposes a scheme whereby every SGW can have a flexibly configurable service area, which is defined by a set of LTE cells or Tracking Areas. The service area of a SGW defines the LTE area (e.g., cells or TAs) that the SGW can serve.

The remainder of this paper is organized as follows. Section II puts the paper in the context of its state of the art. The main idea of the paper is described in Section III. Section IV presents and discusses simulation results. The paper concludes in Section V.

2. Related Work

In the sphere of attempts to reduce the frequency of IP handovers, a large body of prior work was proposed. The central theme in these pioneering studies pertains to the adoption of hierarchical management strategies using local agents. Hierarchical Mobile IPv6 (HMIPv6) [5] is a notable example. Most proposed protocols employ hierarchies to localize the binding traffic. Determination of the optimal size of local networks is one of the most challenging tasks in hierarchical management procedures. To deal with this task, Xie et al. propose an analytic model based on the average total location update and packet delivery cost [6]. In [7] and [8], decision of the optimal size of regional networks is based on mobility patterns, registration delays, and the CPU processing overhead loaded on the local mobility agents. While most hierarchical techniques are intended to reduce the BU traffic by localizing handoff signaling, they cause additional issues related to network traffic management. Effectively, some local agents get congested with traffic while others are not efficiently utilized. To overcome this deficiency, the choice of network hierarchies should be performed in a dynamic manner [9], [10]. However, one major drawback of the available schemes is that they both deliver packets to users via multiple levels of nodes in RAN, a fact that leads to long packet delivery delay and congestion of the selected RAN nodes with redundant traffic. One possible solution to this issue is to reduce the size of the subnet domains. However, this would lead to frequent inter-domain IP handovers and consequently excessive signaling.

Another approach to solve the issue of frequent IP handovers in hierarchical management procedures is possible by referring to the mobility pattern of users [11], [12]. In [11] for instance, users are classified based on their velocity. Users receive thresholds from the network and compare their velocity to those thresholds. Users with velocities exceeding the propagated thresholds simply register with higher levels of the hierarchies. While this idea is straightforward, it still does not solve the issues of traffic distribution among mobility anchors. Indeed, in case all users have the same feature of mobility, they end up by registering with the same mobility anchors (e.g., SGWs). This will intuitively overload the selected mobility anchors with traffic whereas other mobility anchors remain underutilized. Additionally, the velocity range for each mobility anchor is fixed. To cope with this issue, Chung et. al. [13] considered a dynamic setting of the velocity range of each mobility anchor depending on the actual velocities of MNs which are currently serviced by the mobility anchor. However, a general requirement for mobility management schemes that are based on the velocity of mobile nodes, consists in the guarantee of a high accuracy in the estimation of the velocity of mobile nodes. Such a task is not always simple, resulting, more frequently, in the selection of inappropriate anchors. In [14], the moving range of a mobile node is the main factor in the mobility anchor selection. In this scheme, mobile nodes are assumed to keep track of their moving area. The lowest mobility anchor that covers the entire moving area is deemed to be the most appropriate one for registration. In this scheme, issues related to how to define the moving range of each mobile node, in addition to how the scheme can be applied to mobile nodes that keep changing their moving areas, are yet to be solved. In [15], a newly-defined factor, dubbed session to mobility ratio (SMR), is used as a factor for the selection of the serving mobility anchors. SMR is defined as the ratio of the session arrival rate to the handover frequency. In the SMR-based scheme, the highest mobility anchor is selected for mobile nodes with small values of SMR. An interesting analysis among the above-mentioned approaches can be found in [16].

Although nowadays' user data traffic is not yet significant to overload mobile gateways, it is expected that the ever-growing community of mobile users along with their emerging bandwidth-intensive applications will form such a challenge. Consequently, not only UE's velocities, but also loads consist an important factor in the selection of anchor points. With this regard, the authors proposed in [17] a new approach, called Dynamic and Efficient Mobility Anchor Point Selection (DEMAPS). The proposed scheme works similar to the underlying mobility management procedure when the network is not overloaded. When the network becomes under heavy loads, the selection of mobility anchors becomes based on an estimation of anchor load transition using the Exponential Moving Average (EMA) method. In the context of LTE, the work in [18] presents an inter-GW load balancing protocol that triggers IP handover based on the load state of gateways. In the proposed scheme, gateways in a particular service (or pool) area periodically and mutually exchange load information in a fashion that a gateway is always aware of the load status of its neighboring gateways. When a gateway is about to get congested, it inquires its less congested neighbors if they can accommodate some of the ongoing sessions. In case of an informative response, gateway handovers take place for some selected UEs.

3. Controlling SGW/MME Relocation Triggering

Following the concept of flexibly configurable service areas per SGW, operators have full flexibility in the way they want to define/configure the service area of a particular SGW. Service areas can be as large as the whole PLMN (Public Land Mobile Network), but could also be limited to a certain regional area (e.g., city, metropolitan area, state) or otherwise defined geographic area (e.g., along a train line or motorway). An important point is that SGW service areas are typically heavily overlapping so that individual TAs/cells are in most cases covered by several service areas, which allows the system to choose a SGW that is expected to best service a given UE.

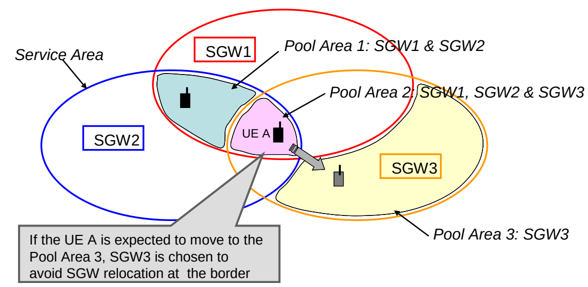

Figure 2. Service areas of neighboring SGWs overlap creating implicit "pool areas".

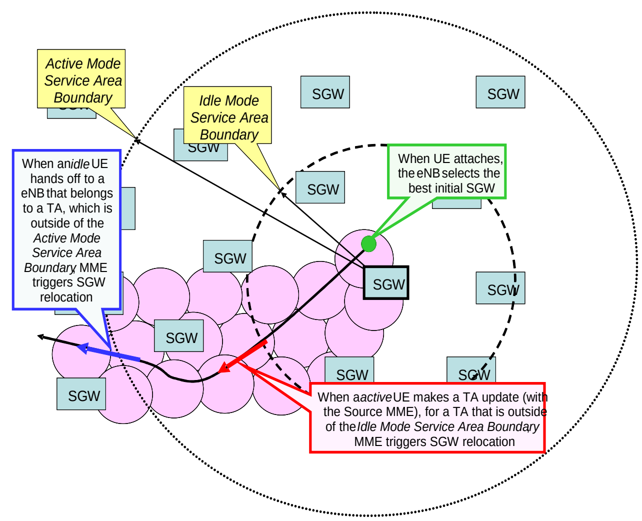

Figure 3. Per-SGW Idle and Active Mode Service Areas.

SGWs that are able to service a certain LTE cell/TA can be considered as the "pool" of SGWs that is available to handle the UEs while camping in that LTE cell/TA. This is illustrated in Figure 2. For example, UE A can be served by SGW1, SGW2 or SGW3 while camping in "pool area" 2, as it is located in the service areas of all three SGWs. However, since the SGW service area concept allows a single SGW to be part of many "pool areas", the crossing of a "pool area" boarder typically does not require a SGW relocation, as the serving SGW (if selected well) is likely able to also serve in the new "pool area".

Since SGW relocation for UEs in active mode should be avoided by all means (i.e., to avoid unnecessary service disruption), relocation should only occur when the UE moves outside the administrative scope of its serving gateway. In idle mode, it is not trivial to know when to relocate SGWs for UEs moving across SGW service areas. Even in this case, it is highly desirable to minimize such SGW relocations (i.e., only when there is sufficient gain) to avoid unnecessary signaling load in the network. For example, SGW relocation after every TA update is generally not desirable and may also lead to unexpected oscillation effects. On the other hand, avoiding SGW relocation for UEs in idle mode until the administrative bounds are reached is also not ideal. Indeed, an earlier SGW relocation (while a UE is still in idle mode) may help to avoid a later relocation when the UE becomes active. Moreover, idle mode SGW relocation can also be desirable in case a UE moves a significant distance from the serving SGW as a more optimal SGW (e.g., less load, geographical proximity) may become available. To conclude, it is desirable to also provide a mechanism that allows for flexible configuration of SGW service areas and supports efficient SGW relocation for UEs in idle mode.

As a solution, this paper introduces the concept of Idle Mode Service Area for SGWs, which define new boundaries where SGW relocation may take place for UEs in idle mode. Effectively, when an idle UE moves beyond the Idle Mode Service Area of the serving SGW, a new SGW will be selected. Figure 3 depicts the concept of active mode and idle mode service area boundaries.

Depending on the UE state (i.e., idle or active), SGW relocation will be triggered at different points when the UE is moving. For idle UEs, the relocation procedure will be activated when the UE leaves the Idle Mode Service Area of the serving SGW. For active UEs, crossing this boundary has no impact. Only when moving beyond the Active Mode Service Area Boundary, SGW relocation will take place. In case a UE crosses the Idle Mode Service Area Boundary while it is in active mode, but changes to idle afterwards, then the UE state change will trigger the SGW relocation (since the UE is no longer inside the Idle Mode Service Area). Recall that the main rational beneath this enhancement is that an earlier idle mode relocation can avoid the need for a later active mode relocations, which is usually more costly in terms of different factors. With such two boundary types of service areas, operators are able to flexibly control when SGW relocation should take place for active and idle mode UEs, on a per SGW basis.

4. Simulation Results and Discussion

In order to emulate the SGW relocation for a moving UE in different states (idle or active), a MatLab simulator was developed to simulate the scenario described in the previous section. The simulator consists of three main parts, namely the input variables handling, the network emulation service and the statistical output computation. The input variables handling entity is responsible for batch simulation runs with varying parameter settings. The main functions of the simulator are in the network emulation, which is responsible for the correct simulation of the network and contains two functionalities. One is the traffic model, which includes the UE state handling and the ON/OFF path length computation, and the other one is the handover handling, which checks whether the UE is still in the idle or active area of the SGW in dependency with the traffic model state of the UE. In case of handover decision, the SGW-Reselection module is called to select a new serving SGW.

The basic concept of the network topology consists of a uniformly distributed matrix of SGWs with a constant Inter-SGW distance (*UId*) in both *x*- and *y*-directions. A UE is initially placed at a randomly selected position (*Sx*, *Sy*) into the network. It then moves with a constant velocity *v* in direction of the x-axis for a certain length *l*. While the UE is moving, its state can change according to the voice traffic model. Depending on the simulation mode, the start position of the UE, its speed, its moving distance and its generated traffic can vary. With no specific purposes in mind, the UE is not simulated to change its moving direction in the considered mobility model: it always keeps moving along the x-axis.

The traffic model of the UE consists of two states:

- ON State: the UE is transmitting data/talking.
- OFF State: the UE is idle.

Both states are controlled by a negative exponential Markov process with a PDF $f(x) = \mu \cdot e^{-\mu \cdot t}$. The variable μ is split into the variables μ_{ON}, which describes the average talking time, i.e., the duration of one call and the variable μ_{OFF}, the average silent time between two calls. Therefore the UE stays with a probability $P_{ON}(t)$ in ON state and switches to OFF state with a probability $(1 - P_{ON}(t))$. The same is valid for the OFF state, i.e., the UE stays in OFF state with a probability $P_{OFF}(t)$ and leaves the OFF state to the ON state with a probability $(1 - P_{OFF}(t))$. In the simulations, it is useful to know at which place the UE is in ON or OFF state, while being on the move. Therefore, the time-dependency of the calls is mapped to the location-dependency of the UE. This mapping is achieved using the velocity of the UE. Table 1 lists up settings of some of the considered scenarios.

Table 1 Settings of simulated scenarios.

Scenario	Distance	Distribution Function	Speed (km/h)	Distribution Function	Probability
Walking	50m-10Km	exponential	2-8	uniform	35%
Short/Slow Run (Car/Tram/Bike /Bus)	500m-10Km	uniform	20-50	uniform	20%
Short Drive (Car/Train/Bus)	10-20km	uniform	30-80	uniform	20%
Medium Drive (Car/Train/Bus)	20-80km	uniform	20-100	Gaussian	20%
Long Drive (Car/Train/Bus)	100-1000km	exponential	20-300	Gaussian	5%

As comparison term, we compare our approach against the legacy concept whereby there are only active mode service areas. Our approach is simulated as follows. In the "Idle/Active mode" model, the UE is served by its selected SGW until it leaves the "idle mode" radius Ri. This is in case its traffic model is also in "OFF-state". If the traffic model is in "ON-state" (i.e., the UE is active), it is served by the selected SGW until it leaves the "active mode" radius Ra. For the special scenario whereby the UE changes its traffic model state from "ON" to "OFF" while moving in the area Ri < x < Ra, it has to carry a SGW handover to select a new valid "idle mode" area.

To sum up, the SGW-reselection algorithm is called when the UE leaves either its idle radius Ri in OFF state or its active radius Ra in ON state or when it switches from ON to OFF in the area between Ri and Ra or at the beginning, when the UE is placed in the network. In these cases, the UE has to select a (new) serving SGW.

To evaluate the performance of the proposed scheme in reducing the number of forced active mode relocations by UEs that reach the service area border, we run a number of simulations. The results are presented in Figures 4 and 5. The results are based on the configurations as shown in Table 2.

The figures compare the average number of handovers experienced by the UE when the additional idle mode service area is adopted and when it is not. The users are moving around with different "walking lengths" and at different speeds, simulating different categories of users (e.g. pedestrians or the "long drive" of ~200 km by train or car – Table 1). Figures 4 and 5 plot the number of handoffs (averaged over the total simulation runs) that were carried out when the UE was in idle mode and in active mode, respectively, and that is under the five scenarios of Table 1.

Table 2 Network configuration.

Parameter	Value
Ra Radius of active mode service area	100 Km
Ri Radius of idle mode service area	65 Km
UId Distance between uniformly distributed SGWs	100 Km
Average on time	300 Sec
Average off time	2000 Sec

The results clearly indicate that the proposed optimization reduces the number of forced active mode SGW relocations at least in the order of factor 10 compared to the legacy approach for the long drive, the "worst case" scenario. On the other hand, the number of additional idle mode relocations increases only by 30% in this scenario. For scenarios 1-4, our optimized solution even manages to completely avoid any active mode relocations, whereas in case of the conventional method there is always some low, but constant probability of active mode relocations (e.g., 13% for the common medium drive ~ 50km scenario). Considering the fact that it is likely that in these scenarios the same users often experience the active mode relocation, even a small probability can lead to an extremely poor user experience.

5. Conclusion

In the Evolved Packet Core, grouping SGWs in particular service areas that overlap is an interesting concept that solves many issues regarding route optimization and service reliability. However, it does not come without its shortcomings. Indeed, defining rigid boundaries for these service areas results in frequent forced SGW relocations. Whilst such forced SGW relocations are still acceptable for UEs in idle mode, they should be avoided for UEs in active mode. Otherwise, the QoE of users may largely get affected.

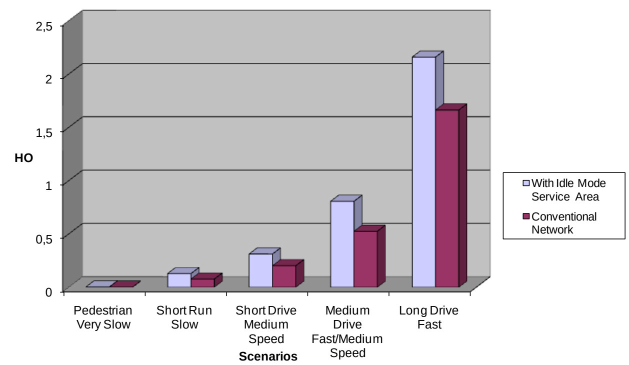

Figure 4. Number of handoffs experienced during idle mode.

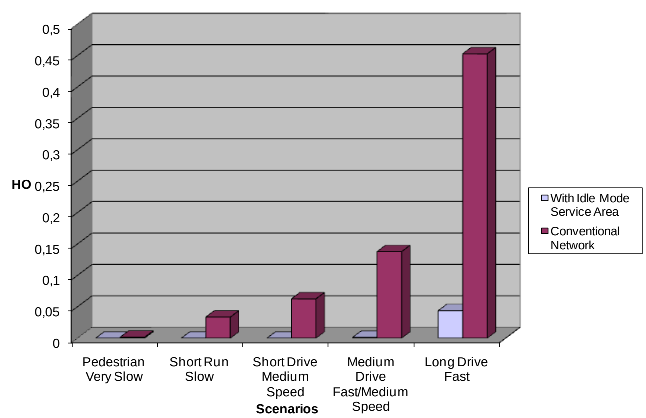

Figure 5. Number of handoffs experienced during active mode.

As a remedy, in this paper, we introduced the concept of additional boundaries that form the so-called Idle Mode Service Area for SGWs. These boundaries indicate where SGW relocation takes place for UEs in idle mode. The main rational beneath this optimization is that an *earlier* idle mode relocation is worthwhile to avoid a *later* active mode relocation.

The performance of the proposed optimization is evaluated with computer simulations considering different scenarios whereby UEs move at different speeds, ranging from pedestrian speed to highway speed. The obtained results demonstrated that the proposed optimization reduces remarkably the number of forced SGW relocations for UEs in active mode although this comes at a slight increase in the number of relocations for UEs in idle mode. Finally, whilst the present evaluation considered only the handoff occurrence frequency as a performance metric, investigating the interactions of the proposed approach with the overall QoS metrics (e.g., delay, packet drops, throughput, etc) forms the basis of our future research directions regarding this topic.

6. REFERENCES

[1] 3rd Generation Partnership Project, "Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," 3GPP TS 23.401.

[2] 3rd Generation Partnership Project, "Technical Specification Group Services and System Aspects; Architecture enhancements for non-3GPP accesses," 3GPP TS 23.402.

[3] Mihovska, A., E. Mino, E. Tragos, C. mensing, G. Vivier, and R. Fracchia, "Policy-Based Mobility Management for Heterogeneous Networks," Proc. of 16th IST Mobile Summit 2007, Budapest, Hungary, Jul. 2007.

[4] A. Klockar, E. Mihovska, E. Mino, and E. Tragos, "Network-Controlled Mobility Management with Policy Enforcement towards IMT-A," in Proc. of ICCCAS 2008, Xiamen, China, May 2008.

[5] H. Soliman, C. Catelluccia, K. El Malki, and L. Bellier, "Hierarchical Mobile IPv6 Mobility Management (HMIPv6)", Network Working Group, RFC 4140, Aug. 2005.

[6] J. Xie and I.F. Akyildiz, "A Distributed Dynamic Regional Location Management Scheme for Mobile IP", In Proc. IEEE INFOCOM, NY, USA, Jun. 2002.

[7] M. Woo, "Performance Analysis of Mobile IP Regional Registration", IEICE Trans. Commun., Vol.E86-B, No. 2, Feb. 2003. pp. 472-478.

[8] S. Pack and Y. Choi, "A Study on Performance of Hierarchical Mobile IPv6 in IP-based Cellular Networks", IEICE Trans. Commun., Vol.E87-B, No. 3, Mar. 2004. pp. 462-469.

[9] C.W. Pyo, J. Li, and H. Kameda, "Dynamic and Distributed Domain based Mobility Management Method for Mobile IPv6", In Proc. IEEE VTC Fall'03, Orlando, Florida, Oct. 2003.

[10] W. Ma and Y. Fang, "Dynamic Hierarchical Mobility Management Strategy for Mobile IP Networks", IEEE JSAC, Vol. 22, No. 4, May 2004, pp. 664-676.

[11] K. Kawano, K. Kinoshita, and K. Murakami, "A Mobility-based Terminal Management in IPv6 Networks", IEICE Trans. Commun., Vol. E85-B, No. 10, Oct. 2002, pp. 2090-2099.

[12] E. Natalizio, A. Scicchitano, and S. Marano, "Mobility Anchor Point Selection Based on User Mobility in HMIPv6 Integrated with Fast Handover Mechanism", in Proc. of IEEE WCNC, New Orleans, USA, Mar. 2005.

[13] W. Chung and S. Lee, "Improving Performance of HMIPv6 Networks with Adaptive MAP Selection Scheme", in IEICE Trans. Commun., Vol.E90-B, No. 4, Apr. 2007, pp.769-776.

[14] M.H. Liu and C.C. Yang, "A Multicast Extension to HMIPv6 with Efficient MAP Selection", in Proc. IEEE VTC 2005-Fall, Dallas, Tx, USA, Sep. 2005.

[15] S. Pack, M. Nam, T. Kwon, and Y. Choi, "An Adaptive Mobility Anchor Point Selection Scheme in Hierarchical Mobile IPv6 Networks", Computer Communications, Vol. 29, No. 16, Oct. 2006, pp. 3066-3078.

[16] 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. pp. 1630-1642.

[17] T. Taleb, A. Jamalipour, Y. Nemoto, and N. Kato, "DEMAPS: A Load-Transition Based Mobility Management Scheme for an Efficient Selection of MAP in Mobile IPv6 Networks," in IEEE TVT journal, Vol. 58, No. 2, Feb. 2009. Pp. 954-965

[18] X. Cheng, J. Luo, R. Halfmann, E. Schulz, and C. Hartmann, "Inter GW Load Balancing for Next Generation Mobile Networks with Flat Architecture," in Proc.69th IEEE VTC 2009-Spring, Barcelona, Spain, Apr. 2009.