Signaling Rate and Performance for Authentication, Authorization, and Accounting (AAA) Systems in All-IP Cellular Networks

Said Zaghloul, *Student Member, IEEE*, and Admela Jukan, *Senior Member, IEEE*

***Abstract*—Authentication, Authorization, and Accounting (AAA) systems are one of the most significant architectural components of the current cellular networks and within the emerging IP Multimedia Subsystem (IMS) standard. Despite the operators' experience in AAA operations, very little is fundamentally known about the expected signaling rate towards the AAA system. In this paper, using stochastic and renewal theoretic techniques, we develop the first analytical model for the AAA signaling rate as a function of protocol parameters, users' access rates, session durations, and mobility. We provide model approximations and evaluate their accuracy under various operational conditions. Our results show that the AAA signaling rate is a monotonic non-linear function of the mobility rate and asymptotically converges to the AAA signaling rate in fixed networks. We also show that by adjusting the accounting interim and the authorization-lifetime intervals from half to full mean session duration, it is possible to define an AAA operational range for accounting messages that minimizes the signaling rate fluctuations due to likely perturbations in session and mobility statistics. The results also include the effect of the session dropping during handoffs and shows that is marginal in operational networks.**

***Index Terms*—AAA signaling rate, server load, diameter, RADIUS, accounting, stochastic techniques.**

I. Introduction

RECENTLY, several emerging all-IP cellular network technologies, such as WiMAX[1], 3GPP2 EVDO[2], and Long Term Evolution (LTE) [3], have incorporated authentication, authorization, and accounting (AAA) systems in their current and future releases as a substitute to the legacy home location register (HLR) and accounting platforms. Such trend was further boosted by the large adoption of AAA protocols for charging and policy control signaling in the IP Multimedia Subsystem (IMS) standards [4]. In the current systems, AAA plays a crucial role in granting users the required access and in facilitating the collection of accounting data which reflect the users' usage of the network resources. Further on the horizon, with the rapid introduction of new services within the IMS framework and the expected increase in user sessions durations, the signaling traffic towards the AAA system is expected to increase appreciably.

Manuscript received January 24, 2008; revised November 21, 2008; accepted March 2, 2009. The associate editor coordinating the review of this paper and approving it for publication was G. Mandyam.

The authors are with Technische Universität Carolo-Wilhelmina zu Braunschweig (e-mail: {zaghloul, jukan}@ida.ing.tu-bs.de).

Digital Object Identifier 10.1109/TWC.2009.080105

Today, large operators plan the signaling rate of their AAA systems by over-provisioning. This is mainly because under provisioned systems would result in blocking users from access or dropping accounting messages, leading to loss of revenue. Although the growth of the AAA signaling is imminent, which s expected to turn over-provisioning inefficient and hard to scale, alternative design guidelines are currently missing. Such guidelines would also benefit other systems that require the knowledge of projected AAA rates. For instance, testing and certifying equipment performance use AAA protocol settings as their input parameters [5]. The AAA signaling also reflects on the projected sizes of other systems and their configurations, such as Wireless Application Protocol (WAP) gateways and content switches [6], [7]. In such systems, accounting messages are typically used to convey session status information, for instance to facilitate IP address to username mapping. Last but not least, over-provisioning typically results in poor energy and space efficiency in operator's production data centers, which is a growing concern. For all these reasons, it is essential to fundamentally understand the implications on the signaling rate at the AAA system in order to allow for its proper design.

In this study, we develop the first analytical model for the mean AAA signaling rate in fixed and mobile systems using stochastic techniques and renewal theory. Our model exploits commonly accepted concepts of residence and call holding times which were used earlier to evaluate call performance and location update rates in cellular networks [8], [9]. Specifically, we develop a model that incorporates protocol parameters such as the accounting interim interval and the authorization lifetime (see [10]–[12]) and use renewal theoretic techniques to account for the session duration and mobility. We also develop approximations that further simplify the use of the proposed model in practical scenarios. In addition, we investigate the signaling load and on the mean time between accounting updates which has implications on accounting reliability. Due to its closed-form result, our model can help to practically quantify the signaling performance of AAA systems and associated network components, for various session and mobility parameters.

This paper is organized as follows. Section II discusses related work. Section III presents background on AAA protocols relevant to the model. Section IV lists the model's assumptions and presents the AAA analytical framework. Section V presents the model's validation and results. Section VI

concludes the paper and gives directions for further research. We draw the readers' attention to the fact that in this study, we restrict our scope to the AAA performance in regular network operation for post-paid (offline) billing models [10], [13]. In other words, neither the security analysis relative to the authentications nor prepaid charging are covered here, as they deserve dedicated studies that must consider security threats and the users' quota, the debiting rate, etc, which is beyond the scope of this article (see [14] for details).

II. Related Work

A foundational framework that characterizes AAA signaling rate in mobile environments is currently missing in the literature. Several studies have addressed AAA systems from different aspects, including third party AAA applications [15], AAA for 802.16e networks [9], and management extensions for Diameter [16]. In [17]–[19], the authors focus on quota and user account management aspects in prepaid systems for GPRS and UMTS systems. In [20], the authors present a study of authentication mechanisms in wireless networks, with the goal to evaluate the cost of authentication messages per user. To the best of our knowledge, our study is the first to comprehensively analyze AAA signaling analytically, by taking into consideration protocol specifics, service session duration, mobility, and network size. The model presented in this paper is a major extension of our past work in [21], [22].

In our work, we follow the statistical approach of characterizing mobility akin to the cellular residence time in traditional 2G networks, but we deviate from the other models in the solution methodology. Our analysis not only requires consideration of network size and protocol aspects, but also a closed form estimation of the portion of each session spent in an area covered by an access gateway. Past work from the classical circuit switched cellular domain does not to apply in the AAA context and as such requires significant modifications. Early work [8], [23], [24] mainly focused on call blocking and dropping, as well as effective call duration. In [8], the authors provide generic theorems that allow the calculation of the effective call duration when the Laplace transform of the distributions of the call and the residence times can be written in a rational form. In [23], the authors use a Hyper-Erlang model to derive the channel occupancy time for micro and macro cellular environments. Reference [24] uses multidimensional birth-death processes to generically study the cellular system performance, where the channel holding time distributions are not given in closed forms, and as pointed in [23] they require many parameters for statistical fitting of the residence time. Common to all the works is that the call holding time per cell for an active call is either offered in the Laplace domain or in non-closed forms. As such, they may result in large complexity when adapting them to estimate the AAA traffic.

III. AAA Background

The most widely adopted AAA signaling protocols are RADIUS[25] and its successor Diameter[10], [11]. Since both protocols largely incorporate the same AAA signaling message types and protocol procedures, in the following discussion we adopt the message names from Diameter. Figs. 1 and 2 illustrate a generic AAA system architecture and the signaling flow [3], [26]. In this architecture, access gateways (AGW) serve multiple base station areas[1]. When a user initiates a mobile session, the radio access network triggers AAA signaling at the corresponding AGW towards the AAA system, shown in Fig.2. When a session is established (step 1), Diameter authentication exchanges (i.e., AA-Mobile-Node-Request, AMR) are conducted with the AAA system to authenticate and/or authorize the incoming session. The authentication response (i.e., AA-Mobile-Node-Answer, AMA) carries the user's profile and network settings back to the requesting gateway. One of these settings is the Authorization-Lifetime attribute used to indicate the time by which the mobile node must re-authenticate once the Authorization-Lifetime expires. In our example, a re-authentication takes place in step 4. Upon successful authentication, an Accounting Request message, ACR type Start, is sent (step 2). The AAA acknowledges the receipt of the ACR message by sending an accounting answer message (ACA). The accounting ACR Start message is typically followed by periodic ACR type (Interim) messages reporting the latest subscriber's usage every Acct-Interim-Interval (steps 3, 5)[10]. Accounting interim messages are used to periodically meter users' sessions and thus minimize revenue losses should the network suffer from unexpected failures [12]. When a handoff occurs between AGW 1 and 2, the accounting session at the source AGW is terminated with an ACR type Stop message (step 6), while a new accounting session is sent by the target AGW after optionally authenticating the user (step 7). Steps similar to (1-6) take place at the new AGW. Once the session is terminated (step 14), an ACR type (Stop) message is sent reporting the final subscriber's usage.

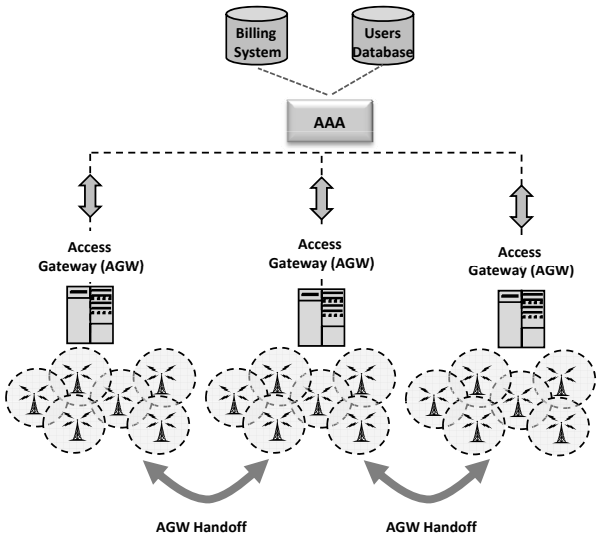

Fig. 1. A simplified "All-IP" system.

[1]AGW is a generic term used to refer to any first IP gateway; examples of AGWs are Access Serving Node Gateway (ASN-GW) in WiMAX, Packet Data Serving Node (PDSN) in 3GPP2 networks, or Serving Gateways in 3GPP Rel6+ systems.

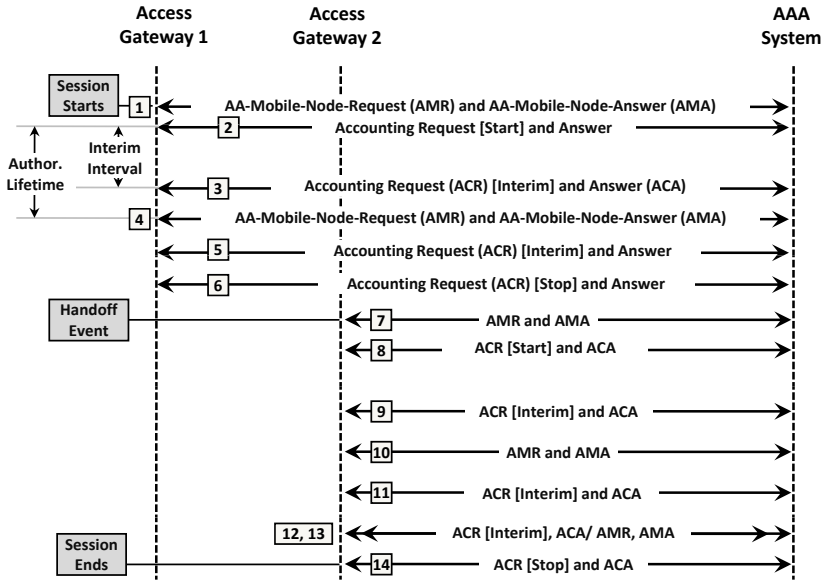

Fig. 2. Typical diameter signaling messages.

IV. Modeling and Analysis

In this section, we start by describing AAA signaling traffic towards the AAA system in fixed (or, very low mobility) scenarios. We then derive a model and an approximation for the signaling traffic in highly mobile scenarios. At the end of the section, we analyze the time between accounting updates.

A. Assumptions

- The new session arrival process from the i^{th} AGW is Poissonian with a mean rate of $\lambda^{(i)}$.
- The session time duration, S, is negative exponentially distributed with an average of E_s. This assumption is used for tractability and is relaxed in the Section V.
- The AGW residence times, R, are independent and identically distributed following the Gamma distribution with a mean of $E_r = k_r\theta_r$. k_r and θ_r are the shape and scale parameters, respectively. Here, the shape parameter is the reciprocal of the coefficient of variation while the scale parameter tells how large the distribution is spread-out[2].
- The packet error rate in the link between the AGWs and the AAA is negligible and hence the packet loss effect on the mean signaling rate is insignificant [22].
- Without loss of generality, reauthentications are always successful for authenticated users.
- The AGW's capacity is very large resulting in negligible blocking. This is realistic as AGW capacities can support a large number of users (e.g., 500,000 [28]).

B. AAA Signaling Rate Model in Fixed Environments

From Fig. 2, the AAA signaling traffic rate consists of the aggregate rate of the authentication, re-authentication, accounting start, accounting interim, and accounting stop messages denoted as ξ_A, ξ_R, ξ_{Start}, ξ_{Int}, and ξ_{Stop} respectively. Let p_a denote the authentication success rate, then the mean signaling rate, $E[\xi]$, can be expressed as,

$$E[\xi] = E[\xi_A] + \big(E[\xi_R] + E[\xi_{Start}] + E[\xi_{Int}] + E[\xi_{Stop}]\big)p_a \quad (1)$$

For a successfully authenticated session[3], if the mobile does not change its AGW during the session, then we will only have one authentication, accounting start, and accounting stop exchanges. From our assumptions, it follows that for any (AGW i), the AAA signaling process can be viewed as a compound Poisson process of the random number of messages sent during the gateway holding times, i.e.,

$$E[\xi_A] = \sum_{i=1}^{N_{AGW}} \lambda^{(i)} = p_a^{-1} E[\xi_{Start}] = p_a^{-1} E[\xi_{Stop}] \quad (2)$$

Notice that we make a distinction between authentications and reauthentications in our model; however in reality, the same AMR message is sent in both cases. Specifically, authentication messages are only triggered when a mobile node enters the service area of an AGW. On the other hand, reauthentication messages are periodically triggered every Authorization-Lifetime time units while the session is being served similar to the accounting interim messages. Thus, the AMR messaging rate is simply the sum of the authentications and the reauthentications rates and can be written as $E[\xi_A] + p_a E[\xi_R]$. Denoting the interim interval as Δ_T and the authorization lifetime as Δ_M, the mean number of interim and reauthentication messages in a session, S, are,

$$E[\xi_{Int}] = \sum_{i=1}^{N_{AGW}} \lambda^{(i)} \lfloor \frac{S}{\Delta_T} \rfloor \,,\; E[\xi_R] = \sum_{i=1}^{N_{AGW}} \lambda^{(i)} \lfloor \frac{S}{\Delta_M} \rfloor \quad (3)$$

[2]Similar to [20], we choose the Gamma distribution for the AGW residence time as it is known to offer a good approximation for the lognormal distribution [27], the widely encountered distribution for cell residence times from field measurements [8].

[3]Without loss of generality, we assume the authentication scheme used in 3GPP2 systems as in [2]. Other authentication schemes can be simply incorporated by multiplying $E[\xi_A]$ and $E[\xi_R]$ by a constant reflecting the number of used messages.

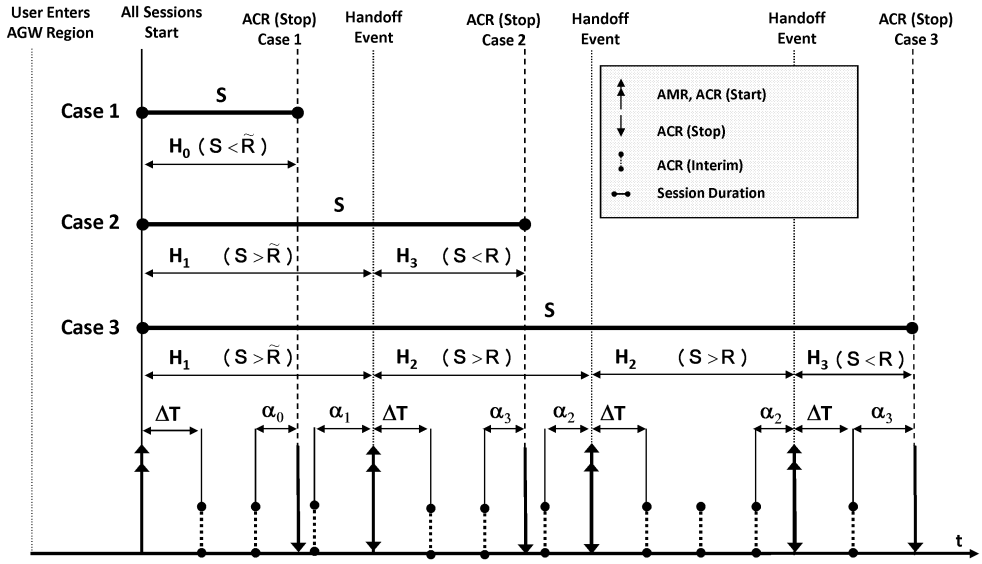

Fig. 3. Diameter signaling traffic model [note that *Reauthentications* (AMR) are omitted for clarity].

Taking the expectation of eq.3 and calculating the mean number of interims as described in Appendix A, eq. 27, it can be shown that the mean number of the interims and reauthentications in a session can be calculated as the infinite sum of the complementary function, $\bar{F}_S(s)$, evaluated at the discrete points ($n\Delta_T$ and $n\Delta_M$). For exponentially distributed sessions we get,

$$E[\xi_{Int}] = \sum_{i=1}^{N_{AGW}} \lambda^{(i)} \sum_{n=1}^{\infty} \bar{F}_S(n\Delta_T) = \sum_{i=1}^{N_{AGW}} \lambda^{(i)} \frac{1}{e^{\Delta_T/E_s} - 1}$$

$$E[\xi_R] = \sum_{i=1}^{N_{AGW}} \lambda^{(i)} \frac{1}{e^{\Delta_M/E_s} - 1} \quad (4)$$

Thus, the total AAA signaling rate in fixed environments is,

$$E[\xi] = \sum_{i=1}^{N_{AGW}} \lambda^{(i)} \left[1 + p_a \left(2 + \frac{1}{e^{\frac{\Delta_T}{E_s}} - 1} + \frac{1}{e^{\frac{\Delta_M}{E_s}} - 1} \right) \right] \quad (5)$$

C. AAA Signaling Rate Model in Mobile Environments

Now that we have characterized the AAA signaling rate in fixed environments, we extend the result in eq.5 to include the effect of mobility. To do so, we define the AGW's session holding time H as the time from a session start or a handoff event until the next handoff event or the session termination. Based on [8], [23], [24] we also define the AGW residence time, R, as the time the user spends in an access gateway area irrespective of the session's activity; the residual of the residence time is denoted as $\tilde{R}$. As shown in Fig. 3, any user's session, S, falls under one of the three following categories with respect to residence time: (1) no handoffs (2) only one handoff (3) multiple handoffs. Case (1) occurs when the session duration is shorter than the remaining time for the user to leave the AGW region (i.e., residual residence time). This is commonly assumed in the literature [8], [23] because the moments when users emerge into an area and when they initiate sessions are not necessarily aligned. Cases (2) and (3) occur if the user makes at least one AGW handoff during her session. Notice that we consider such cases for analytical purposes in order to obtain the distribution of the AGW session holding time (i.e., $H_i, i \in 0, .., 3$). If there are no handoffs in the session (i.e., case 1), the AGW session holding time (i.e., H_0 here) is the conditional duration of the user's session being smaller or equal to the residual residence time (i.e., $S \leq \tilde{R}$). If the session includes only one handoff (i.e., case 2), then we have two distinctive types of the AGW session holding times namely, before a handoff H_1 and, afterwards, H_3. Thus, H_1 is the conditional duration of the residual residence time is less than or equal to session time (i.e., $\tilde{R} \leq S$), whereas H_3 corresponds to $S < R$. Because of the memoryless property of the exponential distribution, the residual of the session time statistically equals the session duration. Thus, if the session includes multiple handoffs (i.e., case 3), we generalize case 2 by adding a new type of a session holding time (H_2) where the conditional duration of the session is greater than the residence time (i.e., $S > R$). The number of the H_2 periods is then one less than the number of handoffs, K (i.e., $K - 1$). Fig. 3 also shows two important parameters used in our model, i.e., the interim-interval, Δ_T, and the times between the last interim interval and an accounting stop for a given AGW session holding time, denoted by α_i. The latter parameter will be used for the derivation of the time between accounting updates. Furthermore, since there is only one message of the type (authentications, accounting starts and stops) in a given session, it follows that at steady state, the mean rate of these types is approximately equal. This is because for operational networks the authentication success rate is around unity (i.e., $p_a \approx 1$) and hence accounting start and stop messages are almost always generated for successfully authenticated users. Thus, the rates of these messages is given as,

$$E[\xi_A] = \sum_{i=1}^{N_{\text{AGWs}}} (E[K] + 1) \lambda^{(i)}$$
$$= p_a^{-1} E[\xi_{Start}] = p_a^{-1} E[\xi_{Stop}] \quad (6)$$

At this point, we turn our attention to the evaluation of the interim and reauthentication rates. To do so, we first evaluate

the distributions of the AGW session holding times, H_i, for the three session cases in Fig. 3 and use the results to find the number of interims and reauthentications.

1) Case 1 (No handoffs): In this case, we have only one part of the AGW session holding time, H_0, where the session time is less than (or equal to) the residual of the residence time (i.e., $S \leq \tilde{R}$). The residual of the residence time, $\tilde{R}$, is,

$$f_{\tilde{R}}(\tilde{r}) = \frac{\bar{F}_R(\tilde{r})}{E_r} = \frac{\Gamma\left(k_r, \frac{\tilde{r}}{\theta_r}\right)}{E_r \Gamma(k_r)} \tag{7}$$

Similar to [8], we assume that the residual time, just like the user's mobility, is an independent random process which does not align with the user's session initiation. The distribution of H_0 is found by integrating the joint probability of S and $\tilde{R}$ in the region limited by h_0 and by dividing the result by the probability that $(S \leq \tilde{R})$ (see Appendix B eq.31) as,

$$\begin{aligned} F_{H_0}(h_0) &= Pr\left(H_0 \leq h_0 \mid S \leq \tilde{R}\right) \\ &= \frac{\int_0^{\infty} \int_0^{min(y,h_0)} f_S(x)\, dx f_{\tilde{R}}(y)\, dy}{Pr\left(S \leq \tilde{R}\right)} \end{aligned} \tag{8}$$

To evaluate $F_{H_0}(h_0)$, we note that,

$$\begin{aligned} &\int_0^{\infty} \int_0^{min(y,h_0)} f_S(x)\, dx f_{\tilde{R}}(y)\, dy \\ &= \int_0^{h_0} \int_0^{y} f_S(x)\, dx f_{\tilde{R}}(y)\, dy + \int_{h_0}^{\infty} \int_0^{h_0} f_S(x)\, dx f_{\tilde{R}}(y)\, dy \\ &= \int_0^{h_0} F_S(y)\, dx f_{\tilde{R}}(y)\, dy + \int_{h_0}^{\infty} F_S(h_0)\, dx f_{\tilde{R}}(y)\, dy \end{aligned}$$

Using eqs.(34,35) from Appendix C, then integrating by parts it can be shown that $F_{H_0}(h_0)$ is,

$$\begin{aligned} F_{H_0}(h_0) &= \frac{B_0 - E_s\left(\frac{\theta_h}{\theta_r}\right)^{k_r} \Gamma\left(k_r, \frac{h_0}{\theta_h}\right)}{B_0} \\ &+ e^{\frac{-h_0}{E_s}} \frac{(h_0 + E_s)\,\Gamma\left(k_r, \frac{h_0}{\theta_r}\right) - \theta_r \Gamma\left(k_r + 1, \frac{h_0}{\theta_r}\right)}{B_0} \\ B_0 &= \left[E_r + E_s\left(\left(\frac{\theta_h}{\theta_r}\right)^{k_r} - 1\right)\right] \Gamma(k_r)\ ,\ \theta_h = \frac{\theta_r E_s}{E_s + \theta_r} \end{aligned} \tag{9}$$

The number of interims in H_0 can be written as $I_{H_0} = \lfloor \frac{H_0}{\Delta_T} \rfloor$. Taking the expectation of I_{H_0} using eq.27 (see Appendix A), it can be shown that the mean number of the interims in H_0 can be calculated as the infinite sum of the complementary function, $\bar{F}_{H_0}(n\Delta_T)$ as,

$$E[I_{H_0}] = \sum_{n=1}^{\infty} \bar{F}_{H_0}(n\Delta_T) \tag{10}$$

2) Case 2 (One handoff): As shown in Fig. 3, in this case we consider two time periods: one before the handoff (i.e., H_1) and another after the handoff event (i.e., H_3). The number of interims in this case is the sum of interims in both periods. Similar to eq. 9, we write the distribution of H_1 as the integration of the joint probability of S and $\tilde{R}$ in the region limited by h_1 and dividing the result by the probability that $(\tilde{R} \leq S)$ given as $Pr\left(\tilde{R} \leq S\right) = \int_0^{\infty} \int_0^{y} f_{\tilde{R}}(x)\, dx f_S(y)\, dy$,

$$\begin{aligned} F_{H_1}(h_1) &= Pr\left(H_1 \leq h_1 \mid \tilde{R} \leq S\right) \\ &= \frac{\int_0^{\infty} \int_0^{min(y,h_1)} f_{\tilde{R}}(x)\, dx f_S(y)\, dy}{Pr\left(\tilde{R} \leq S\right)} \end{aligned}$$

Integrating by parts and using eqs.(34-35) from Appendix C, $F_{H_1}(h_1)$ can be written as,

$$\begin{aligned} F_{H_1}(h_1) &= B_1^{-1}\left(\frac{\theta_h}{\theta_r}\right)^{k_r} \gamma\left(k_r, \frac{h_1}{\theta_h}\right) - \Gamma(k_r) + \\ &e^{\frac{-h_1}{E_s}} \Gamma\left(k_r, \frac{h_1}{\theta_r}\right)\ ,\ B_1 = \Gamma(k_r)\left(\left(\frac{\theta_h}{\theta_r}\right)^{k_r} - 1\right) \end{aligned} \tag{11}$$

Similarly, we write the distribution of H_3 as the integration of the joint probability of S and R in the region limited by h_3 and dividing the result by the probability that $(S \leq R)$ given as $Pr(S \leq R) = \int_0^{\infty} \int_0^{y} f_S(x)\, dx f_R(y)\, dy$.

$$\begin{aligned} F_{H_3}(h_3) &= Pr(H_3 \leq h_3 \mid S \leq R) \\ &= \frac{\int_0^{\infty} \int_0^{min(y,h_3)} f_S(x)\, dx f_R(y)\, dy}{Pr(S \leq R)} \end{aligned}$$

Carrying the integrations in eq.12, it can be shown that $F_{H_3}(h_3) = F_{H_1}(h_3)$. Thus, similar to the derivation of eq.10, it follows that the mean number of interims are,

$$E[I_{H_1}] = E[I_{H_3}] = \sum_{n=1}^{\infty} \bar{F}_{H_1}(n\Delta_T) \tag{12}$$

3) Case 3 (Multiple handoffs): This case is a generalization for case 2 as it includes three typical types of holding times: H_1, H_2, and H_3 (see Fig. 3). It is required because in a single handoff the mobile does not spend the full residence time at any AGW, which requires an evaluation of the distribution of H_2; the latter occurs only after spending the full residence time of an AGW, as well as the number of handoffs, K, since we have $K+1$ holding time periods. In a similar fashion to H_1, the AGW session holding duration, H_2 is expressed as the integration of the joint probability of S and R in the region limited by h_2 and dividing the result by the probability that $(R \leq S)$ given as $Pr(R \leq S) = \int_0^{\infty} \int_0^{y} f_R(x)\, dx f_S(y)\, dy$ as,

$$\begin{aligned} F_{H_2}(h_2) &= Pr(H_2 \leq h_2 \mid R \leq S) \\ &= \frac{\int_0^{\infty} \int_0^{min(y,h_2)} f_R(x)\, dx f_S(y)\, dy}{Pr(R \leq S)} \end{aligned}$$

Using integration by parts, it can be shown that H_2 follows the Gamma distribution with shape and scale parameters of k_r and θ_h, as $F_{H_2}(h_2) = \frac{\gamma\left(k_r, \frac{h_2}{\theta_h}\right)}{\Gamma(k_r)}$. Consequently, the mean number of interims in H_2 is given as,

$$E[I_{H_2}] = \sum_{n=1}^{\infty} \bar{F}_{H_2}(n\Delta_T) \tag{13}$$

Thus, the average number of interims $E[I]$ in an arbitrary session S can be evaluated by combining the results from all

the three cases weighted by their expected number of occurrences. Let the probability of no handoffs be $p_0 = Pr\,(K = 0)$ as defined in eq.32. Then it follows that,

$$\begin{aligned} E\,[I] &= p_0 E\,[I_{H_0}] + 2\,(1-p_0)\,E\,[I_{H_1}] \\ &+ (E\,[K] - 1 + p_0)\,E\,[I_{H_2}] \end{aligned} \tag{14}$$

Similar to eq.6, using the compound Poisson model of arrivals from all AGWs, the aggregate interim rate is found by evaluating $E\,[I]$ from all AGWs as,

$$E\,[\xi_{Int}] = \sum_{i=1}^{N_{\text{AGWs}}} \lambda^{(i)} E\,[I] \tag{15}$$

Notice that the same analysis for the interims applies to the number of reauthentications, M, by replacing Δ_T in all terms in eq.14 by Δ_M. Thus, the reauthentication rate, $E\,[\xi_R]$, is,

$$E\,[\xi_R] = \sum_{i=1}^{N_{\text{AGWs}}} \lambda^{(i)} E\,[M] \tag{16}$$

The only left parameter to evaluate the interim and reauthentication signaling rates is the distribution of the number of handoffs, K. It can be shown (see Appendix B), that the probability of the number of handoffs in a session is,

$$Pr\,(K = k) = \begin{cases} 1 + \dfrac{\left(\left(\frac{\theta_h}{\theta_r}\right)^{k_r} - 1\right) E_s}{E_r} & k = 0 \\ G\,(k) - G\,(k+1) & k \geq 1 \end{cases} \tag{17}$$

*w*here $G(k)$ is defined in Appendix B, eq.29. From the above, the mean number of handoffs can be written as,

$$E\,[K] = \sum_{n=0}^{\infty} nPr(K = n) = \sum_{n=1}^{\infty} G(K = n) = E_s E_r^{-1} \tag{18}$$

Substituting eq.18 into eqs.(15-16), the mean interim and reauthentication rates are evaluated. Therefore, the mean AAA signaling rate can be obtained by substituting eqs.(6, 15, 16) into eq.1 as,

$$\begin{aligned} E\,[\xi] = \sum_{i=1}^{N_{\text{AGWs}}} \lambda^{(i)} \Big[&(1 + 2p_a)\,(E\,[K] + 1) \\ &+ p_a\,(E\,[I] + E\,[M]) \Big] \end{aligned} \tag{19}$$

As will be shown in the results, an approximation of eq.19 by assuming exponential residence times may be reasonable. In this case, due to the memoryless property, $H_0 = H_1 = H_2$ and is exponentially distributed. Hence, the number of interim messages in such periods can be found in closed form as in eq.5. It follows that the AAA signaling rate can be written as,

$$\begin{aligned} E\,[\xi] = \sum_{i=1}^{N_{\text{AGWs}}} \lambda^{(i)}\,(E\,[K] + 1) \Bigg[1 + p_a \Big(2 + \frac{1}{e^{\frac{\Delta_T}{E_H}} - 1} \\ + \frac{1}{e^{\frac{\Delta_M}{E_H}} - 1} \Big) \Bigg], \; E_H = \frac{E_s}{E\,[K] + 1} \end{aligned} \tag{20}$$

In cases where context transfer is used between AGWs to facilitate handoffs, the authentication upon handoffs is no longer necessary as only one authentication message is sent per session. If reauthentications are triggered based on the session start time rather than the handoff moments, then the number of authentication and reauthentications is obtained using the fixed model in eq.4, otherwise the reauthentications rate is found using eq.16 as $\sum_{i=1}^{N_{\text{AGWs}}} \lambda^{(i)}\,(1 + E[M])$. Since context transfer signaling takes place in the core IP network, the effect of signaling packet losses is here insignificant. In the following subsections, we address three special scenarios for applicability of the derived model.

D. The Effect of Session Dropping on the AAA Signaling Rate

In some cases, sessions are dropped during handoffs due to excessive handoff delays or due to other factors such as unavailability of wireless resources. Such effects can be generically incorporated in the model in eq.19, by using the likelihood function of session dropping, denoted as ϱ. For instance, for excessively long handoff delays (i.e., longer than d_a time units), the session is dropped with a probability of $\varrho = \Pr(d > d_a)$. The probability ϱ can be obtained from available analytical models such as [29], [30] or from measurements, as the handoff delay highly depends on the access technology and the used handoff mechanisms. For instance, according to [31], [32], the handoff delay is given as the sum of the AltPPP Sync, AltPPP Request, AltPPP Reply, and ICMPv6 Router Advertisement messages. In EVDO systems, observations showed air link latencies of 99 ms and a standard deviation of 48 ms[33]. Since each message is sufficiently small to fit in one radio frame, and assuming a typical 50 ms handoff delay at the EVDO layer, 10ms RTT delay between the AAA system and the AGW, and 10 ms for authentication, the resulting mean delay is 466 with a standard deviation of 100ms. Using moment matching and assuming a Gamma fit, it can be obtained $\varrho = \Gamma\,(k_0, d_a/\theta_0)/\Gamma(k_0), k_d = (466/100)^2 = 21.72, \theta_0 = 466/k_d = 21.46$.

By going back to Fig.3 and considering each of the three cases separately, once for a complete session and another for an incomplete one using the session dropping probability ϱ, the following results can be derived. For case 1, no modification is needed since the sessions are not dropped (no handoffs). For case 2 (i.e., one handoff), if the session is dropped, then the first period H_1 occurs with probability $Pr(K \geq 1)(1-\varrho) = G(1)(1-\varrho)$, or else two periods H_1, H_1 occur with probability $Pr(K = 1)\varrho = (G(1) - G(2))\varrho$ (see eq.17). In case 3 (multiple handoffs), for an incomplete session dropping at the m^{th} handoff, the period H_1 is followed by $(m-1)$ H_2 periods. Thus, it can be shown that the probability of dropping at the m^{th} handoff is $Pr(K \geq m)\varrho^{m-1}(1-\varrho) = G(m)\varrho^{m-1}(1-\varrho)$. For a session including m handoffs the probability is $Pr(K = m)\varrho^m = (G(m) - G(m+1))\varrho^m$. Consequently, the effective number of handoffs $E[K_e]$ with session dropping is,

$$\begin{aligned} E[K_e] &= \sum_{k=1}^{\infty} k\Big((G(k) - G(k+1))\varrho^k + G(k)\varrho^{k-1}(1-\varrho) \Big) \\ &= \frac{E_s}{E_r} \frac{1 - \left(\frac{\theta_h}{\theta_r}\right)^{k_r}}{1 - \left(\frac{\theta_h}{\theta_r}\right)^{k_r} \varrho} \end{aligned}$$

The last period H_1 will occur only if the session is not dropped. In mathematical form we have,

$$Pr\left(\textit{Last period exists}\right) = p_l = \sum_{k=1}^{\infty} \left(G(k) - G(k+1)\right) \varrho^k = \left(1 - \left(\theta_h \theta_r^{-1}\right)^{k_r}\right) E[K_e] \quad (21)$$

Since for k handoffs, we have $(k-1)$ H_2 periods, the mean number of H_2 periods is,

$$E[N_{H_2}] = \sum_{k=1}^{\infty} (k-1) \Bigg(\left(G(k) - G(k+1)\right)\varrho^k + G(k)\varrho^{k-1}(1-\varrho) \Bigg) = E[K_e] - (1 - p_0) \quad (22)$$

Thus, the mean number of interim messages is given as,

$$E[I] = p_0 E[I_{H_0}] + (1 - p_0 + p_l) E[I_{H_1}] + E[N_{H_2}] E[I_{H_2}]$$

The mean number of reauthentications, $E[M]$, can be evaluated similar to $E[I]$ by using Δ_M instead of Δ_T when evaluating $E[I_{H_0}]$, $E[I_{H_1}]$, and $E[I_{H_2}]$ respectively, i.e.,

$$E\left[\xi\right] = \sum_{i=1}^{N_{\text{AGWs}}} \lambda^{(i)} \Bigg[\left(1 + 2p_a\right) \left(E\left[K_e\right] + 1\right) + p_a \left(E\left[I\right] + E\left[M\right]\right) \Bigg] \quad (23)$$

E. The Effect of Roaming Users

The model presented so far captures the signaling rate due to home users, in line with the current AAA architecture which mainly incorporates home user's statistic, regardless of whether they are served by the home or the foreign networks. Although the signaling rate due users from the roaming partners is insignificant relative to that due to the home users, where the model in eq.19 can serve as a good approximation, the question arises of how the model can capture an increase in the number of roaming users. This may be of particular interest in specialized scenarios, such as Mobile Virtual Network Operators (MVNO), where third parties with no wireless infrastructure can offer services by using the wireless infrastructures from various wireless operators. In these scenarios, the same AAA signaling system maybe be used to capture roaming and home users, and the overall rate becomes a complex function of the mobility patterns and the number of access gateways. To illustrate this effect, let us consider a scenario with two operators and one MVNO, referred to O1, O2 and OM; here, the goal is to asses the signaling volume at the AAA system belonging to operator O1. For O1-home subscribers, the AAA system *always* receives the authentication and accounting signaling either directly from access gateways within O1 or from O2's AAA systems, when the O1 subscribers roam into O2. Thus, the model in eq.19 always applies. On the other hand, the AAA signaling from roaming users (i.e., network O2's users or the MVNO users served by network O1) is received by the AAA system only when they are served by network O1. Hence, only parts of the AAA signaling pertaining the roaming users' sessions are received by the AAA system in O1. Thus, even if we hypothetically assume that the session arrival rates from the home and roaming users are equal, the AAA signaling pertaining to roaming users will always be significantly less than that for home users. The analytical treatment of the roaming users is rather different and requires extending the model in eq.19 and intertwining it with a mobility model to account for the possibility of leaving the network (see the preliminary work in [34] where a transient Markovian approach was used). To this end, Section V shows preliminary simulation results.

F. The Mean Time Between Accounting Updates

Interim messages are mainly used to meter the service usage in realtime, and thus protect against revenue losses if an AGW fails during the service lifetime. The interim interval, Δ_T, represents the maximum interval where the system is under risk, since the usage is not reported until the end of the interim interval. The larger the interim interval, the larger is the risk, and hence the smaller the signaling rate. Since mobility results in accounting stop messages, mobility can be viewed as a risk alleviating process as it decreases the observed risk interval, κ. A dual argument can also be made for security by considering the time between AMR messages in a similar fashion. The mean update interval is also important for dimensioning content switches and WAP gateways. Thus, $1/\left(E\left[\kappa\right]\right)$ represents the mean signaling rate from the AAA system to such switches. At any instant, the next accounting update can be an interim, an accounting stop due to a handoff, or an accounting stop due to the session termination. From Fig.3, we observe five distinct update interval types (Δ_T and $\alpha_i, i \in \{0,..,3\}$) with different probabilities of occurrence (i.e., p_{Δ_T} and p_{α_i}). Thus using a composition model, the mean update interval is,

$$E\left[\kappa\right] = \Delta_T p_{\Delta_T} + \sum_{i=0}^{3} E\left[\alpha_i\right] p_{\alpha_i} \quad (24)$$

The mean period until next update, $E\left[\alpha_j\right]$, is equal to the remaining time in the AGW holding time, H_j, after the last interim update. This can be expressed as $E\left[\alpha_j\right] = E\left[H_j\right] - E\left[I_{H_j}\right]\Delta_T$. The means of H_j can be obtained by integrating by parts their complementary distributions, i.e.,

$$E\left[H_0\right] = \frac{\Gamma\left[k_r\right] Es}{B_0 \theta_r^{kr} (E_s + \theta_r)^{2k_r+1}} \Bigg(2E_s^2 \left(E_s \theta_r \left(E_s + \theta_r\right)\right)^{k_r} + E_s \theta_r \left(k_r + 2\right) \left(E_s \theta_r \left(E_s + \theta_r\right)\right)^{k_r} + k_r \theta_r^{k_r+2} (E_s + \theta_r)^{2k_r} + E_s \left(\theta_r (E_s + \theta_r)^2\right)^{k_r} \left((k_r - 2)\theta_r - 2E_s\right) \Bigg)$$

$$E\left[H_1\right] = E\left[H_3\right] = E_s + \frac{E_r \left(\theta_h/\theta_r\right)^{k_r+1}}{\left(\theta_h/\theta_r\right)^{k_r} - 1}$$

$$E\left[H_2\right] = k_r \theta_h \quad (25)$$

Finally, the probabilities of each update interval (i.e., Δ_T and α_i) can be approximated by the ratio of the corresponding mean number of update intervals in a session and the total mean number of all updates during the session scaled by its probability of occurrence. Since we have $\left(E\left[K\right] + 1\right)$ accounting stops due to the k handoffs and session termination plus $E\left[I\right]$ interim intervals, the total number of update

intervals within a session holding time is $(E[I]+E[K]+1)$. Therefore, the probability of the update interval of Δ_T is given as $E[I]C_0$ where $C_0=(1+E[I]+E[K])^{-1}$. Since the periods α_0, α_1, and α_3 occur only once within a session holding time, where α_0 occurs if no handoffs take place while α_1 and α_3 appear otherwise, their corresponding likelihoods of occurrence are p_0C_0 and $(1-p_0)C_0$ respectively. Since these probabilities sum to unity (i.e., $p_{\Delta_T}+\sum_{i=0}^{3}p_{\alpha_i}=1$), we have,

$$p_{\Delta_T}\approx E[I]C_0 \quad , \quad p_{\alpha_0}\approx p_0C_0 \tag{26}$$
$$p_{\alpha_1}=p_{\alpha_3}\approx(1-p_0)C_0\ ,\ p_{\alpha_2}\approx 1-(p_{\Delta_T}+p_{\alpha_0}+2p_{\alpha_1})$$

Thus, by substituting eqs.(25-26) into eq.24, we obtain the mean update interval, $E[\kappa]$.

V. Model Validation and Results

The goal of this section is to validate the model by simulations and to study the effects of the residence time and the interim interval on the total signaling rate. The simulations are based on a C++ event driven simulator developed to generate the users' requests for a given area covered by 5 AGWs arranged in a torus. Each AGW serves a rectangular area of A x B cells. Each cell has a lognormally distributed residence time to match findings from field results [20]. We simulate macro cell sizes with users following fully random mobility between the cells [26]. The number of cells and the parameters of the residence times are varied to reflect different AGW residence times. When users initiate sessions, an authentication message is sent by the serving AGW to the AAA server. If the authentication is successful (i.e., by tossing a random variable and comparing to p_a), an accounting start message is scheduled to be sent after a short delay of the order of roundtrip time between the AAA and the AGW systems. The session duration is generated following either Exponential or lognormal distributions. Successive residence times representing handoff instants are obtained by carrying out movements between cells in the AGW areas until the session duration is exceeded. At the handoff moments, accounting stop and start, and authentication messages are scheduled to be sent. Accounting interims and reauthentications are also scheduled between handoff events. Simulation results are compared to the analysis by finding the parameters for the AGW residence time through matching the first two moments.

In Fig. 4, we show the effect of the residence time on the mean AAA signaling rate. We compare simulation results with the analytical, approximate, and fixed models in eq.19, eq.20, and eq.5. The analytical results (lines) match simulation results (dots) within $< 2\%$ error. The approximate model also gives good estimates ($< 5\%$ error). This comes from the observation that changing the coefficient of variation ($C_R = k_r^{-0.5}$) for the residence time does not change the resulting signaling rate considerably, suggesting that exponential approximations for the residence time perform reasonably well. We also notice that as the residence time to session duration ratio increases, the signaling rate approaches the fixed rate model asymptotically. This is because when residence times are large, handoffs are unlikely and hence the fixed rate model applies.

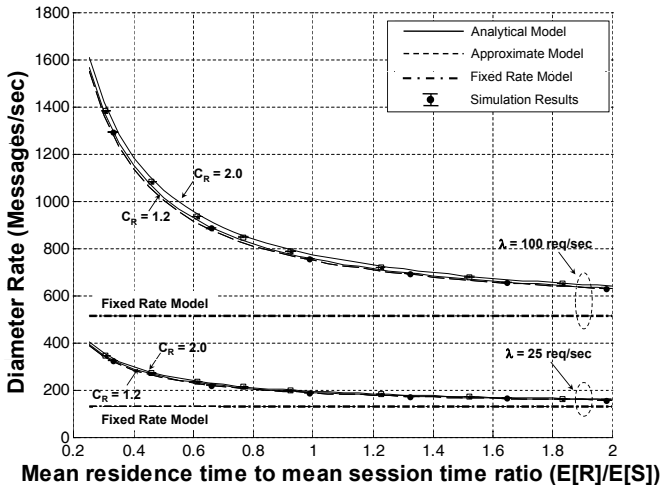

Fig. 4. Residence time effect on the mean signaling rate. Simulation parameters [5 AGWs, $E_S=\Delta_M=40$ min, $\Delta_T=20$ min, 5×5 cells per AGW with residence times varying from 2.5 - 15 mins per cell (lognormal coeff. of var. $\in\{2,3\}$), mean batch method, 30 batches, 10 hr long simulation, 95% confidence (error bars are within the marker sizes)].

In high mobility scenarios, ignoring mobility can therefore result in large errors.

In Fig.5, we show the effect of changing the mean residence time duration characterizing mobility on the authentication and the accounting traffic loads for two different values of the authorization lifetime. We see the traffic split and observe that the accounting traffic load is higher than the authentication traffic load for the authentication scheme used in this paper, based on [2]. For practical authorization lifetime settings, the number of accounting messages is usually larger than authentication messages. As observed in Fig.4, we also see that using the fixed model results in a large estimation error when the ratio of the mean residence time to the mean session duration is low (i.e., high mobility). We also show results for the case of context transfer. If reauthentications are triggered based on handoff instants rather than the session start time, the corresponding authentication rate approaches that of the fixed model. This is because when the E_r/E_s ratio is low, reauthentications are barely triggered. However when $E_r > E_s$, the number of re-authentications is limited by the session duration rather than the residence time. We clearly see for the cases $\Delta_M = E_s$ and $\Delta_M = .5E_s$ that the model for fixed networks (see eq.4) can be practically used to estimate the authentications rate if context transfers are used.

The effect of the Interim-Interval, Δ_T, on the mean signaling rate is shown in Fig. 6. We observe that, irrespective of the residence times, the signaling rate decreases with the increasing interim interval. Notice that setting Δ_T to values below $0.25E_s$ can result in a very large increase in the signaling rate, while setting it too low may defeat its purpose of protection against billing record losses due to unexpected failures. In all cases in Fig. 6, we observe that ranges from $0.5E_s$ to E_s offer an appropriate tradeoff between signaling rate and reliability, which is a result of significant practical importance to the mobile operators. Finally, notice that networks with low AGW residence times become insensitive to Δ_T faster than those with higher residence times. This is due to the fact that

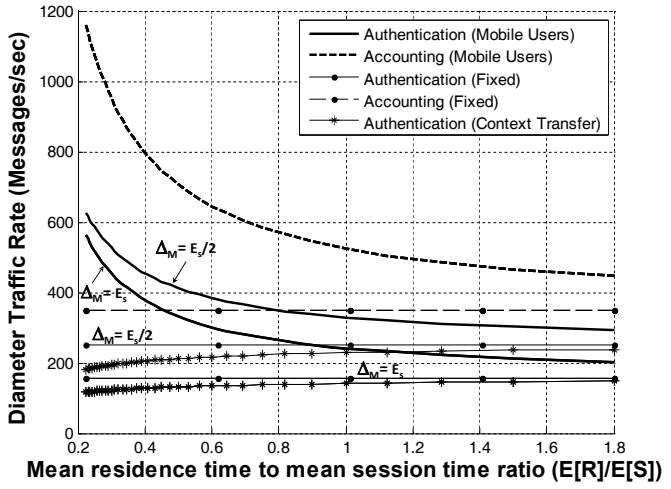

Fig. 5. Numerical results for the residence time effect on the mean signaling rate for the authentication and the accounting traffic. [E_s = 40 min, C_R = 2, $\Delta_T = E_s/2$, $\Delta_M \in \{E_s/2, E_s\}$].

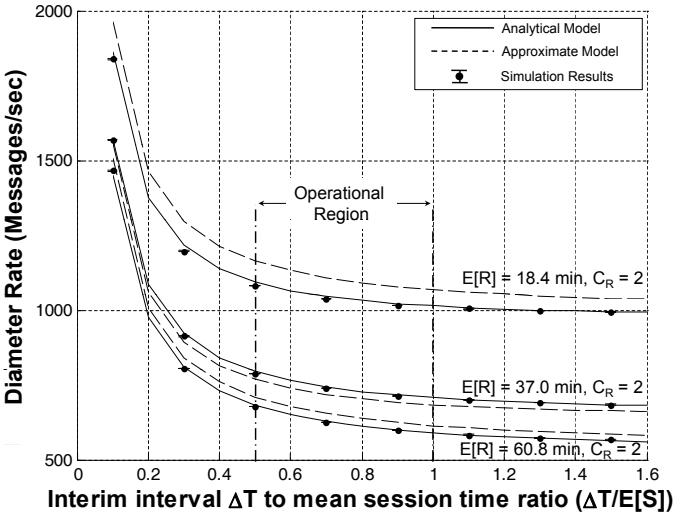

Fig. 6. Numerical results for the Interim-Interval effect on the mean AAA signaling rate. Simulation parameters [5 AGWs, $\lambda^{(i,j)} = 20$ req/sec, $E_S = \Delta_M = 40$ min, 5×5 cells per AGW with residence times varying from 2.5 - 15 mins per cell (lognormal coeff. of var. = 3), mean batch method, 30 batches, 10 hr long simulation, 95% confidence (error bars are within the marker sizes)].

in such networks much less interim updates are triggered as Δ_T increases. Similar results can be obtained by varying the authorization lifetime.

The effect of the mobility on the billing update interval is shown in Fig.7. Clearly, the increased mobility (i.e., smaller residence times) results in a reduced billing update interval and hence in an increased billing reliability. However, shorter update intervals may reflect on higher capacity requirements for gateways relying on such accounting updates as well as on higher AAA system capacity (see Fig. 4). This is due to the fact that as Δ_T increases beyond the mean residence time, E_r, the mean number of interims per session holding time ($E[I_{H_i}]$) decreases (see eqs. (3, 12, 13)). It follows that p_{Δ_T} decreases and hence the mean update interval becomes shorter as shown in Fig.7. Using the same reasoning, one observes that as Δ_T/E_r ratio decreases, the interim updates will dominate

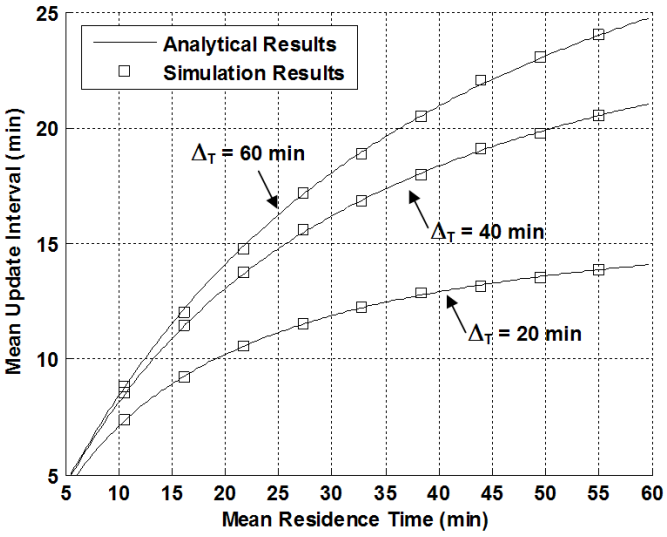

Fig. 7. The mobility effect on the mean update interval $E[\kappa]$. Simulation parameters [$E_S = \Delta_M = 40$ min, $k_r = 0.25$, 100,000 sessions, 5 simulation runs, 95% confidence levels (error bars are within marker's size)].

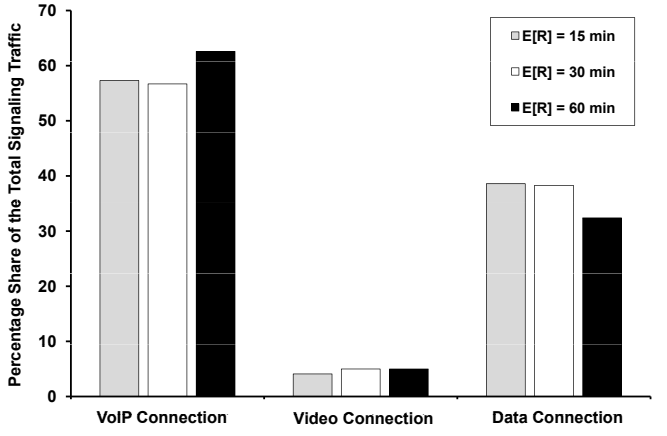

Fig. 8. Service shares of the mean signaling rate, [$\Delta_T \in \{2.5, 10, 30\}$ and $\Delta_M \in \{5, 20, 60\}$ for VoIP, video, and data resp.]

and hence the update interval approaches Δ_T.

Results shown in Fig.8 illustrate the effect of the session durations on the mean signaling rate, for an exemplary mix of services of 70% VoIP, 5% video, and 25% data with session durations of 5, 20, and 60 mins, respectively. The total signaling rate is the sum of the rates from all services, assuming that radio resource admission control and link allocation are always successful. The signaling rate due to each service is calculated based on eq.19 with the corresponding $\lambda^{(i)}$ values set according to the service traffic proportion, the interim interval and the authorization lifetime are set to $0.5E_s$ and E_s respectively. As shown in Fig. 8, the signaling rate shares are not necessarily proportional to the service arrival rates. In other words, we neither have AAA rate shares of 70% for VoIP (i.e., 57-63% instead) nor 25% for data (i.e., 32-38% instead). We notice that higher residence times result in an increase in the share of the short duration services (i.e., VoIP) while reducing the shares for longer session durations (i.e., data) due to the lower number of handoffs and higher residence times.

Figs. 9.a and 9.b show the effect of the handoff signaling

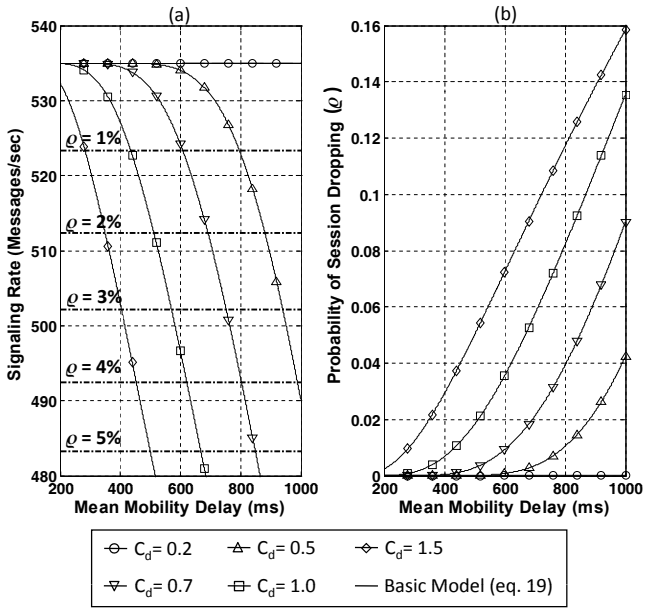

Fig. 9. Signaling rate vs mean mobility delay [Parameters: E_s = 40 min, Δ_T=20 min, Δ_M = 40 min, E_R = 18.4 min, C_R = 2, p_a=0.97, d_a = 2.0 sec, c_d = 1.3, λ = 10 req/s, N_{AGW} = 5].

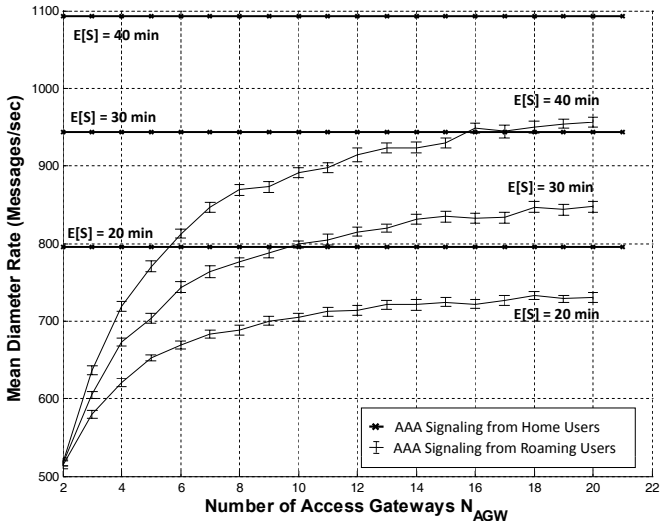

Fig. 10. The AAA signaling load as a function of the number of access gateways in the network under consideration. [E_r = 18.4 min, C_R = 2, $\Delta_M = E_s$, $\Delta_T = 0.5E_s$, session arrival rates for home and roaming users is 100 req/sec. For roaming sessions: 80 req/s initiate in the network, 20 req/s enter the network with established sessions. The mean batch method is used (30 batches, 95% confidence intervals), 10 hr long simulations].

delay on the resulting AAA signaling rate and the session dropping probability. We study the effect of the mean handoff delay as well as the variance of the delay characterized by the coefficient of variation C_d using Gamma fits. The sessions are dropped if the handoff signaling delay d_a>2s. We observe that as the handoff delay and the session dropping probability increase, the corresponding AAA signaling rate decreases. We also see that highly varying handoff delays (i.e., large C_d) result in higher session dropping. For nominal target session dropping rates of ($< 2\%$), the resulting AAA signaling rate can be approximated by the model in eq.19 rather than eq.23.

Fig. 10 compares the AAA signaling loads due to both the home and the roaming users of similar arrival rates by simulations. The simulation is performed similar to [34] by defining a network with linear arrangement of AGWs and assuming a random mobility pattern between AGWs (i.e., the probabilities of going east or west are equal). The network is surrounded by two roaming partners situated at its eastern and western borders. Roaming users may initiate their sessions from within the network under consideration or enter the network with already established sessions. From Fig.10, we observe that the AAA signaling due to home users does not depend on the size of the network while the signaling due to roaming users depends on the size of the network characterized by the number of its access gateways. As the number of AGWs in the network under consideration is increased, longer portions of the roaming sessions are served by the network and hence more signaling is received at the AAA system. When the network becomes very large, the signaling behavior of roaming users approaches that of home users. A similar trend is also observed (not shown here) when the residence time of the AGWs is large (i.e., very large AGW areas) as roaming users make little or no handoffs during their sessions. Further research is needed in better understanding of the spacial, mobility and session statistics effects, to assess the performance of networks with sizable number of roaming users, such as in MVNO scenarios.

Finally, in Table I, we compare our model's results by simulating Log-Normally distributed session times (i.e., non-exponential) with relatively large coefficient of variation ($C_s = 2$). Due to the analytical complexity, the exact analytical consideration of generic session times requires an elaborate approach and is a future item for this research. Table I shows that the error between the analytical model and the simulations is less than 13%. We argue that such error due to the exponential distribution can be considered within the practical 20% design margin and does not necessarily result in excessive over provisioning of the system. Therefore, also the exponential model offers reasonably tolerable accuracy, even for generic sessions with high variance ($C_s = 2$).

VI. Summary and Conclusion

In this study, we presented the first analytical model for the mean AAA traffic rate by considering the users' access-request rates, Diameter protocol specifics, service session duration, and mobility. An approximation for the model was also derived. In addition, we quantified the mean interval between accounting updates in cellular systems critical to the reliability of the billing process in presence of mobility. Our results showed a monotonic decrease in the AAA signaling rate as the residence time increases. We demonstrated that excluding mobility when estimating the AAA rate can result in large errors in presence of high mobility. We also showed that by adjusting the accounting interim and the authorization-lifetime intervals from half to full mean session duration, it is possible to find an AAA operational range that minimizes the signaling rate fluctuations due to likely perturbations in session and mobility statistics. In addition, we showed that when context transfers between AGWs are possible, the resulting authentications rate can be approximated by the AAA signaling model for fixed networks. We also demonstrated that

TABLE I
A COMPARISON BETWEEN THE ANALYTICAL MODEL IN EQ.19 AND SIMULATIONS WITH LOGNORMALLY DISTRIBUTED SESSION TIMES WITH COEFFICIENT OF VARIATION OF 2 [E_r = 18.4 MIN, C_R = 2, $\Delta M = E_s$, $\lambda = 100$ REQUESTS/SEC, SIMULATION'S MEAN BATCH METHOD, 30 BATCHES, 95% CONFIDENCE INTERVALS]

Interim Interval ($\mathbf{\Delta T}$)	E[S] = 40 min		E[S] = 30 min		E[S] = 20 min		E[S] = 5 min	
	Ana.	Error	Ana.	Error	Ana.	Error	Ana.	Error
$\mathbf{E_s/4}$	1278	7.42%	1132	6.78%	987	6.95%	777	6.90%
$\mathbf{E_s/2}$	1093	8.60%	943	6.41%	796	9.01%	580	11.61%
$\mathbf{E_s}$	1013	7.41%	860	6.57%	708	6.73%	486	12.27%

the impact of session dropping on the AAA signaling rate due to excessive mobility delay is insignificant for operational networks. Although our model offers practically good estimates even for generic session times, future work includes the analytical considerations of generic session time distributions and analytically investigating the cases for roaming users.

APPENDIX A
THE FLOOR OF A RANDOM VARIABLE

In this section, we give a generic expression for the distribution and the mean of the number of interims, J, in any random interval X, with a density function $f_X(x)$. Let $J = \lfloor \frac{X}{\Delta_T} \rfloor$. The density function of J can be written as $f_J(j) = \int_{j\Delta_T}^{(j+1)\Delta_T} f_X(x)\,dx = F_X((j+1)\Delta_T) - F_X(j\Delta_T)$. With simple algebraic manipulations and change of variables in the summation, we have,

$$E[J] = \sum_{j=1}^{\infty} j f_J(j) = \sum_{j=1}^{\infty} \bar{F}_X(j\Delta_T) \tag{27}$$

APPENDIX B
THE NUMBER OF HANDOFFS IN A SESSION

In this section, we derive the density function, $Pr(K=k) = f_K(k)$, of the number of handoffs in a session S. Let $R_0^{(n)} = \tilde{R} + \sum_{j=1}^{n-1} R$. If we define $G(k) = Pr\left[S > R_0^{(k)}\right]$ (i.e., the probability that S includes at least k handoffs), then $G(k)$ is given as,

$$G(k) = \int_0^{\infty} \bar{F}_S(x) \left(f_{\tilde{R}}(x) \otimes \overbrace{f_R(x) \otimes .. \otimes f_R(x)}^{(k-1)^{th}-fold} \right) dx \tag{28}$$

Since the n^{th} fold convolution of Gamma density functions is Gamma distributed with shape and scale parameters of nk_r and θ_r respectively, then eq.28 can be expressing using the Laplace transform of $R_0^{(k)}$ as,

$$\begin{aligned} G(k) &= \frac{\theta_r \mathcal{L}\left\{\Gamma(k_r, y) \otimes \text{PDF}_\gamma\left((k-1)k_r, y\right)\right\} \mid_{\hat{s}=\frac{\theta_r}{E_s}}}{\Gamma(k_r) E_r} \\ &= \frac{E_s}{E_r}\left(1 - \left(\frac{\theta_h}{\theta_r}\right)^{k_r}\right)\left(\frac{\theta_h}{\theta_r}\right)^{k_r(k-1)} \end{aligned} \tag{29}$$

It follows that the probability that a session contains k handoffs (where $k \geq 1$) is written as,

$$\begin{aligned} f_K(k) &= Pr\left(R_0^{(k+1)} > S \geq R_0^{(k)}\right) \\ &= Pr\left(S > R_0^{(k)}\right) - Pr\left(S > R_0^{(k+1)}\right) \\ &= G(k) - G(k+1) \quad , \quad k \geq 1 \end{aligned} \tag{30}$$

Finally, the probability that no handoffs p_0 occur in S (i.e., $p_0 = Pr\left(S \leq \tilde{R}\right)$) is given as,

$$\begin{aligned} p_0 &= \int_0^{\infty}\int_0^{y} f_S(x)\,dx f_{\tilde{R}}(y)\,dy = \int_0^{\infty} F_S(x) f_{\tilde{R}}(x)\,dx \\ &= \int_0^{\infty} F_S(x) \frac{\bar{F}_R(x)}{E_r} dx \end{aligned} \tag{31}$$

Similar to eq.29, eq.31 can be written using the Laplace transform as $p_0 = 1 - \frac{\theta_r \mathcal{L}\{\Gamma(k_r,y)\}|_{\hat{s}=\frac{\theta_r}{E_s}}}{E_r \Gamma(k_r)}$. Using eq.33 and after simplifying, we have,

$$p_0 = 1 + \left(\left(\theta_h \theta_r^{-1}\right)^{k_r} - 1\right) E_s E_r^{-1} \tag{32}$$

APPENDIX C
MISCELLANEOUS RELATIONSHIPS

The incomplete gamma function falls under two categories: The lower and the upper incomplete gamma functions denoted as $\gamma(k,x)$ and $\Gamma(k,x)$ respectively and are defined as $\gamma(k,x) = \int_0^x t^{k-1}e^{-t}dt$, $\Gamma(k,x) = \int_x^{\infty} t^{k-1}e^{-t}dt = \Gamma(k) - \gamma(k,x)$. When $x = 0$, we usually write $\Gamma(k) = \Gamma(k,0) = \int_0^{\infty} t^{k-1}e^{-t}dt$. The Gamma probability density function is given as $\text{PDF}_\gamma\left(a, \frac{x}{b}\right) = \frac{x^{a-1}e^{\frac{-x}{b}}}{b^a\Gamma(a)}$. The Laplace transforms of the upper and lower Gamma functions as well as a useful integral are given as,

$$\begin{aligned} \mathcal{L}\{\gamma(k,x)\} &= \Gamma(k)\frac{(1+\hat{s})^{-k}}{\hat{s}} \\ \mathcal{L}\{\Gamma(k,x)\} &= \Gamma(k)\frac{1-(1+\hat{s})^{-k}}{\hat{s}} \end{aligned} \tag{33}$$

A useful relationship that we use in this article is given as,

$$\int \Gamma\left(a, \frac{x}{b}\right) dx = x\Gamma\left(a, \frac{x}{b}\right) - b\Gamma\left(a+1, \frac{x}{b}\right) \tag{34}$$

The integration $\int x\Gamma\left(a, \frac{x}{b}\right) e^{\frac{-x}{c}}$ is needed in this paper and is solved using integration by parts and incorporating eq.34. Defining $d = bc(b+c)^{-1}$, we have,

$$\begin{aligned} &\int x\Gamma\left(a, \frac{x}{b}\right) e^{\frac{-x}{c}} = -ce^{-\frac{x}{c}}(c+x)\Gamma\left(a, \frac{x}{b}\right) \\ &+ \frac{c^2\left(\frac{x}{b}\right)^a\left(\frac{x}{d}\right)^{-a}\left((b+c)\Gamma\left(a, \frac{x}{d}\right) + b\Gamma\left(a+1, \frac{x}{d}\right)\right)}{b+c} \end{aligned} \tag{35}$$

ACKNOWLEDGMENT

We would like to thank Dr. Wolfgang Bziuk for reviewing the early versions of this paper and for his invaluable comments and suggestions.

REFERENCES

[1] "WiMAX Forum Network Architecture - Stage 2 Part 2 - Release 1.1.0,' [Online]. Available: http://www.wimaxforum.org/technology/documents/.

[2] 3GPP2 X.S0011-005-C, "Accounting Services and 3GPP2 RADIUS VSAs," vol. 1.0, Aug. 2003.

[3] 3GPP TS 22.258, "Service Requirements for the All-IP Network (AIPN)," vol. 8.0.0, Mar. 2006.

[4] G. Camarillo and M. Garcia-Martin, *The 3G IP Multimedia Subsystem (IMS)*. John Wiley & Sons, 2004.

[5] "Spirent Communications," [Online]. Available: http://www.spirent.com/about/technology.cfm?az-c=ab\&media=7\&ws=343\&ss=290.

[6] "Cisco content services gateway installation and configuration guide," R3.1, Jan, 2007.

[7] "Openwave Mobile Access Gateway," [Online]. Available: http://www.openwave.com/us/products/gateway_products/mobile_access_gateway/.

[8] Y. Fang, I. Chlamtac, and Y. Lin, "Modeling PCS networks under general call holding time and cell residence time distributions," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, Dec. 1997.

[9] J. Na, Y. Chung, M. Yun, and Y. Kim, "An efficient diameter-based accounting scheme for wireless metropolitan area network (WMAN)," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, 2004.

[10] "RFC 3588-Diameter Base Protocol," [Online]. Available: http://www.faqs.org/rfcs/rfc3588.html.

[11] P. Calhoun, *et al.*, "Diameter Mobile IPv4 Application," *RFC 4004*, Aug. 2005.

[12] "RADIUS accounting - interim accounting record extension," [Online]. Available: www.freeradius.org/rfc/draft-ietf-radius-acct-interim-01.txt.

[13] Ralph Kühne, *et al.*, "Charging in the IP multimedia subsystem: A tutorial," *IEEE Commun. Mag.*, July 2007.

[14] "RFC 4006-Diameter Credit-Control Application," [Online]. Available: http://www.faqs.org/rfcs/rfc4006.html.

[15] F. McEvoy, *et al.*, "New third-party AAA architecture and diameter application for 4GWW," in *Proc. IEEE PIMRC'05*, 2005.

[16] F. Eyermann, *et al.*, "Diameter-based accounting management for wireless services," in *Proc. IEEE WCNC*, 2006.

[17] W. Yang, et al, "Performance modeling of integrated mobile prepaid services," *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, Mar. 2007.

[18] S.-I. Sou, et al, "Modeling prepaid application server of VoIP and messaging services for UMTS," *IEEE Trans. Veh. Technol.*, vol. 56, no. 3, May. 2007.

[19] S. Sou, *et al.*, "Modeling credit reservation procedure for UMTS online charging system," *IEEE Trans. Wireless Commun.*, vol. 6, no. 11, Nov. 2007.

[20] W. Liang and W. Wang, "On performance analysis of challenge/response based authentication in wireless networks," *Elsevier Computer Networks J.*, vol. 48, 2005.

[21] S. Zaghloul and A. Jukan, "On the performance of the AAA systems in 3G cellular networks," in *Proc. IEEE ICC'07 Conf.*, 2007.

[22] S. Zaghloul and A. Jukan, "Relating the AAA and the radio access rates in 3G cellular networks," *IEEE Commun. Lett.*, Apr. 2007.

[23] K. Yeo and C. Jun, "Modeling and analysis of hierarchical cellular networks with general distributions of call and cell residence times," *IEEE Trans. Veh. Technol.*, vol. 51, no. 6, Nov. 2002.

[24] P. Orlik and S. S. Rappaport, "A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions," *IEEE J. Select. Areas Commun.*, vol. 16, June 1998.

[25] "RFC 2865-Remote Authentication Dial In User Service (RADIUS)," [Online]. Available: http://www.faqs.org/rfcs/rfc2865.html

[26] 3GPP2 X.S0013-000-B, "All-IP core network multimedia domain," vol. 1.0, Dec. 2007.

[27] A. Abdi and M. Kaveh, "K distribution: an appropriate substitute for Rayleigh-lognormal distribution in fading-shadowing wireless channels," *IEEE Electronic Lett.*, vol. 34, pp. 851-852, 1998.

[28] "Starent ST16 PDSN&HA Datasheet," http://www.starentnetworks.com/pdf/StarentNetworks_PDSN_Datasheet_0904.pdf [Online].

[29] S. Mohanty and I. Akyildiz, "Performance analysis of handoff techniques based on mobile IP, TCP-migrate, and SIP," *IEEE Trans. Mobile Computing*, vol. 6, no. 7, July 2007.

[30] M. Ylianttila, J. Mäkelä, and K. Pahlavan, "Analysis of handoff in a location-aware vertical multi-access network," *Elsevier Computer Nets*, vol. 47, 2005.

[31] H. Yokota *et al.*, "RFC5271-Mobile IPv6 Fast Handovers for 3G CDMA Networks," June 2008.

[32] 3GPP2 X.S0040-0, "PPP-alternate protocol (AltPPP) for cdma2000, wireless IP network standard," vol. 1.0, Jan. 2007.

[33] M. Claypool *et al.*, "Characterization by measurement of a CDMA 1x EVDO network," *ACM WICON'06 Conf.*, Aug. 2006.

[34] S. Zaghloul, W. Bziuk, and A. Jukan, "Signaling and handoff rates at the policy control function (PCF) in IP multimedia subsystem (IMS)," *IEEE Commun. Lett.*, July 2008.

Said Zaghloul (StM) is a PhD candidate at the Technical University Carolo-Wilhelmina of Braunschweig in Germany. Prior to his PhD studies, he was with Sprint-Nextel as a Telecommunication Design Engineer where he was a major contributor to the design and testing of Sprint's wireless data network architectures in several areas including MVNO and roaming solutions, dual GPRS/CDMA solutions and hybrid WiFi/CDMA phone products. In 2003, Said was granted a Fulbright Scholarship to pursue his MSc studies at the University of Kansas. In 2005, he received his MSc degree with honors in computer engineering, for his work in the area of system design and protocol performance of inverse multiplexed satellite connections. In 2002, he received the first IEE award for BSc senior projects in Jordan for his undergraduate project work dedicated to the development of a UMTS cellular planning tool. Said's current research interests include: next generation all-IP wireless architectures, signaling plane performance, mobility, wireless protocols, and wireless communications.

Admela Jukan (SM) is W3 Professor of Electrical and Computer Engineering at the Technical University Carolo-Wilhelmina of Brunswick (Braunschweig) in Germany. Prior to coming to Brunswick, she was research faculty at the Institut National de la Recherche Scientifique (INRS), University of Illinois at Urbana Champaign (UIUC) and Georgia Tech (GaTech). From 2002-2004, she served as Program Director in Computer and Networks System Research at the National Science Foundation (NSF) in Arlington, VA. She received the M.Sc. degree in Information Technologies and Computer Science from the Polytechnic of Milan, Italy, and the Ph.D. degree (cum laude) in Electrical and Computer Engineering from the Vienna University of Technology (TU Wien) in Austria. Dr. Jukan is the author of numerous papers in the field of networking, and she has authored and edited several books. She serves as a member of the External Advisory Board of the EU Network of Excellence BONE. Dr. Jukan has chaired and co-chaired several international conferences, including IFIP ONDM, IEEE ICC and IEEE GLOBECOM. She serves as Associate Technical Editor for IEEE COMMUNICATIONS SURVEYS, IEEE COMMUNICATIONS MAGAZINE and IEEE NETWORK.