

An IETF-based Evolved Packet System beyond the 3GPP Release 8

Suresh Krishnan, Laurent Marchand, Gunnar Nilsson Cassel

Abstract — 3GPP is working on the new SAE Evolved Packet System (EPS) for Release 8 (R8), scheduled for completion in 2008. The target is a low-latency, higher data-rate, all-IP core network capable to support real-time packet services over multiple access technologies.

Two network architecture solutions are defined within the umbrella of the project:

- GTP-based solution (i.e., 3GPP TS 23.401)
- PMIP-based solution (i.e., 3GPP TS 23.402)

It is expected that a number of operators will deploy the first solution for various considerations while others will go for the second solution due to other reasons. Both solutions will be supported by the industry. This paper is strictly focusing on the IETF/PMIP-based network solution – while the further evolution of the GTP-based solution is by purpose left outside the scope of the paper.

The EPS R8 represents a major step forward towards the realization of the 3GPP objectives and vision. However due to time constraints, it may still leave room for improvements in future releases.

The purpose of this paper is to introduce a possible mobility architecture which goes beyond the current R8 architecture. It is also describing the key drivers for such an evolution.

Index Terms—

EPS	Evolved Packet System
LMA	Local Mobility Anchor
MAG	Mobility Access Gateway
PDN	Packet Data Network
P-GW	PDN Gateway
SAE	System Architecture Evolution
S-GW	Serving Gateway

I. INTRODUCTION

An important goal of the IETF/PMIP-based Evolved Packet System (EPS) is to provide a converged network architecture which integrates common mobility, security and QoS mechanisms for key fixed and mobile broadband accesses. The network architecture described in this document is different in some regards from the GTP-based solution. The key differences are:

- A different mobility architecture based on Proxy Mobile IPv6 (PMIPv6) and Mobile IPv6 (MIPv6).
- A different QoS approach referred as “bearer-less”

to indicate that a one-to-one mapping between the radio bearers and the tunnels between Serving and PDN Gateways is not needed. A single PMIP tunnel per Quality-of-Service Class between gateways is sufficient.

- An enhanced ‘off-path’ policy control framework with policy enforcement functions in both the Serving and PDN Gateways.

In this paper, we are providing the main motivations for a further evolution of the EPS beyond R8. Subsequently, we are describing a possible targeted mobility architecture and associated main reference points.

II. MAIN MOTIVATIONS

According to various industry sources the number of mobile subscribers will grow from 3 billions to 3.6 billions by the end of 2010. Therefore, the number of mobile phone users is expected to grow by less than 20%, while backhaul expenses risk increasing dramatically due to an exponential increase in bandwidth required for video and multimedia applications, as well as the need to support multiple technologies. Moreover, the Average Revenue Per User (ARPU) is likely to remain fairly flat. A growing data services ARPU will essentially compensate for the circuit-voice ARPU decline.

In 2006, the equivalent of 2-8 E1s/T1s was sufficient per cell site. In 2008, 10 to 30 Mbps is needed to sustained cell site equipped with HSPA. Emerging mobile broadband technologies such as LTE will drive the requirements to over 500 Mbps per cell site. Without technical considerations in the access network and EPS, the Internet inter-connect and transport costs could rise significantly, seen as part of the operators network OPEX. Operators must target improved efficiency in their networks by dropping the cost per Mbps of bandwidth.

An evolution of the EPS can contribute to this target by avoiding un-necessary routing (e.g., selecting the shortest path between user-equipments), filtering un-wanted IP traffic (e.g., SPAM, Malware, Denial of Service, virus, etc.), introducing tunneling reduction mechanisms for the selected mobility protocols, and by providing better control of peer-to-peer services.

The 3GPP EPS Release 8 is supporting a network-based mobility protocol (e.g., PMIPv6) and a Client-based mobility protocol (e.g., CMIPv6). Since there is no clear separation

between local and global mobility, it is not currently possible to simultaneously use both mechanisms. This limitation is significantly important in a Fixed Mobile Convergence (FMC) scenario. As an example, a user is served by a Residential Gateway equipped with a fiber modem and a WLAN Access Point (AP). Let's also assume that this subscriber is using a multi-access terminal with support for LTE and WLAN. If this user initiates an IMS (Internet Multi-media System) VoIP (Voice over IP) call in his/her home, it is probable that the voice will be carried over the WLAN up to the Residential Gateway. From the Residential Gateway, the fiber access will be used to carry the voice data traffic. If this subscriber is walking out the door while speaking on his mobile phone, when the signal strength toward his private WLAN AP is fading, a seamless handover toward the LTE access is desirable. If the selected mobility protocol is PMIPv6 – there is a problem. The EPS only sees the Fiber and LTE accesses. It has no knowledge about the existence of the WLAN. In this particular scenario, only the mobile node can initiate a handover procedure. **A mixed mobility mode** is one possible solution – allowing an operator to combine the advantages of network-based and client-based mobility protocols.

The overhead linked to PMIPv6 and CMIPv6 could be significant when taking into account the huge amount of IP traffic sustained by fixed and mobile broadband accesses. **Tunneling optimization** mechanisms could be used to reduce the overall transport costs associated with the utilization of these mobility protocols.

In a R8 (PMIP) roaming scenario, a PMIP tunnel is established from the LMA in the Home P-GW to the MAG in the Visited S-GW. If the selection of the S-GW is under the control of the P-GW in the home network, it leads to a situation where the home network has the major role to play in the access selection for its roaming subscriber. Unfortunately, the home has no knowledge of the actual network conditions in the various supported accesses within the visited network. One particular access type could be congested while another one is underused. Since the visited operator is the actual owner of the access resources – EPS architecture should ensure that the visited network plays a relevant role in the access selection procedure.

In some roaming scenarios, the utilization of the “home tunneled” approach where the data traffic of each subscriber is forwarded back to their respective home network can result in high latency and consequently “poor” end-user experiences. **Local Breakout** is a mechanism which permits a user-equipment to be connected to one PDN Gateway in the home and to another PDN Gateway in the visited network. The latency and transport costs could be reduced by the utilization of the local P-GW. However, client applications in the mobile node must be able to select the appropriate PDN for the various services. Beside the “home tunneled” and “local breakout” mechanisms, an additional solution could be introduced to avoid the unnecessary complexity in the user equipment.

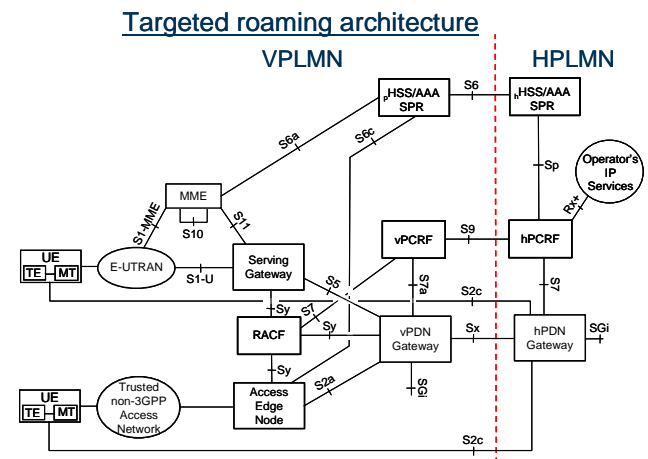
When several mobile operators are competing in a market

where cellular penetration is close to 100%, it is becoming difficult for one operator to differentiate itself from another. Furthermore, even with a large set of services, it is far from certain that an operator with its associated brand name will be capable to attract all segments of the population. Therefore it is expected that a number of operators may establish distinct divisions or work with separate companies in order to tackle particular groups of consumers (e.g., teenagers, retirees, etc.) with specialized service offerings. Moreover, a nation-wide deployment of a new radio technology represents a significant investment. For those reasons, it is highly probable that, at least in certain countries, several operators may share a common access network infrastructure. The EPS needs the ability to support **a clear separation between access networks and service providers**, to efficiently support such scenarios.

Finally, a simple **access agnostic QoS application protocol** which could be used by peer-to-peer services and be properly “regulated” by the policy control infrastructure would further contribute to the improvement of the 3GPP EPS.

III. WAY FORWARD FOR THE EPS ARCHITECTURE

In this section, we are presenting a possible targeted architecture of the Evolved Packet System beyond Release 8. Due to space constraints and simplicity considerations, we are only covering a subset of the overall EPS network architecture – we are mainly addressing the roaming architecture. We are not covering the support of non-3GPP un-trusted networks, or any GTP-based solution. Only the main relevant reference points are presented in the simplified targeted roaming architecture diagram below:



A first possible evolution is related to the introduction of a local anchor in the VPLMN – a local anchor is by the way already included in some of the possible network scenarios in R8. This anchor allows us to have a clearer separation between local (e.g., S5 and S2a) and global mobility (e.g., Sx or S2c). One potential evolution path to introduce this local anchor could be defined by simply further developing the local PDN gateway used for Local Breakout. *Please note that several others, technically equivalent, evolution scenarios are*

also possible.

The utilization of a local anchor will further improve the overall flexibility of the EPS by allowing an independent evolution of the local and global mobility protocols. Additionally, it will allow combining, in an efficient fashion, network-based and client-based mobility mechanisms.

Another possible evolution consists in a better separation of the service policy control functions (e.g., PCRF) from the Resource and Admission Control Functions (e.g., RACF). This approach, promoted by the ETSI TISPAN organization, permits distinct operators to offer different sets of IMS and non-IMS services over a shared EPS and access network infrastructure. In the above situation, we have two operators acting as Network Service Providers (NSPs) while a separate entity is acting as the Access Network Provider (ANP). Naturally the same operator can simultaneously perform the NSP and ANP roles. Each NSP has a PCRF to “police” services while the ANP has a RACF for performing admission control and efficiently managing access network resources.

These two differences in relation to EPS R8 are described in more details in section IV and V.

In the described evolved roaming architecture, the Visited PDN GW is serving both non-3GPP Access Edge nodes (e.g., BRAS, ASN GW, PDSN, etc.) and 3GPP Serving GWs. Roaming subscribers can move between accesses without the need for signaling back to their respective home network. The local mobility protocol could be complemented by using either a global network-based (e.g., Sx) or a client-based mobility protocol (e.g., S2c).

The full roaming interface consists of the S9, S6, and S2c or Sx reference points – it is able to support both 3GPP and non-3GPP accesses.

IV. MOBILITY ARCHITECTURE

This mobility architecture of an EPS beyond Release 8 describes the mobility protocols that could be used for implementing the various reference points defined in the overall architecture. Two basic classes of mobility protocols – **network-based and client-based** – have to be considered for solving these mobility issues.

1) *Network based mobility protocols; In the case the local mobility is network-based, all the mobility signaling is performed between the EPS network nodes, and the UE is not involved. The advantage of these protocols is that mobility services can be provided to terminal equipment that is not mobility aware. These protocols also reduce the amount of signaling and data tunneling overhead on the radio interface. The downside is that the applicability of these protocols is limited, and some features, like latency or transport cost reduction, may be hard to implement. The network-based mobility protocol that we will use to describe the architecture is Proxy Mobile IPv6 (PMIPv6).*

2) *Client-based mobility protocols; In the case the local mobility is client-based, all the mobility signaling is initi-*

ated by the UE. These protocols provide more features than the network-based mobility protocols and can gracefully handle more complex scenarios. Since the signaling and data tunneling are initiated by the UE, there is some associated waste of radio resources. There are mechanisms that have been designed to reduce, if not eliminate, the additional overhead brought on by client-based mobility protocols, but they are not widely deployed yet. The client-based mobility protocol that we will use to describe the architecture is Dual Stack Mobile IPv6 (DSMIPv6).

The solution for mobility in such networks needs to address two distinct challenges. The first of them is local mobility, or micro-mobility, that is concerned with the movement of the mobile node within the visited network (or a network domain). The second challenge is global mobility, or macro-mobility, that is concerned with the movement of the mobile node across multiple visited networks (or network domains). Given this way of dividing the mobility challenge, it is conceivable that the protocols for solving local mobility may be different from those for solving global mobility.

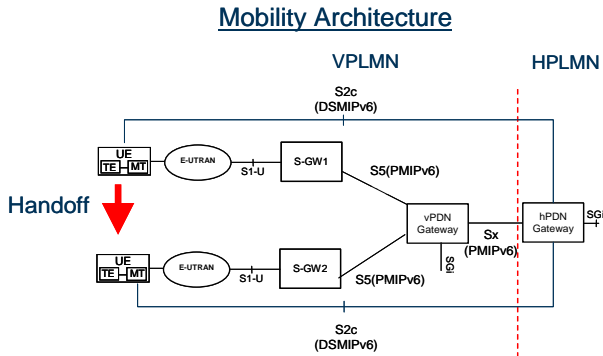
A. Local mobility

This term describes the kind of mobility when a mobile performs a handover within an administratively (and perhaps geographically) restricted domain – e.g., within a VPLMN. When a mobile is roaming in a visited network, micro-mobility protocols assure that the attachment changes in the visited network need not, unless requested, be communicated to the home network. Hence, micro-mobility signaling can be accomplished in a relatively quick fashion so as to reduce the handoff latency. Since the local mobility challenge is usually constrained to a smaller geographical region, the latency and transport costs are usually lower. Hence network-based mobility protocols work very well in this scenario, since their limitations are not so relevant. Hence we recommend that network-based mobility protocols are used for local mobility – and the preferred IETF protocol to implement this reference point is PMIPv6. Once the MME triggers the S-GW, the S-GW initiates a Proxy Binding Update (PBU) message towards the P-GW in order to notify it of the new location of the UE. The P-GW updates its binding cache and sends an acknowledgement (PBA) message to the S-GW. At this point a tunnel is established between the S-GW and P-GW to carry the user traffic. After this, traffic can flow from and to the UE through the P-GW. Please note that the tunnel can be pre-established to reduce the handover latency, and might be shared with other UEs for scalability reasons.

B. Global mobility

This term describes the kind of mobility when the mobile moves across administrative (e.g., geographical) boundaries. In this case the signaling needs to be performed towards the home network of the mobile. Hence, macro-mobility signaling cannot usually be performed as quickly. There are two approaches to solving global mobility. We can either use network- or client-based mobility protocols for this purpose.

The network-based mobility protocols, as defined today, do not support route optimization and hence they risk increasing latency and/or transport costs. Hence, using client-based mobility protocols is natural for global mobility. As an alternative, network-based mobility protocols can be enhanced to support, e.g., route optimization features and/or handover performance.



C. Desirable features for enhanced mobility

1) Local Anchor

The roaming scenario described earlier can be implemented using only global mobility protocols – i.e., all the signaling is passed back home. But having a local anchor and separating local mobility from global mobility offers several advantages

- Keeps VPLMN in control for access selection. The visited network can observe network conditions and possibly move the UE between accesses
- Allows for charging and lawful intercept in the visited network
- In a wholesale business model, where there are many non-3GPP access network operators – reduce the number of roaming agreements to sign
- It makes route optimization possible and hence provides the possibility to reduce latency and transport costs
- Allows the visited network to provide services like localized services, emergency calls etc.
- Reduces mobility signaling back to the home network

Given all these features that are made possible, it is highly recommended to use a local anchor in the visited network. It is possible to locate the local anchor at the visited P-GW, since it plays a corresponding role for non-roaming UEs when in their home network – and hence basically no, or at least a minimum of, additional functionality needs to be specified.

2) Tunneling Optimization

As described earlier, client-based mobility protocols come with a considerable tunneling overhead on the radio interface. One way to tackle this problem is to figure out a way to reduce the tunneling overhead. In the existing solution for

mobility using bidirectional tunneling all the data flowing between the UE and another node it is communicating with, is tunneled inside another IPv6 packet which carries the data between the UE and the P-GW. This means that there is a minimum 40 byte tunneling overhead per packet leaving the UE. The UE is on a wireless link typically on the last hop, and these 40 bytes per packet could result in a huge wastage of the available spectrum. It should furthermore be noted that this overhead exists also in the EPC network between the S-GW and the P-GW even with network based mobility protocols, since they use the same tunneling mechanism. When we analyze the overhead we can observe that the content of the tunneling packet does not change very much over the session. Thus we can abstract out the common data carried in every tunneled packet and signal it to the P-GW when we need to initiate communication to a new node. Using these optimizations we can greatly reduce, if not eliminate, the tunneling overhead between mobile entities.

3) Traffic Localization

Consider the case where two mobiles are attached to the same S-GW in an EPS network. If these two mobiles need to communicate with each other, the traffic flows from the S-GW to the P-GW and back to the same S-GW. There might be valid reasons for this to happen – e.g., the P-GW would like to inspect all traffic – but this leads otherwise to unnecessary latency for the communication and takes up bandwidth that may be better used for other traffic. So, if the network policy permits, it should be possible for the P-GW to signal to the S-GW that these types of traffic should be localized to the S-GW and pass directly through, instead of first passing to the P-GW and back again. This scenario can also be extended to cover two mobiles attached to two distinct S-GWs that communicate with the P-GW. If there is a direct lower cost path between the S-GWs, it should be used if the P-GW directs the S-GWs to do so.

V. QoS AND POLICY CONTROL

The main objective of a QoS-enabled infrastructure is to ensure that the users get the experience they expect, while the operator who manages the network can make optimal usage of limited network resources. Several QoS models have been defined by IETF. The main ones are:

- Over-provisioning
- Static provisioning
- Signaled provisioning (e.g., RSVP, NSIS)
- Differentiated services
- Measurement Based Admission Control (MBAC)
- Resource and Policy Based Admission Control

By and large, these mechanisms are complementary. The issue for the industry is determining which mechanisms to use, in which combinations, and in which parts of their networks.

The 3GPP has already selected, for the EPS, a QoS approach based on a combination of the Resource and Policy Based Admission Control model, Differentiated services,

Signalled provisioning and MBAC.

Even if many IETF QoS standards exist, unfortunately they tend to be used within a single network domain only, without any standardized service classes for the entire network. The absence of widely accepted standard service classes is a major problem. A service class is essentially a specific set of QoS parameters grouped together to achieve a particular type of traffic handling. Fortunately, 3GPP has addressed this problem by introducing the concept of QoS Class Identifiers (QCI) already in Release 7.

We intend to promote the utilization of the QoS classes (QCIs) defined by 3GPP over all access network technologies supported by the EPS. There is furthermore no technical reason found to change the R8 EPS QoS architecture. Our intent is therefore to strictly further improved the QoS model.

A fundamental reference point for roaming is the S9. When an Application Function (AF) resides in the home network, and Local Breakout or Route Optimization is used, the AF may have to rely on network services from the VPLMN for efficiently delivering services. As an example, a roaming subscriber abroad is calling a taxi from his mobile phone. The taxi company may be equipped with an IP phone system relying on a 64 kbps codec. Since mobile phones are not typically supporting such a codec, trans-coding functions from the VPLMN are needed. The S9 could allow a VPLMN to advertise to the PLMN(s) the various network services which it is offering. Moreover, it permits applications in the HPLMN, via the Rx+ reference point, to request the utilization of those services for each specific media flow associated to a particular session. The S9 reference point is already defined in R8 – beyond R8, we simply could add the support for additional services, as needed

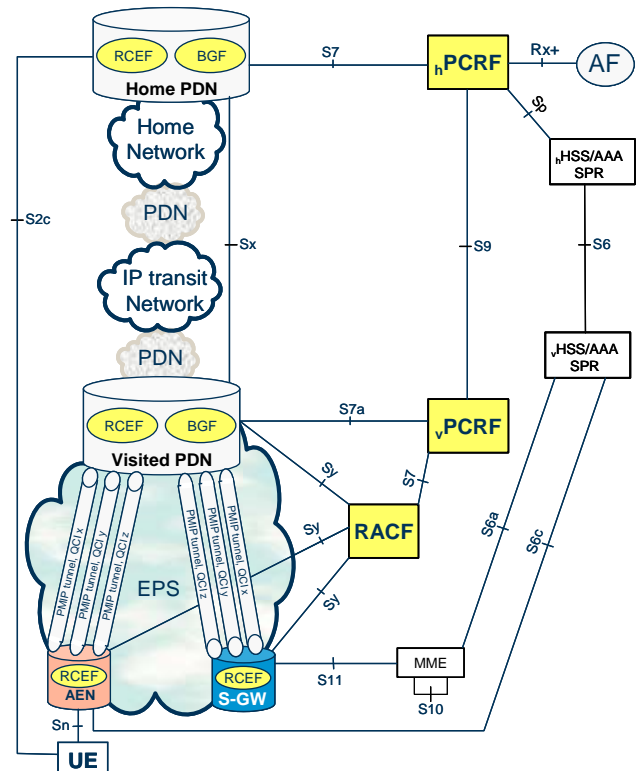
The picture to the right depicts the overall proposed PCC and QoS architecture.

The main considered enhancements consist of the following:

1) Introduction of a logical RACF entity along with a new reference point (Sy)

As previously described in this paper, Fixed and Mobile Convergence is leading towards the ability for a better separation of some business roles, better support for shared networks, and a clear separation of services from the actual access networks. Consequently we are looking into bringing the best aspects of the 3GPP PCC and the ETSI TISPAN architectures together – in order to define a future converged policy control infrastructure. As already adopted by TISPAN, we are recommending a clearer separation between service policy control and admission and resources management functions in the reference architecture.

Please note that whether such a separation is needed or will be implemented for all networks, e.g., in mobile networks, remains to be proven. A separate RACF function might negatively impact, e.g., the set-up time.



2) Introduction of a new IETF NSIS-based access agnostic QoS application protocol (Sn)

With the increase usage of peer-to-peer services, a flexible QoS architecture cannot be limited to only sustain a network-initiated QoS approach. We must support both a network-initiated and a terminal-initiated procedure. Unfortunately, the current mechanisms in the industry are inadequate. RSVP was not designed for a mobile environment. The current proposed NSIS QoS application protocol is still too complex (e.g., too many parameters) – and other available solutions are access specific. Consequently, we are considering introducing a new simpler and access agnostic QoS protocol which would run on top of the NSIS Network Transport Layer Protocol. We are looking to fully embrace the 3GPP QCI approach by allowing an application to only specify the needed QCI, Requested Bit Rate, and Minimum Bit Rate. We are also considering integrating this new NSIS QoS protocol with the policy control infrastructure. This approach would permit operators to be in control regardless of whether the QoS is initiated from a service in the network or in the terminal.

All services that run over a network have an expected quality and response time, which can be met if sufficient network resources and relevant procedures are available. If the expectations are met, the user will perceive “good” quality of the service – and if not, insufficient quality of service.

Network operators wish, of course, to maximize the utilization of its network resources. Ideally, all resources would be utilized to one hundred percent so that all the invested capacity could be used (at least during peak demand). To do this all traffic would be buffered – regardless of traffic type, quality and delay consideration – and put on the network

when capacity is available.

This leads to a dilemma. If we want to maximize network utilization, we have to ignore the service requirements and pack data on equipment and links as much as possible. However, if we do this, the perceived user service experience would be “poor”. If we want the user to have a good experience – then we need to make sure that more than sufficient capacity is always available, by heavily over-dimensioning the system. If we do this – we risk, of course, making inefficient use of the network resources.

The solution to this dilemma is to introduce needed but pragmatic Quality of Service mechanisms in the network. These could ensure that different application/user-specific traffic is intelligently allowed into and treated within the network, in order for the users to get the experience they desire (and pay for) while utilizing the network resources efficiently.

VI. CONCLUSIONS

The utilization of a common core network to serve a number of wired and wireless broadband access technologies will represent a major step forward in the communication industry. This paper presents a number of motivations and potential requirements on an evolution of the Evolved Packet System (EPS) beyond Release 8. It is also proposing a number of mechanisms to address these issues.

The possible Beyond R8 EPS architecture is designed to sustain latency-sensitive services over high data-rate IP networks. Moreover the proposed mobility protocols and mechanisms have been selected with the intent to deliver services over multiple fixed and mobile broadband access technologies. The aim is to support inter-technology seamless handover between all supported access types by the EPS.

The introduction of new access technologies such as LTE, along with flat-rate, can significantly increase the transport and Internet inter-connect costs. Consequently, an important concept supported by this architecture is to typically forward the application signaling from a subscriber back to his home network service infrastructure. However, the payload data path for this particular application is regularly optimized, unless specifically requested otherwise, e.g., by an application or policy function.

Several architecture choices have been made to reduce the amount of signaling back to the home network. Several optimization mechanisms are also proposed to reduce OPEX and CAPEX while improving traffic characteristics: Local Break-Out, Tunneling Optimization, Traffic Localization, and Route Optimization.

Which of the discussed new requirements that eventually will be agreed and associated with an EPS Beyond R8 is, however, still an open question. This paper has, nevertheless, outlined a possible way to accommodate them in IETF/PMIP-based EPS networking specifications of the future.

REFERENCES

- [1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".
- [2] 3GPP TS 23.401: "GPRS Enhancements for E-UTRAN Access".
- [3] 3GPP TS 23.402: "Architecture Enhancements for non-3GPP accesses".
- [4] IETF Internet-Draft, draft-ietf-netlmm-proxymip6-05.txt, "Proxy Mobile IPv6" work in progress
- [5] IETF RFC 3775, "Mobility Support in IPv6".
- [6] IETF Internet-Draft, draft-ietf-netlmm-pmip6-ipv4-support-00.txt, "IPv4 Support for Proxy Mobile IPv6" work in progress.
- [7] TS 23.203: "Policy and Charging Control Architecture".
- [8] 3GPP TS 22.278: "Service requirements for evolution of the system architecture".
- [9] IETF Internet-Draft, draft-ietf-mip6-nemo-v4traversal-03.txt, "Mobile IPv6 support for dual stack Hosts and Routers (DSMIPv6)" work in progress
- [10] IETF RFC 4555, "IKEv2 Mobility and Multi-homing Protocol (MOBIKE)"