Towards Supporting Highly Mobile Nodes in Decentralized Mobile Operator Networks

Tarik Taleb*, Konstantinos Samdanis*, Fethi Filali**
*NEC Europe, Germany
{taleb,samdanis}@neclab.eu
**QUWIC, Doha, Qatar
{filali}@quwic.com

Abstract— **There is a general trend towards the decentralization of mobile operator networks. Such network decentralization will not be efficient without rethinking mobility management schemes, particularly for users moving for a long distance and/or at a high speed (e.g., vehicles). To support such highly mobile users, this paper introduces a data anchor gateway relocation method based on user mobility, history information, and user activity patterns. The performance of the proposed schemes is evaluated through simulations and encouraging results are obtained.**

I. Introduction

Along with the ever-growing community of mobile users and the tremendous increase in the traffic associated with a wide plethora of emerging bandwidth-intensive mobile applications, mobile operators are facing a challenging task to accommodate such huge traffic volumes, far beyond the original network capacity [1] [2]. Effectively, current mobile networks are highly centralized, not optimized for high-volume data applications that shall come with 4G and beyond technologies.

In these networks, central gateways handle all mobile IP traffic; all traffic needs to be tunneled to the core, and no caching or data offload is supported at the network edge. As shown in Fig. 1, this highly centralized network architecture leads to i) high demand on central location due to "backhauling" of all data traffic, ii) dramatic increase in bandwidth requirements and processing load leading to undesirable bottlenecks, and last but not least, to iii) long communication paths between users and servers leading to waste of core network resources, undesirable delay, and poor Quality of Experience (QoE).

A straightforward solution to this issue may consist in having operators invest in speed; some claim that todays backbone will be tomorrows edge. Operators could also upgrade their core network nodes and build a scalable core network that can accommodate peak hours of these emerging bandwidth-intensive mobile applications. Whilst these are technically and technologically possible solutions, they economically represent a significant challenge for operators, particularly due to the fact that the Average Revenues per Users (ARPU) are dropping dramatically given the trend towards flat rate business models. Operators are thus investigating cost-effective methods for accommodating the increasing mobile network traffic with minimal investment into the existing infrastructure.

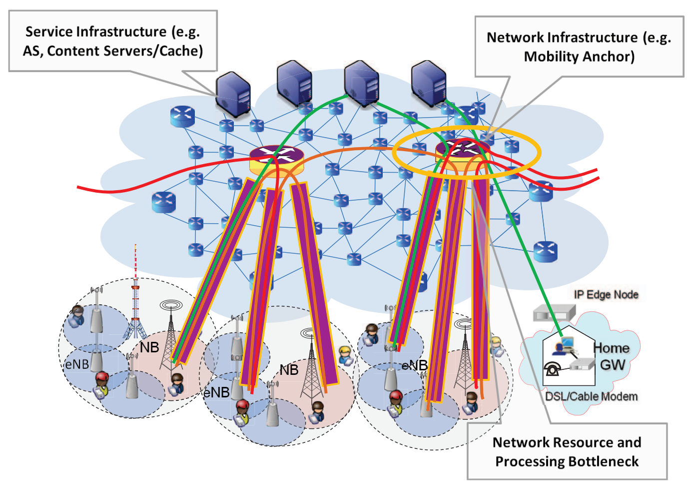


Fig. 1. Today's centralized mobile operator networks.

Selectively offloading traffic as close to the Radio Access Network (RAN) as possible is one of the key solutions, in which many operators have shown interest. According to [3], such traffic offload shall be achieved based on local data anchor gateways located close to the RAN. This essentially leads to a decentralized mobile network deployment as depicted in Fig. 2. In such a decentralized network, Packet Data Network Gateways (PDN-GWs), Serving GWs (S-GWs), and Mobility Management Entities (MMEs), are locally deployed to serve a local community of users. In this paper, we focus on the Evolved Packet System (EPS) but the general description can be equally applied to other 3GPP's mobile networks such as the General Packet Radio Service (GPRS).

The benefits of decentralized mobile networks are manifold. Indeed, with such decentralized networks, operators will be able to optimize usage of their resources by selectively breaking out IP traffic near to the network edge using Selected IP Traffic Offload (SIPTO) solutions [3]. Such vision fits the flat charging paradigm, while being also in line with the quest for a network architecture flatter than what has been achieved with the Evolved Packet Core (EPC) [4][5]. Effectively, by breaking out selected traffic at entities close to the moving terminal, operators will be able to avoid overloading their scarce core network resources (i.e., Gateway GPRS Support

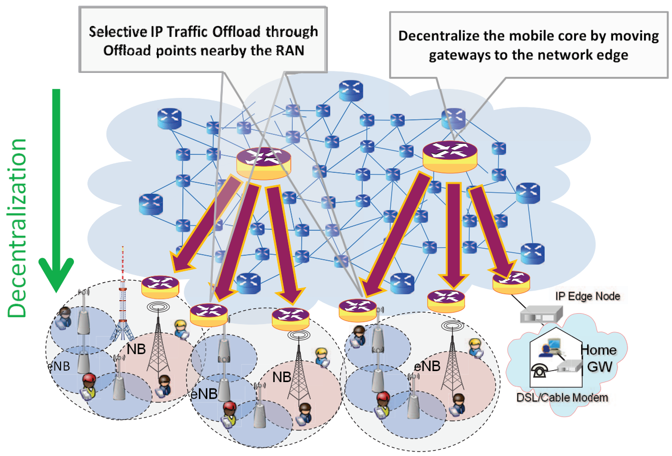


Fig. 2. The trendy decentralized mobile operator networks.

Nodes - GGSNs, Serving Gateway Support Nodes - SGSNs, and PDN/S-GWs).

The discussions and analysis in 3GPP have been on the definition of the architecture, i.e., on where the point of local breakout/traffic offload should be placed, in addition to issues regarding security, charging, mobility, traffic control/handling [6], and optimal gateway selection [7]. The efficient usage of the network resources in such decentralized architecture requires an optimal selection of data anchor gateway (in terms of geographical proximity and/or load) for User Equipment (UEs) [7]. In current standards, notifying a UE, particularly in idle mode, of the availability of such optimal gateway is immediately followed by enforcing the UE to disconnect and reconnect to the network. During this operation, the UE is relocated to the currently optimal gateway. For UEs in idle mode, moving fast and/or for a long distance (e.g., vehicles), this solution may lead to frequent unnecessary gateway relocations. To resolve this issue, a number of solutions can be envisioned. The objective of this paper is to compare among the different solutions and discuss their advantages and pitfalls.

The remainder of this paper is structured as follows. Section II describes some related work. Section III introduces our proposed solutions to the above-mentioned issues. The performance evaluation of the proposed solutions is discussed in Section VI. Section V concludes the paper.

II. Related Work

As discussed earlier, the introduction of the 3GPP Long Term Evolution (LTE) brought forward fundamental architecture and management features towards the deployment of a more decentralized cellular infrastructure. Certain functions like the Radio Network Controller (RNC) is merged within the evolved Node Base (eNB) flattening and simplifying the prior Universal Mobile Telecommunications System (UMTS) network architecture [4], while the network management follows a Self-Organized paradigm with several functions operating in a distributed manner [8]. Further flattening of the network architecture is also envisioned for the UMTS network in [9], where the GGSN and SGSN are integrated in the NBs, while in [10] a similar approach is introduced for EPS with the aim of bringing gateway functions to the edge of the network, merging, in this way, the RAN-CN (Core Network) split.

Traditionally centralized or hierarchical architectures were developed to share resources, while keeping the base station cost low [9]. Nowadays the evolution of computer equipment in combination with the increased mobile-originated as well as mobile-terminated data volumes has eliminated such economic reasons. In contrast, operators utilize data offloading solutions and Content Distributed Network (CDN) services in positions towards the network edge by introducing P/S-GW functionality close to eNBs. Further details on data offloading from a 3GPP perspective focusing on LIPA/SIPTO are provided in [11], illustrating specific network architectures and service requirements that meet the current decentralized needs, while enlightening the network management and deployment issues focusing on QoS and service continuity. More benefits associated with the use of decentralized architectures are documented in [12][13], with the most important one being route optimization and increased robustness.

The current research efforts regarding the decentralized cellular networks focus on mobility and service continuity. The enhanced handover performance and signaling minimization for active users avoiding the use of core network are among the most significant benefits, but the use of distributed mobility management is essential. In [14], a dynamic and distributed mobility management scheme is presented. It merges the mobility anchor and base station and uses tunneling to forward traffic upon a handover, allowing also the user to establish flows via different mobility anchors for efficient resource usage. A similar approach is also followed in [7], with additional mechanisms for performing load balancing ensuring QoS targets. An alternative solution that distributes the contents of a centralized mobility anchor to a set of distributed mobility agents using the concept of virtual routers and Distributed Hash Tables (DHT) is introduced in [15], with main benefits in load balancing and resiliency.

Considering the PDN connectivity of idle users, the work in [7] states the main challenges behind keeping PDN connectivity and examines the following PDN connection re-establishment solutions including periodic, upon a tracking area update, and upon network indication. This paper complements such solution introducing further parameters to perform PDN connection re-establishment based on user mobility and history information as well as user activity patterns. Although such selection criteria have been encountered before for centralized schemes as documented in [16], their scope was mostly focused on active sessions. Their performance is expected to be different in decentralized schemes, particularly in the case of UEs being in idle mode. The reason is mainly the different focus that concentrates on the timing issues of the new session establishment and whether the "always on" connectivity of UEs is maintained; parameters that affect the signaling overhead and negatively impact the session establishment delay.

III. Proposed Schemes

Before delving into details about the solutions we intend proposing towards supporting highly mobile nodes in decentralized mobile operator networks, we shall stress out that whilst the description in this paper relates to eUTRAN (evolved UMTS Terrestrial Radio Access) the same applies to UMTS and other types of networks. In case of UMTS, MME, eNB, S-GW, P-GW, and Service Area map onto SGSN, BS or RNC, SGSN, GGSN, and Routing Area or Location Area, respectively. The contributions of this paper relates to the notification of optimal P-GW availability and enforcement of P/S-GW relocation for highly mobile UEs being in idle mode.

In current standards, the network (i.e., MME) disconnects a UE whenever an optimal P-GW becomes available (i.e., without any intelligence). Since the network has no knowledge of the type of applications being active on the UE, their state (e.g. when they refresh their registration/subscription) and the impact of the forced disconnection, the UEs must in most cases immediately re-establish the PDN connection and also trigger certain applications to re-register/subscribe again (e.g. Instant Messaging or SIP applications). For fast/far-moving UEs (e.g., vehicles), this will clearly result in a large number of unnecessary signaling that have certainly some impact on the UE battery. Admittedly, the solution is straightforward and has minimal impact on the MME. It is, hereunder, referred to as the baseline gateway relocation.

As an alternative solution, the MME disconnects the UE with some intelligent logic. For example, the MME takes into account the history of UE mobility (e.g., number of handoffs/Tracking Area Updates (TAUs) performed) to decide whether to immediately disconnect the UE or to delay the disconnection request till after a predetermined period of time. In comparison to the above described baseline gateway relocation approach, this UE mobility-aware gateway relocation method reduces the frequency of PDN disconnection requests and consequently has less impact on the UE battery lifetime. However, the solution relies on the intelligent logic at the MME, as the MME needs to keep track of UE mobility.

In another solution, the MME indicates the availability of a more optimal P-GW in the TAU response without enforcing the disconnection. The disconnection is enforced only at service area boundaries. Compared to the baseline gateway relocation and the UE mobility-aware gateway relocation approaches, this solution is clearly efficient in terms of reduced signaling, as it enforces PDN disconnection only at service area boundaries and does not require any intelligent logic considering the UE mobility pattern at the MME (i.e., simple implementation).

Furthermore, it also enables the UE to take into account the type of applications that are active and their state in the decision when to re-establish the PDN connection. This allows the UE, for example, to re-establish the PDN connection, if it was previously indicated by the network, when the application updates its periodic registration/subscription status, and not just after, which would require another re-registration/subscription. In another approach combining the above solutions, the MME disconnects the UE for a given number of times as in the baseline gateway relocation method and after that the MME indicates the availability of a more optimal P-GW in the TAU response without enforcing the disconnection.

IV. Performance Evaluation

Following the description of the various S-GW and P-GW relocation schemes, this section provides a comparison study based on simulation using Matlab. In the simulations, the wireless access network is represented as a graph G(V,E) with V nodes denoting eNBs and E edges indicating adjacency between neighbor eNBs.

The Erdos-Renyi model [17], G(V,p) is used to create a random instance for the simulation topology with V = 60 number of eNBs and p = 0.12 probability of adjacency between two cells. Such model is adopted to create a relatively dense topology where each eNB has on average four neighbors representing a network formation composed by small cells, where network decentralization is envisioned. The coverage among neighboring eNBs is assumed to be ideal with no interference, while eNBs are assumed to impose no load limits in terms of the supporting number of users and the resource availability.

For the performance evaluation of the different P-GW relocation schemes, we consider a highly decentralized mobile network assuming P-GWs being collocated with eNBs. For the simulation study we considered the following details for each of the previously described solutions:

- Solution 1 (Optimal) maintains the optimal P-GW for a UE disconnecting it every time it performs a handover.
- Solution 2 (Dist-based) disconnects a UE once it performs a handover towards an eNB with distance more than two hops away from the one that is associated with the current P-GW.
- Solution 3 (Ses-driven) disconnects a UE only when they initiate a connection while not having an optimal P-GW.
- Solution 4 (Hybrid Dist/Ses) maintains the optimal P-GW disconnecting UEs for 3 consecutive times and then disconnects UEs only when they initiate a connection while not having an optimal P-GW.
- Solution 5 (Info-based) maintains the optimal P-GW disconnecting UEs that exhibit a history record for initializing connections within certain periods, while disconnecting the remaining UEs only when they initiate a connection while not having an optimal P-GW.

In the following, we assume that UEs enter the wireless network while being in idle mode following a Poisson distribution with a mean rate λ = 20 UE per minute and are assumed to remain inside the network according to an exponential distribution with mean μ = 10 minutes, which is half of the simulation duration of 20 minutes. UEs may initiate a connection with a probability p_{act} = 0.5 following a uniform distribution within the range of $[t_A, t_L]$, where the t_A is the

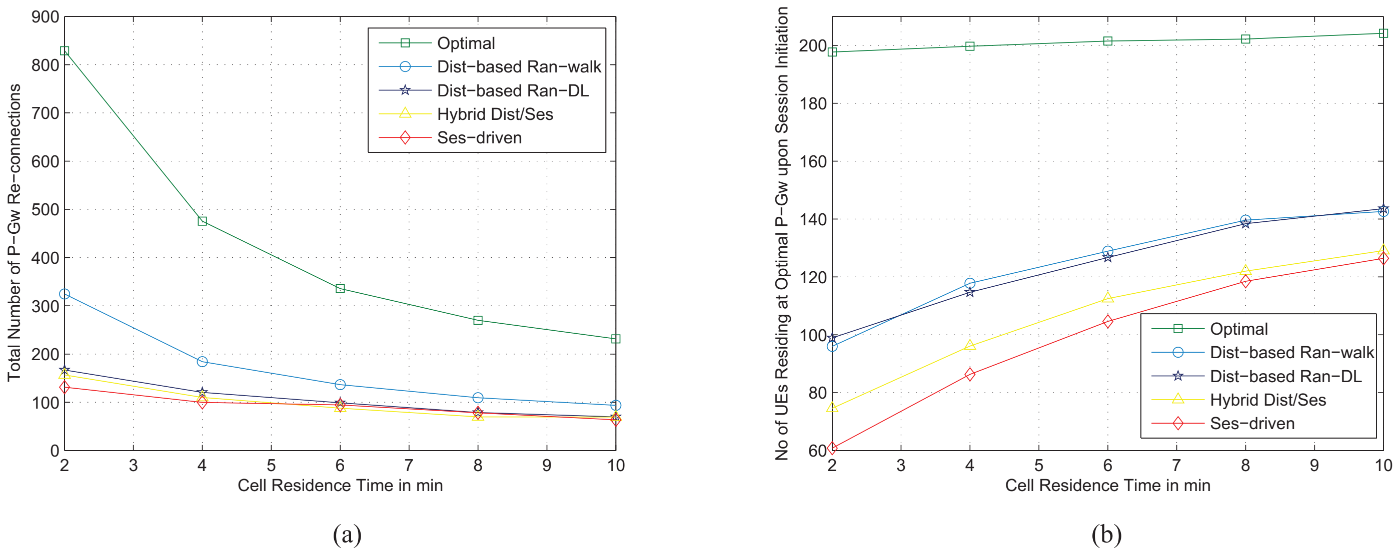


Fig. 3. Performance evaluation of the different P-GW relocation methods without taking into account any contextual information about users: (a) Signaling overhead, (b) P-GW optimality.

arrival time and t_L is the leaving time. Once a UE initiates a connection, it is assumed active for the remaining time until it leaves the network.

The evaluation of the proposed methods is performed considering the trade-off between the P-GW update overhead and P-GW optimality upon initiating a session. In particular, the P-GW update overhead is measured as the total number of updates, i.e. P-GW re-connections in the entire network, while the P-GW optimality is measured as the number of UEs that are associated with the optimal P-GW at the time when they initiate a new session. The simulation is performed altering the mean cell residence time $n = \{2, 4, 6, 8, 10\}$ minutes in order to evaluate the performance of the proposed schemes for different mobility speeds.

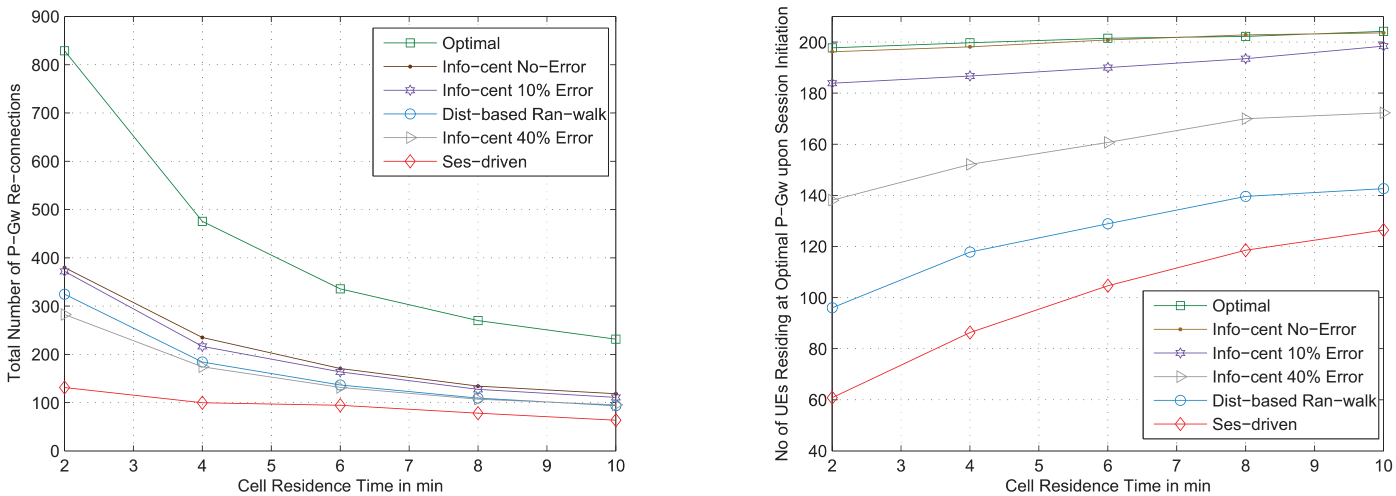
Fig.3 illustrates the signaling overhead and the P-GW optimality for Solutions 1-4. A general observation is the strong correlation between the P-GW overhead and the P-GW optimality: a high number of P-GW updates produces a high degree of P-GW optimality. Solution 1 ensures constant P-GW optimality irrespective of speed, at the cost of the highest signaling overhead, while the non-optimal approaches are 50% to 70% less accurate with Solution 3 introducing the smallest overhead at the cost of the lowest P-GW optimality and increased connection set-up delay.

All other solutions avoid increasing the connection set-up delay since UEs are constantly associated with a certain P-GW, though they introduce sub-optimal routing and thus result in relatively increased end-to-end delays. For higher speeds or shorter cell residence times, the difference in the signaling overhead between optimal and the remaining approaches increases, whilst for lower speeds all non-optimal approaches exhibit a similar amount of signaling overhead.

The performance of the distance-based approach, i.e. Solution 2, hinges on the UEs mobility pattern and on the traveled distance threshold after which P-GW updates occur. This solution is beneficial for UEs with relatively low mobility features; users moving not far away from the location of the initial P-GW, whilst for other cases its performance is proportional to the P-GW update frequency. Effectively, Fig. 3(a) demonstrates that the distance based approach with a random walk mobility model (Dist-based Ran-walk) produces much higher overhead compared to the equivalent one whereby movement towards an eNB in the vicinity of the current P-GW has a higher probability (Dist-based Ran-DL) despite the fact that both approaches have the same distance limit.

It should be noted that in terms of P-GW optimality, both approaches exhibit similar results as shown in Fig. 3(b). The hybrid and session driven approaches, i.e. Solutions 3 and 4, exhibit nearly the same signaling overhead. However, in terms of P-GW optimality, the hybrid approach outperforms the session-driven one. The magnitude of this better performance decreases as the cell residence time increases. The information centric approach, i.e. Solution 5, exhibits high P-GW optimality, almost similar to the optimal solution, while resulting in only half signaling overhead, and that is when the information regarding the behavior of particular UEs is accurate (Fig. 4).

However, in practice, it is all but impossible to predict the UE connection establishment behavior with no error. For this reason, we deliberately inserted different level of errors in the prediction of user behavior. Fig. 4 plots the signaling overhead and P-GW optimality in case user behavior is predicted with no-error, with 10% error and with 40% error. As comparison terms, we use the optimal solution, the session driven solution and the distance based random walk solution. When there are errors in the user behavior prediction, the P-GW optimality when the information centric approach is used degrades. This degradation increases along with errors in the prediction. How-



(a) (b)

Fig. 4. Enhancement in the performance of the P-GW relocation methods when taking into account contextual information about users: (a) Signaling overhead, (b) P-GW optimality.

ever, the signaling overhead remains enacted or comparable to the other schemes. These results elucidate that with more accurate knowledge on user behavior, an operator can largely improve the usage of its resources.

V. Conclusions

To cope with the emerging 4G traffic, many mobile operators are interested in decentralizing their networks. This network decentralization will not be effective, unless mobility management techniques are also rethought. In this vein, this paper introduced a set of solutions, particularly designed for users with high mobility features, such as vehicles.

Indeed for UEs in idle mode and traveling for a long distance and/or at a high speed, a number of solutions were proposed to notify these UEs of the availability of optimal data anchor gateways while minimizing unnecessary PDN disconnections. The proposed optimizations aim all for avoiding unnecessary signaling for the PDN re-connection to an optimal data anchor gateway, and unnecessary application layer signaling for registration/subscription based applications.

From the conducted simulations and the obtained results, it has become clear that approaches that allow the network to simply indicate to UEs the availability of an optimal PDN connection, without enforcing PDN disconnection exhibit better tradeoff in terms of P-GW optimality and signaling overhead. In some of these approaches, disconnection can still be enforced whenever needed (e.g. at service area boundaries). It has become also clear that knowledge of direction and speed of movement may further enhance the network performance and minimize the signaling overhead. Furthermore, knowledge of the user activity within certain time periods may help operators to categorize the users in order to reduce the signaling overhead and to enhance the user experience.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2009-2014," White Paper, Feb. 2010.

[2] "MOBILE TRAFFIC GROWTH + COST PRESSURES = NEW SOLUTIONS?," New Mobile, Jan 2010.

[3] 3GPP TR 23.829, "TS Group Services and System Aspects; Local IP Access and Selected IP Traffic offload", May 2010.

[4] 3GPP TS 23.401, "General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access", Jan 2011.

[5] 3GPP TS 23.402, "Architecture Enhancements for non-3GPP Accesses", Jun 2010.

[6] T. Taleb, K. Samdanis, S. Schmid, "DNS-based Solution for Operator Control of Selected IP Traffic Offload", IEEE ICC Kyoto, Jun. 2011.

[7] T. Taleb, Y. Hadjadj-Aoul, S. Schmid, "Geographical Location and Load based Gateway Selection for Optimal Traffic Offload in Mobile Networks", IFIP Networking, Valencia, May 2011.

[8] 3GPP TR 36.902, "TS Group Radio Access Network, Evolved Universal Terrestrial Radio Access Network (E-URAN), Self-configuring and Self-optimizing network (SON) Use cases and Solutions", Apr. 2011.

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

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

[11] K. Samdanis, T. Taleb, S. Schmid, "Traffic Offload Enhancements for eUTRAN", IEEE Communication Surveys and Tutorials, 2011.

[12] H. Chan, Problem statement for distributed and dynamic mobility management, IETF Internet Draft, Jul 2011.

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

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

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

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

[17] B. Bollobas, "Random Graphs", Cambridge University Press, 2nd Edition, 2001.