Comparison of MME Signaling Loads for Long-Term-Evolution Architectures

Indra Widjaja*, Peter Bosch†, Humberto La Roche‡
*Bell Labs, Alcatel-Lucent, Murray Hill, NJ 07974
†Bell Labs, Alcatel-Lucent, Antwerp, Belgium
‡LTE Solutions Group, Alcatel-Lucent, Murray Hill, NJ 07974

***Abstract*—The LTE architecture consists of eNBs that provide wireless connectivity to each UE, S-GWs that anchor the user plane for mobility, MMEs that provide control-plane support to each UE in its domain such as handover, paging, bearer management and tracking area updates and finally, PDNs that bridge the cellular network to the Internet. In this paper, we investigate LTE architectures with distributed MMEs and with a centralized MME. We present an analysis of the signaling load performance at the MME serving multimedia-capable UEs and compare the performance under both architectures. We propose a multicast paging procedure to alleviate the MME signaling load and compare the MME performance with multicast and unicast paging procedures. Our results show significant benefits of multicasting.**

I. Introduction

The evolution of the current 3G wireless technology under the name of Long-Term Evolution (LTE) has been considered by the Third Generation Partnership Project (3GPP) to meet future demand and maintain its competitive position for the future [1]. 3GPP has developed a framework for an evolved UTRAN (E-UTRAN) with high-level objectives [2][3] that include: higher data rate, improved spectrum efficiency, reduced latency in user plane and control plane, inter-working with other (3GPP or non-3GPP) wireless systems, and reduced cost.

Some key network characteristics of LTE include flatter architecture and usage of IP-only protocol for signaling and bearer transport. Fig. 1 shows the overall network access architecture that has two parts: E-UTRAN consisting of evolved Node-Bs (eNBs) and Evolved Packet Core (EPC) consisting of Serving Gateways (S-GWs), PDN-Gateways (P-GWs) and Mobility Management Entities (MMEs) [4]. The eNBs provide wireless connectivity to User Equipment (UE or terminal) and are inter-connected with each other via the X2 interface to handle mobility management, load management, and general management and error handling situations among the eNBs [5] The eNBs are also connected to the EPC via the S1 logical interface [6]. The S1 interface consists of S1-MME (or S1-C) interface for control-plane traffic and S1-U for user-plane traffic.

The separation between control and user planes in LTE poses an interesting question on the distribution of control-plane elements. As shown in Fig.1(a), MMEs can be distributed in local offices to reduce signaling delays. Communication among them is facilitated through S10 interface. Fig.1(b) shows an alternative architecture where a centralized

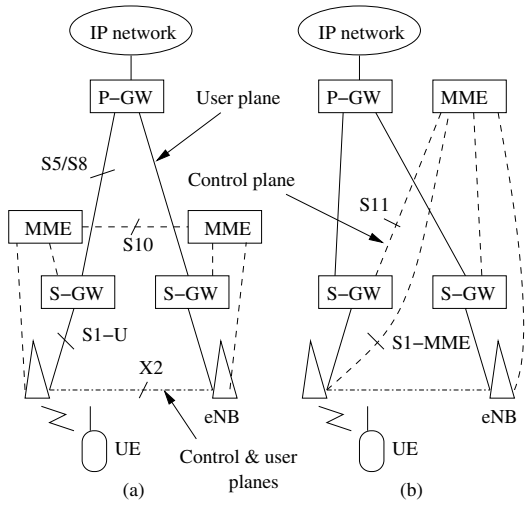

Fig. 1. LTE architecture with (a) distributed MMEs and (b) a centralized MME.

MME (or a centralized MME organized as a *pool* of MMEs) is located in a regional office. MME centralization serves the signaling functions for all the S-GWs in its domain. This architecture is attractive since it takes advantage of sharing and eliminates/reduces inter-MME signaling.

The rest of this paper is organized as follows. We first describe signaling flows that have appreciable impact on MME signaling load. We investigate both unicast paging as specified in the standard and multicast paging as a proposed enhancement. We then provide an analytical model to calculate the signaling load at the MME that serves UEs. Finally, we present our comparative results on MME signaling loads.

II. MME Signaling Flows

In this section, we describe the signaling flows that can have appreciable impact on MME processing load. There are four main events that contribute to the signaling load at the MME, which we discuss below.

A. UE-originated session

The event is triggered through a UE-originated session, as shown in Fig.2. This message sequence aims to set up a communication path, or EPS/SAE bearer between the UE and the Internet. This bearer can be Quality-of-Service (QoS) controlled and runs between the P-GW, through the S-GW, and the eNB.

978-1-4244-2515-0/09/$25.00 ©2009 IEEE

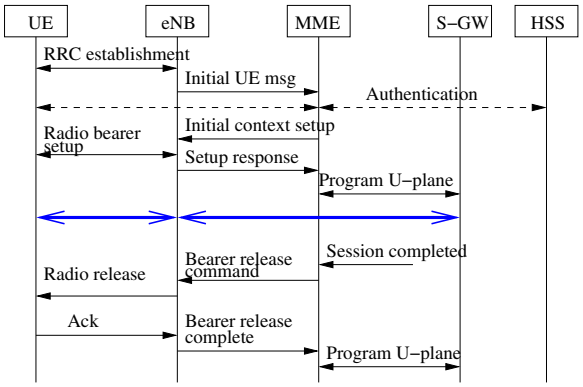

Fig. 2. UE-originated session (user-plane flow in blue).

In UE-originated session setup procedure, we assume that a session is successfully completed and no failure arises in any signaling exchange. The scenario in the figure assumes that UE's connection management is initially in IDLE state where no RRC connection exists (see [7]). Upon the establishment of the RRC connection and a service request from UE, the eNB selects a serving MME and sends an initial UE message to the MME to trigger an initial context setup. The message includes a unique UE identity within the eNB as is described separately [8] and any Non-Access-Stratum (NAS)-Packet Data Unit (PDU) that is transparently relayed by the eNB from UE to MME. Authentication to the Home Subscriber Server (HSS) is optional unless when no UE context exists in the network (e.g., when the UE first powers up).

Upon reception of the UE-originated session, the MME sends an initial context setup request containing UE identity within the eNB-MME pair, bearer identity, bearer level QoS parameters, receiving IP address and GTP-U tunnel identifier (TEID) for identifying multiplexed uplink traffic over the tunnel between SGW and eNB. If initial context setup is successful, the MME then programs the user-plane of the S-GW accordingly to set up the bearer between the eNB and the S-GW. Finally, the MME can terminate a session by sending the bearer release command to the eNB to release all bearer-related resources.

B. UE-terminated session

Fig. 3 shows the signaling flow for UE-terminated session where it is assumed that the connection management at the MME for the UE is initially in IDLE state. A new incoming session may be triggered by explicit signaling to establish a connection or by a downlink data packet arriving at the serving S-GW that does not have the associated TEID for the packet. In IDLE state, the MME only knows the location of the targeted UE within its tracking area and has to initiate a paging procedure to find and activate the UE.

3GPP describes the use of SCTP to transport S1-MME control messages and only allows unicast message transmission from MME to eNB. This implies for paging control messages that copies of paging information will be have to be transmitted to all the eNBs within the tracking area and each eNB is addressed separately. However, paging message

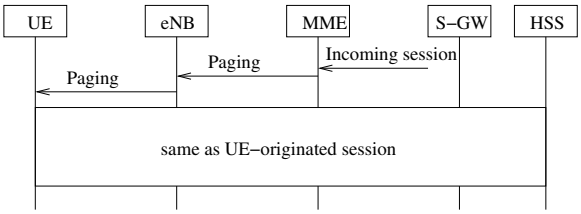

Fig. 3. UE-terminated session.

delivery is a typical example of an application where all eNBs simultaneously need to be addressed. One of the eNBs that eventually locates the UE responds with an initial UE message and the subsequent signaling flow is identical to that for UE-originated session.

Since unicast paging can be potentially expensive in terms of MME signaling load, we propose using multicast for transmitting paging information from the MME [1]. While 3GPP only defines a SCTP connection between the MME and eNB, we propose to use UDP over ESP for paging message delivery at the eNBs. Other messages between MME and eNB are still transmitted in a unicast fashion over SCTP. We will quantify the impact of paging multicast on the signaling load performance later in this paper.

When an incoming session arrives while the UE is in CONNECTED state, the MME knows the location of the UE within the eNB and thus can simply update the bearer at the eNB. The MME also needs to reprogram the S-GW. In this paper we assume a model where sessions are initiated only when data needs to be transferred to/from the UE. The data session persists until it either times-out or is torn-down at the application layer.

C. Handover

This event is due to a UE handover that handles mobility while UE is in CONNECTED state[2]. There are several scenarios for intra-LTE handovers:

- Intra-eNB handover: A UE performs a handover from one cell sector to another within the same eNB. In this case, there is no signaling at MME and thus has no impact on the MME load.
- Inter-eNB handover in the same MME pool area: A UE performs a handover from one cell to another under the control of the same serving MME. There are two possible cases. In one case, the UE stays with the same S-GW (S-GW pool area). In another case, the UE may need to change to another S-GW.
- Inter-eNB handover with MME change: A UE performs a handover from one cell served by a source MME to another cell served by a target MME within the same LTE.

The first scenario can be trivially ignored by looking at an expanded cell served by an eNB. The second scenario occurs most likely with a centralized MME that serves a very large

[1]Threats and counter measures using IP multicast have been noted in [9].

[2]In IDLE state, a UE performs cell reselection to change cells, which does not involve MME [10].

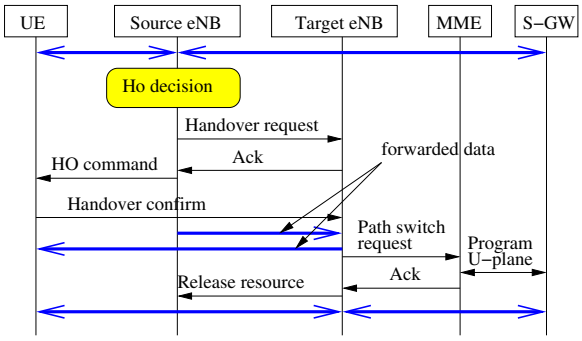

Fig. 4. Intra-MME handover.

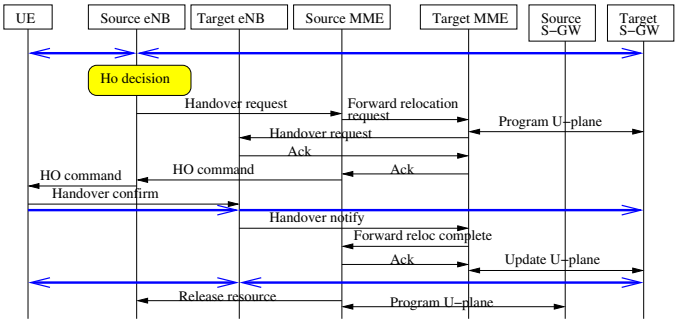

Fig. 5. Inter-MME handover.

number of eNBs within a region. It can also occur with the distributed-MME architecture since an MME typically serves multiple eNBs. The third scenario occurs with the distributed-MME architecture when MME relocation is needed. It can also occur with the centralized-MME architecture when a UE moves outside the region controlled by the centralized MME. We assume this scenario is less frequent compared to handovers within a large region. It is also worth noting that during transition phases when current 3G technologies are still prevalent, it is likely that inter-technology handovers may also occur frequently. This scenario will be ignored in this paper as this is likely only a transitional phase.

Fig. 4 shows the signaling flow for inter-eNB handover in the same MME pool and S-GW pool. Handover decision is initiated by the source eNB upon receiving measurement results from the UE and triggers the need for a handover to a target eNB. The source eNB then sends a handover request to the target eNB over the X2 interface. If admission control passes, the target eNB will return with a handover request acknowledgement with the bearers whose resources have been prepared. The UE then switches to the target eNB when handover is confirmed. Subsequent signaling messages occur in order to prevent data loss during handover [7]. In particular, while the uplink traffic travels directly through the target eNB to the S-GW, the downlink traffic from the S-GW will be forwarded by the source eNB to the target eNB before path switching is completed. When the MME receives the path switch request message from the target eNB, it determines the appropriate S-GW. If the same S-GW can continue to serve the target eNB, the MME simply programs the S-GW so that downlink traffic will be switched from the source eNB to the target eNB. On the other hand, if S-GW relocation is needed, the handover signaling flow takes two pairs of message exchange between MME and S-GWs: one pair to set up bearer at the target S-GW and another pair to release bearer at the source S-GW.

An inter-MME handover can happen in the distributed-MME architecture. Fig. 5 shows a scenario where the source eNB learns from its configuration that the handover initiation requires signaling over the S1 interface instead of the X2 interface. First, the source eNB sends a handover required message to the source MME, which then selects and notifies the target MME with a relocation message. Assuming that S-GW relocation is also needed in the distributed-MME architecture, the target MME subsequently programs the bearer for uplink traffic at the target S-GW and exchanges handover request messages to setup the bearer for uplink traffic at the target eNB. After the bearer at the target entity has been setup, the target MME returns an ACK to the source MME, which then issues a handover command so that the UE communicates with the target eNB. At this point, downlink traffic travels through source S-GW, source eNB, target eNB and UE, while the uplink traffic travels from UE through target eNB and target S-GW. After the target MME receives a handover notify message, it exchanges relocation complete messages with the source MME. In response, the source MME sets up a timer to release the resources at the source eNB and source S-GW. The target MME finally programs the bearer for downlink traffic at the target S-GW.

D. Tracking area update

The last important event occurs when a UE moves and detects that its current tracking area (which is identified by a tracking area code that is broadcast in a cell on channel BCCH) is not in the tracking area or list of tracking areas that it has registered with the network. In such a case, the UE updates its new tracking area to the network regardless of whether it is in IDLE or CONNECTED state.

For the case when there is an MME change, the associated signaling flow is depicted in Fig. 6. Here the UE initiates a tracking area update (TAU) message upon detecting a new tracking area. Upon receiving the TAU request from the UE, the eNB selects the MME responsible for the new tracking area. The figure assumes a new target MME, which will then update the location of the UE to the HSS so that any future incoming session can be properly routed to the correct MME. The process includes canceling the location information at the source MME and inserting subscription data at the target MME. Finally, the target MME validates the TAU with the UE.

If the MME is the same serving MME, the MME simply records the UE new location and accepts the TAU. There is no signaling to the HSS in this case. If TAU occurs when the UE is in CONNECTED state, the MME also needs to program the affected S-GWs. However, these message exchanges are already taken into account in a handover and thus are not shown in the figure.

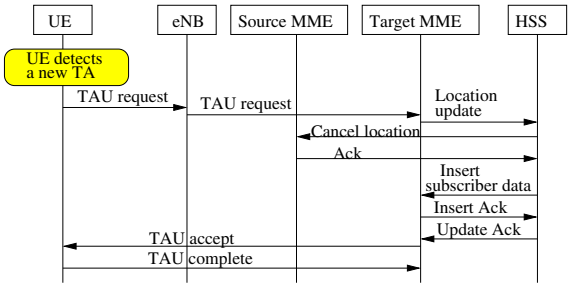

Fig. 6. Inter-MME tracking area update.

III. MME Signaling Load

In this section, we provide a simple analysis to quantify the signaling load at MME due to the four events described in the previous section. We assume that each UE is a smart phone supporting K application types such as voice calling, web browsing, SMS/MMS, email, etc. Let λ_k be the average arrival (originating and terminating) rate of type-k session at a UE (sessions/hour/UE). Since a session can be either originated by a UE or by its peer, we denote $Pr\{O_k\}$ as the probability that a type-k session is originated by a UE. The average type-k session duration is denoted by μ_k^{-1}. Further, let C denote the total number of eNBs in a region under consideration, A denote the area of a cell (consisting of multiple sectors) for a given eNB and ρ denote the UE density in number of UEs per km^2. For simplicity, we assume cells are uniform.

From Fig. 2, the total number of messages (entering and leaving) processed at MME per hour due to UE-originated type-k sessions for both architectures is given by

$$N_o(k) = 10\lambda_k Pr\{O_k\}\rho AC. \quad (1)$$

The UE-terminated case with paging as shown in Fig. 3 requires 11 messages at MME in addition to the paging messages when UE is in IDLE state. The number of messages is 10 if paging is not needed when UE is in CONNECTED state (e.g., when a user receives a phone call while browsing a web). Thus, the total number of messages at the MME per hour for UE-terminated type-k sessions for both architectures is given by

$$\begin{aligned} N_t(k) &= [(11+C_a)R_p p_I + 10(1-p_I)] \\ &\quad \lambda_k(1-Pr\{O_k\})\rho AC, \end{aligned} \quad (2)$$

where C_a is the number of eNBs per tracking area and R_p is the average number of paging transmissions per page (due to paging lost in the backhaul or at air interface). To compute the probability that UE is in IDLE state, p_I, we note that the process $(Xi, Yi,)$, for $i = 1, 2, \cdots$, can represent CONNECTED and IDLE periods with $E[X_i] = \mu^{-1}$ and $E[Y_i] = \lambda^{-1}$. From the theory of alternating renewal process and independence assumption of applications, we have $p_I = \prod_{k=1}^{K} \mu_k/(\lambda_k + \mu_k)$. Similarly, UE is in CONNECTED state if one or more applications are in active session and thus occurring with probability $(1 - p_I)$.

If paging multicast is used instead, the corresponding number of messages at MME is given by

$$N_t(k) = [12R_p p_I + 10(1-p_I)]\lambda_k(1-Pr\{O_k\})\rho AC, \quad (3)$$

To compute MME message load due to handover with average UE velocity of V in km/hr, we adopt a widely-used fluid flow model (e.g., [11]) for computing mobile crossing rate out of an enclosed region with perimeter length L as $R = \rho V L/\pi$. Then, the total number of messages per hour processed at MME for the centralized-MME architecture due to intra-MME handovers is given by

$$N_h^c(k) = [4(1-p_R) + 6\,p_R]R(1-p_I)C, \quad (4)$$

where p_R is the S-GW relocation probability and can be well approximated by $1/\sqrt{C_a}$ if each S-GW serves the size of a tracking area.

For the distributed-MME architecture, we assume for simplicity that each S-GW serves the size of a tracking area. Based on Fig. 4 and Fig. 5, the total number of messages per hour processed at MME is given by

$$N_h^d(k) = [4(1-p_R) + 20\,p_R]R(1-p_I)C. \quad (5)$$

The centralized-MME architecture only involves intra-MME tracking area update. Thus, the total number of messages per hour due to tracking area updates can be approximated by

$$N_a^c = 3RC/\sqrt{C_a}, \quad (6)$$

where the rate of crossing out of a tracking area can be approximated by the rate of crossing out of a cell multiplied by $\sqrt{C_a}$. The number of tracking areas is correspondingly reduced by a factor of C_a.

For the distributed-MME architecture, the total number of messages per hour due to tracking area updates can be given by

$$N_a^d(k) = [3(1-p_R) + 9\,p_R]RC/\sqrt{C_a}. \quad (7)$$

Finally, the total number of messages for the centralized-MME architecture due to all events is simply the sum of (1), (2) with unicast paging or (3) with multicast paging, (4) and (6) over all application types. For the distributed-MME architecture, the corresponding number is given by the sum of (1), (2) with unicast paging or (3) with multicast paging, (5) and (7).

IV. Numerical Results

In this section, we present some numerical results on message processing requirements for both architectures and quantify the benefit of multicast paging. We assume an environment with high user density (3 million attached UEs) in NYC region of size 785 km^2 (land area excluding water). The resulting value of ρ is approximately 3820 UEs/km^2, assuming uniform UE distribution. For user mobility, we assume $V = 20$ km/hr.

We assume four main application types (voice, SMS/MMS, email, and web) with their associated busy-hour parameters given in Table I.

TABLE I
SESSION PARAMETER FOR VARIOUS APPLICATIONS

Application	Session arrival rate	Session duration	$Pr\{O\}$
Voice	0.4	0.05	0.5
SMS/MMS	0.1	negligible	0.4
Email	0.05	negligible	0.2
Web browsing	0.05	0.1	1

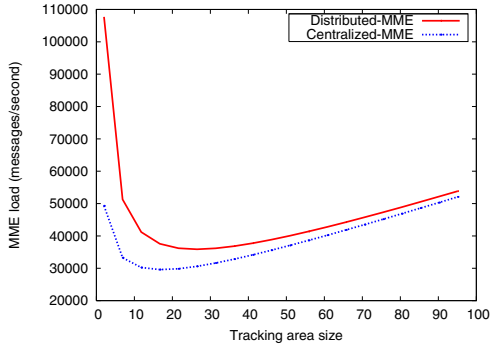

Fig. 7. Comparison of centralized and distributed MMEs.

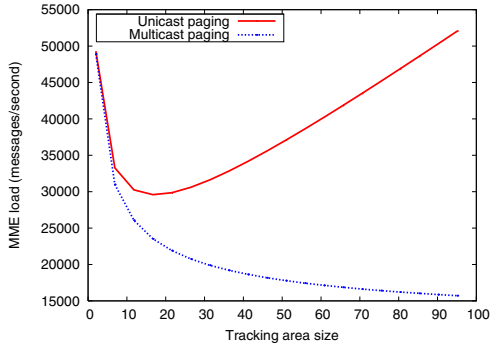

Fig. 8. Comparison of unicast and multicast paging.

Fig. 7 compares the signaling loads for both architectures with unicast paging. We assume that $C = 500$ and $R_p = 1.1$. Assuming uniform circular cells with an overlapping factor γ of 1.2, the required cell radius, r, to cover the entire area is $r = \gamma\sqrt{785/(C\pi)}$ km. Note that the MME load in either architecture initially decreases as C_a increases. This is because the TAU rate decreases rapidly from a very high number when C_a is very small. As C_a increases further, the TAU rate keeps decreasing, but at a slow rate. On the other hand, UE-terminated session rate increases linearly as a function of C_a. Thus, the MME load eventually increases as C_a increases further. The optimal tracking area size can be shown to be proportional to $(V/\sum_k \lambda_k)^{2/3}$. It is interesting to note from the figure that the distributed-MME architecture faces a very high signaling load when Ca is very small. This is because the probability of relocation is very high. When relocations are less frequent, the difference in signaling loads between the two architectures diminishes.

Fig. 8 compares the signaling loads with unicast paging and multicast paging for the centralized-MME architecture. Observe that unlike the plot with unicast paging which defines an optimal number of eNBs, the MME load with multicast paging is a decreasing function of C_a since there is no effect of UE-terminated sessions. Thus, it is beneficial in making the tracking area size much larger with multicast paging.

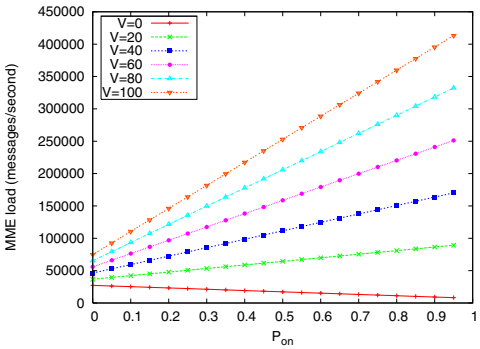

Fig. 9. MME load versus $Pr\{$"always-on"$\}$.

In the future, there will be more "always-on" applications that can potentially keep UEs in CONNECTED state continuously. Intuitively, if UEs are in CONNECTED state, the signaling load would decrease since there is no more paging. On the other hand, the load due to handovers would increase. In Fig. 9, we present the MME signaling load versus the probability that some always-on applications are running on a UE for different UE mobility rates. The graph represents the centralized-MME architecture with unicast paging. As can be seen, always-on applications generally can be detrimental to the MME, except when UEs are non-mobile ($V = 0$). The reason is that while all always-on UEs contribute to handovers, only newly arriving sessions contribute to UE-terminated sessions.

V. CONCLUSION

Distributed-MME architecture incurs higher MME signaling load than the centralized architecture due to inter-MME communication. We have also shown that multicast paging can offer marked improvement over the unicast paging. Since smart phones capable of running a multitude of applications are increasingly more common, these applications may make UEs "always-on". Our results show that these always-on applications can further increase the MME signaling load significantly.

REFERENCES

[1] 3GPP TD RP-040461, "Proposed Study Item on Evolved UTRA and UTRAN", Dec. 2004.
[2] 3GPP TR 25.913, "Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)," Dec. 2008.
[3] Ekstrom et al., "Technical Solutions for the 3G Long-Term Evolution," IEEE Communications Magazine, pp. 38-45, Mar. 2006.
[4] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN)," Mar. 2009.
[5] 3GPP TS 36.420, "X2 General Aspects and Principles," Dec. 2008.
[6] 3GPP TS 36.410, "S1 General Aspects and Principles," Mar 2009.
[7] 3GPP TS 23.401, "GPRS Enhancement for E-UTRAN access", Dec. 2008.
[8] 3GPP TS 36.413, "S1 Application Protocol (S1AP)," Mar. 2009.
[9] TR 33.821, "Rationale and Track of Security Decisions in LTE RAN/3GPP SAE, Mar. 2009.
[10] 3GPP TS 36.304, "User Equipment (UE) Procedures in Idle Mode," Mar. 2009.
[11] D. Lam, D.C. Cox and J. Widom, "Teletraffic Modeling for Personal Communications Services," IEEE Communications Magazine, pp. 79-87, Feb. 1997.