Manuscript submitted to
AIMS' Journals
Volume **X**, Number **0X**, XX **200X**

doi:10.3934/xx.xx.xx.xx

pp. **X–XX**

# MODELING THE SIGNALING OVERHEAD IN HOST IDENTITY PROTOCOL-BASED SECURE MOBILE ARCHITECTURES

Zoltán Faigl

Mobile Innovation Centre
Budapest University of Technology and Economics
Műegyetem rkp. 3.
Budapest, 1111, Hungary

Miklós Telek

MTA-BME Information systems research group
and
Department of Networked Systems and Services
Budapest University of Technology and Economics
Műegyetem rkp. 3.
Budapest, 1111, Hungary

(Communicated by the associate editor name)

Abstract. One of the key issues in recent mobile telecommunication is to increase the scalability of current packet data networks. This comes along with the requirement of reducing the load of signaling related to establishment and handover procedures. This paper establishes an analytical model to analyze the signaling overhead of two different secure mobile architectures. Both are based on the Host Identity Protocol for secure signaling and use IPsec for secure data transport. The paper presents the cumulative distribution function and moments of security association periods and calculates the rate of different signaling procedures in a synthetic network model assuming M/G/$\infty$ process for session establishments between end-nodes. Using the model, it is shown that the Ultra Flat Architecture has significant performance gains over the traditional End-to-End HIP protocol in large-scale mobile environment in the access networks and toward the rendezvous service, but performs worse in the core transport network between the GWs.

1. **Introduction.** By the end of 2014, the number of mobile-connected devices will exceed the number of people on earth, and by 2018 there will be nearly 1.4 mobile devices per capita [1]. Reduction of signaling load related to establishment and handover procedures is hence one of the important challenges for mobile networks architectures.

As a possible approach to enhance scalability of the core network, the Ultra Flat architecture (UFA) has been introduced by Daoud et al. [2]. Fig. 1 illustrates the UFA. UFA represents the ultimate step toward flattening the packet-switched domain of mobile networks. The objective of the UFA design is to distribute core functions into single nodes at the edge of the network, e.g., in local Point of Presences of mobile network operators. Certain control functions could remain centralized, e.g., the subscriber information base, domain name resolution service and addressing

2010 *Mathematics Subject Classification.* Primary: 68M20, 68M11; Secondary: 68M10, 60K20.
*Key words and phrases.* security association period, infinite server queue, Host Identity Protocol, delegation of signaling rights, Ultra Flat Architecture.

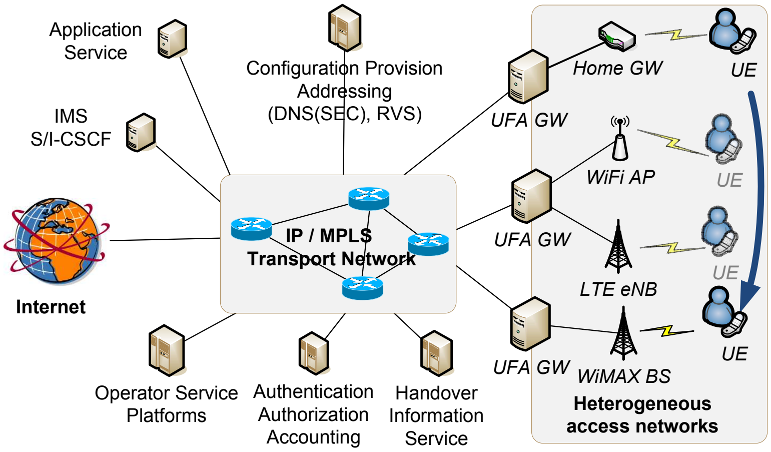

Application Service
IMS S/I-CSCF
Internet
Configuration Provision Addressing (DNS(SEC), RVS)
IP / MPLS Transport Network
Operator Service Platforms
Authentication Authorization Accounting
Handover Information Service
UFA GW
Home GW
UE
WiFi AP
UE
UFA GW
LTE eNB
UE
UFA GW
WiMAX BS
UE
Heterogeneous access networks

FIGURE 1. Ultra Flat Architecture.

service, which resolve application-level identifiers to the IP address of the peers. The intelligent nodes at the edge of the network are called UFA gateways (GWs). The user data traffic is conveyed directly between communicating parties through these GWs. GWs are also breakout points toward content distribution networks and Internet services.

Faigl and Bokor [3, 4] introduced a Host Identity Protocol (HIP)-based [5] signaling scheme for UFA, referred to as UFA HIP in the following. On the other hand, a network deploying the standard HIP control plane will be referred to as end-to-end HIP architecture (E-E HIP) in the following.

Normally, HIP operates in an end-to-end or terminal-based fashion and provides key agreement, Internet Protocol security (IPsec) security association (SA) management and IP mobility management between pairs of endpoints [6]. IPsec is a standardized protocol suite to provide encryption, integrity, message origin authentication and anti-replay protection for IP datagrams between two hosts. An SA is the bundle of algorithms and parameters on two hosts, being used to encrypt and authenticate IP datagrams selected by traffic selectors in one direction. For the protection of bi-directional traffic, an SA pair is required. The initial SA establishment procedure is dubbed as Base Exchange (BEX) in HIP. After initial SA establishment there are forthcoming HIP update procedures for different purposes, such as rekeying, IP address update due to handovers, as described in Section 2.2.

HIP operates between the network and transport layer, and splits the identity and locator role of IP address. It means that the addressing is based on long-term, globally unique host identities (HIs) instead of short-term IP addresses. Host Identity Tag (HIT) is a 128-bit hash generated from the HI, having the same format as an IPv6 address. Legacy applications use the HIT for addressing peers. The HIP stack in the hosts is responsible for the translation of HITs to IP addresses and for the treatment of IP address changes and the presence of multiple network interfaces, seamlessly for the applications. HIP Host Association (HA) is a set of states in the control plane of peers established after a successful BEX. A HA includes the HIT and IP addresses of the peers, the key material, cipher suite for protection of the communication in the control plane, i.e., for protection of HIP communication, and user plane, that is, the parameters of the SA pair.

HIP-enabled hosts can register and keep updated their address at the rendezvous service (RVS). The purpose of RVS is the following. If the HIP stack of a host does not have up-to-date information on the locator of a peer, then the first HIP packet of BEX is sent toward the RVS and forwarded by the RVS to the actual locator of the destination peer. This is typically required for initial reachability of a peer or in case of simultaneous IP address change of the endpoints. If only one endpoint changes its address, then it notifies its peers with HIP update messages for the modification of the IP address in the existing HAs and SAs.

The terminal-based control of E-E HIP has some drawbacks in an operator-controlled environment. The network has no ability to control and decrypt IPsec communication, which encumbers, e.g., traffic control, mobility management, deep packet inspection, legacy interception by the operator. Additionally, a terminal-based control causes unnecessary high network and computational overhead on the UEs and in the access networks.

Hence, UFA HIP utilizes a new control function, the delegation of signaling rights [7, 4] integrated into terminal attachment, session establishment and inter-GW handover procedures. In case of delegation of signaling rights the delegates are temporarily authorized by the delegator to proceed in certain tasks, such as periodic location updates, rekeyings. The delegate gets notifications from the delegator about state changes. In general, delegation may facilitate the optimization of resource utilization between the delegator and the delegate. In UFA HIP, the UEs delegate certain HIP control roles to their access GW. GWs are responsible for sending and receiving BEX messages, location updates and rekeying requests toward and from the peers of the UE, and for notifying the UE about state changes.

UFA HIP divides the E-E SAs, characteristic for E-E HIP, into two segments: one between the UE and the GW and the other between the GW and the peer of the UE. In a network with $N$ attached users and $M$ GWs, E-E HIP and UFA HIP have in maximum $N(N-1)/2$ and $N+M(M-1)/2$ possible SA pairs, respectively. Assuming that $M < N$, it is expected that UFA HIP will induce less network and processing overhead in the control plane than E-E HIP on the whole or in certain parts of the network.

The objective of this paper is to analyze the performance gains achieved by UFA HIP compared to E-E HIP in terms of signaling reduction. Thus, Section 2 presents the main characteristics of signaling procedures for both schemes. Section 3 determines the Cumulative Distribution Function (CDF), BEX rate and moments of the SA period between pairs of hosts and the mean rates of different types of update procedures. Section 4 applies the analytical model, and determines the overhead of the two schemes for a predefined set of input parameters. Section 5 concludes the paper and discusses our plans.

As a side result, Appendix A, describes the calculation of the moments of busy period in infinite server queues, a part of related work that needed some clarification.

2. **Background and related work.** We call SA period the duration of a HA and the related SA pair between two HIP hosts. The length of SA period depends on the busy and idle periods of the connection, which are determined by communication sessions between the two hosts. The exact behavior of SA period will be described in Section 3.1. To calculate the CDF and moments of SA period, we must know the CDF and moments of busy and idle periods.

2.1. **Busy period in M/G/$\infty$ queuing systems.** In our model we assume that the arrival process and the duration of communication sessions originating from the running applications and services in the peers can be described with an M/G/$\infty$ queuing system. The infinite server assumption holds under the theoretic conditions of no buffering and transmission delays in the E-E path. In those conditions the HIP control layer is notified at the same instant about a communication session when that appears on the link. This assumption enables simple analytical modeling of SA periods. Hence, to determine the characteristics of SA periods, first, we must know the behavior of busy periods in M/G/$\infty$ queuing system.

Kulkarni [8] provides exact, closed form solution of the Laplace-Stjieltes transform (LST) of the CDF of the busy period of an M/G/$\infty$ queuing system. This solution can be used to compute any moment of the busy period. It is a drawback of the Laplace transform (LT) solution that finding the inverse LT is inefficient. [8, p. 425] has some errors in the proof and the resulting first moment. Hence, in Appendix A we clarified the calculation of moments. The first and second moments of busy periods are provided by (97) and (98) for general service time distribution and by (102) and (103) for exponentially distributed service time.

Daley [9] gives an integral expression for the complementary CDF (CCDF) of the busy period of M/G/$\infty$ queue in time domain. The CCDF of busy period is (105) in Appendix B. Due to the recursion in the expression the CDF of busy period can only be calculated by a numeric method. Algorithm 3 presents the calculation using $\Delta t$ as the step-size.

2.2. **Base exchange and update procedures.** The base exchange (BEX) procedure is illustrated in Fig. 2. I1 packet starts the procedure, containing the HIT of the initiator and responder. I1 is a basic hello message and part of return routability procedure to check the availability of responder. R1 packet is a pre-created response to I1 packet.

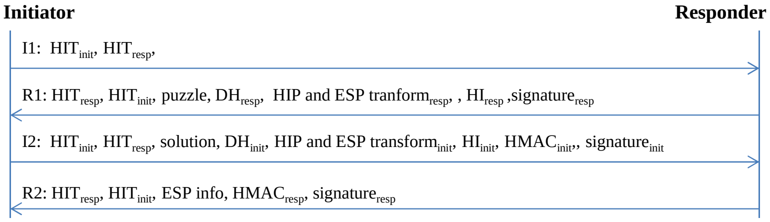

Initiator
Responder
I1: HIT$_{init}$, HIT$_{resp}$,
R1: HIT$_{resp}$, HIT$_{init}$, puzzle, DH$_{resp}$, HIP and ESP tranform$_{resp}$, , HI$_{resp}$ ,signature$_{resp}$
I2: HIT$_{init}$, HIT$_{resp}$, solution, DH$_{init}$, HIP and ESP transform$_{init}$, HI$_{init}$, HMAC$_{init}$,, signature$_{init}$
R2: HIT$_{resp}$, HIT$_{init}$, ESP info, HMAC$_{resp}$, signature$_{resp}$

FIGURE 2. HIP Base Exchange.

The puzzle field in R1 contains a challenge for the initiator. In general, puzzle-based challenge-response mechanisms aim to mitigate denial-of-service attacks initiated from fake initiators against the responder. In HIP, the initiator must find a solution that, if given to a one-way hash function concatenated with this challenge, produces an output starting with a pre-defined number of zeros (e.g., Hash(challenge, solution) $= 0x000ABCDEF$). The expected number of zeros, i.e., difficulty level of the puzzle, is tunable by the responder. To find a good solution, the initiator must execute a brute-force search. The responder can verify the solution sent in I2 packet by only one hash function call. The responder does not establish HA and SA pair with the initiator until the reception of the good solution in I2.

The R1, I2 and R2 messages implement the standard authenticated DH key exchange method [10]. The calculation of DH public values of the peers is the following. $\mathrm{DH}_{\mathrm{resp}} = g^x \mathrm{mod} p$, $\mathrm{DH}_{\mathrm{init}} = g^y \mathrm{mod} p$, where $g = 2$, $p$ is a standard large prime number [11], and $x, y \in [1, p-2]$ are self-generated random values of the peers. The peers calculate a shared key, referred to as DH secret, as $K_{\mathrm{resp}} = DH^x_{\mathrm{init}}$ and $K_{\mathrm{init}} = \mathrm{DH}^y_{\mathrm{resp}}$, i.e., $K_{\mathrm{init}} = K_{\mathrm{resp}} = g^{xy} \mathrm{mod} p$. The session keys are calculated at each peer using a one-way hash function with the DH secret as one of the parameters. Such session keys are e.g., the integrity keys. The Hashed Message Authentication Codes (HMAC) provide message origin authenticity and integrity protection, and can be verified by the other peer knowing the integrity key (HMAC = Hash(message,integrity key)). Public-key signatures provide authentication of the peer, the origin and integrity of the message, and can be checked by any peer using the HI of the signing entity as the public key. The ESP info field contains the security parameter indexes, required for the identification of the SA pair.

Fig. 3 presents the update procedure of HIP. The mandatory fields of an update packet are HITs, HMAC, signatures, the sequence and acknowledgment number. The latter two fields enable detection of packet loss and ordered delivery of update packets. Non-mandatory fields are the following. Notification carries control data, e.g., the new IP address of a mobile peer. DH public key values are sent in case of rekeying for calculation of a fresh DH secret and key material for the HA and the related SA pair. The registration request and response (REG req and resp) enable subscription to a service of the peer, such as delegation of signaling or RVS service. CERT field carries the certificate-chain which proves that the signature is valid and the update procedure is authorized. An update procedure normally contains three packets. However, if the size of CERT field or a Notification field together with the other ones is larger than the maximum transfer unit size of the network then these fields are sent in multiple packets to the other peer. The peers must acknowledge each packet from the other peer by communicating the sequence number of the received packet.

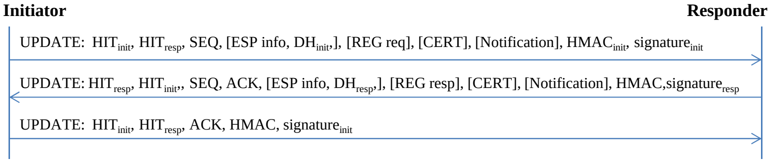

Initiator
Responder
UPDATE: $\mathrm{HIT}_{\mathrm{init}}$, $\mathrm{HIT}_{\mathrm{resp}}$, SEQ, [ESP info, $\mathrm{DH}_{\mathrm{init}}$,], [REG req], [CERT], [Notification], $\mathrm{HMAC}_{\mathrm{init}}$, $\mathrm{signature}_{\mathrm{init}}$
UPDATE: $\mathrm{HIT}_{\mathrm{resp}}$, $\mathrm{HIT}_{\mathrm{init}}$,, SEQ, ACK, [ESP info, $\mathrm{DH}_{\mathrm{resp}}$,], [REG resp], [CERT], [Notification], HMAC,$\mathrm{signature}_{\mathrm{resp}}$
UPDATE: $\mathrm{HIT}_{\mathrm{init}}$, $\mathrm{HIT}_{\mathrm{resp}}$, ACK, HMAC, $\mathrm{signature}_{\mathrm{init}}$

Figure 3. HIP update.

We distinguish three update procedure types for the sake of the paper because they have different network and node processing requirements. (1) UPDATE with DH (UPDATEw/DH) signifies the rekeying procedure, as defined in the standard [12]. It contains the DH public key values and ESP info from the non-mandatory fields. (2) UPDATE with CERT (UPDATEw/CERT) refers to an update containing the CERT field [13]. It is required in the following two subcategories. First, when the delegator, i.e., an UE or a GW, registers to the delegation service of a GW. In this case the CERT field contains the authorization certificate-chain, which authorizes the delegate to act in the name of the delegator in the scope of the authorized roles. Second, when a mandated update procedure is initiated by a delegate towards the peer of the delegator. Mandated means that the signaling is performed by a delegate in the name of the delegator. In this case the CERT field contains the authorization certificate-chain, which certifies for a peer that the delegate is temporarily authorized to proceed in the name of the delegator. A mandated update procedure can have any purpose except rekeying. Such purposes are, e.g., registration of the delegator's IP address at the RVS, update of the peers of the delegator with the new IP address of the delegator, or registration of the delegator to the delegation service of a next delegate. A certificate-chain may be too long for one CERT field within on HIP packet, therefore it may be split into multiple CERT fields that are transferred in more than one HIP update messages. (3) UPDATE signifies all the other types of update procedures, e.g., registration to RVS [14] or location update of the peers [15].

2.3. **Establishment and handover procedures in E-E and UFA HIP.** Tables 1 and 2 describe the main functions that the HIP control plane provides in E-E and UFA HIP.

Fig. 4 outlines the terminal attachment, session establishment and handover procedures from the point of view of a UE. Fig. 5 and 6 the same in UFA HIP. All figures present the triggered HIP signaling procedures and the related control function in parentheses.

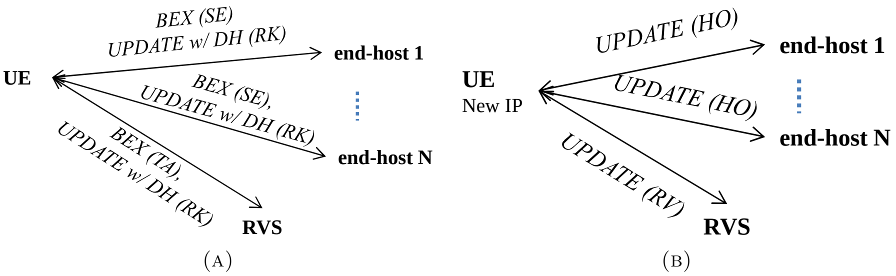


FIGURE 4. (A) Session establishment and (B) handover procedures in end-to-end HIP based network

## 3. Security association period, mean BEX and update rates in HIP-based architectures.

This section describes the main mathematical statements required for the performance evaluation of UFA and E-E HIP.

3.1. **Definitions and notations.** Fig. 7 presents the number of active sessions ($Q(t)$) between two HIP peers as a function of time ($t$), and the behavior of SA periods.

The basic assumptions of our analysis are the following. Sessions above the IP-layer are established between two hosts according to a Poisson process with rate $\lambda$. $Y$ denotes the inter arrival times of the sessions, which are exponentially distributed with parameter $\lambda$. The session duration times, denoted by $S$, are generally distributed. The number of active sessions at time $t$ is denoted by $Q(t)$. The busy periods are defined as the periods when $Q(t) > 0$. $X$ stands for the length of busy period. $\check{Y}$ and $\widehat{Y}$ denote idle periods with the following restrictions. In case of $\check{Y}$ the idle period is shorter than a constant parameter $T$, while in case of $\widehat{Y}$ it is longer

Table 1. HIP control functions in E-E and UFA HIP (first part).

| Abbrev. | Description |
|---|---|
| TA | Terminal attachment (TA) occurs when a HIP-enabled host is switched on or is rebooted. |
| in E-E | The UE executes BEX procedure towards the RVS, as illustrated in Fig. 4a. |
| in UFA | The UE executes a BEX towards the access GW [3], as presented in Fig. 5a. |
| SE | Session establishment (SE) covers the HIP signaling procedures related to communication flows between the HIP peers. |
| in E-E | BEX is triggered before session establishment as long as there is no HA between two peers, as drafted in Fig. 4a. The HA is closed (and the SA is deleted) when the SA is unused by upper-layer communication for a period denoted by $T$. $T$ is the minimum of the unused association lifetimes (UALs) configured at the initiator and responder [5]. |
| in UFA | In UFA HIP, the same SA pair is used for all communication flows between an UE and its delegate GW, independently of the remote peer. Furthermore, the same SA pair between two GWs provides data protection for all service data flows passing through these GWs, independently from the source and destination peers. During SE, an UPDATE procedure is triggered between the UE and its UFA GW, as long as there is no HA between the UE and the remote peer, as presented in Fig. 5b. During that, the GW is notified about the request to establish HA with the remote peer in the name of the UE, and the UE gets feedback on the success of delegated task. In case of lack of SA pair between the GW and the remote peer, a BEX procedure is initiated from the GW to the remote peer in the name of the UE. Otherwise, if the SA pair has earlier been established (by other flows) between the GW and the remote peer, then an UPDATEw/CERT is triggered. If the remote peer turns out to be a delegate GW, then that GW notifies the remote peer about HA creation using an UPDATE procedure. |
| HO | Handover (HO) covers the HIP procedures for mobility management of a UE. A basic assumption is that GWs publish different IP domains. Handover means that a UE is moving from one IP address domain to another IP address domain, by visiting an access network that is connected to a new GW. |
| in E-E | If a host gets a new IP address, it sends the address to its peers using an UPDATE procedure [15], as illustrated in Fig. 4b. This is a reactive mobility management solution. |
| in UFA | The handover procedure has two phases in UFA HIP as illustrated by Fig. 6. The HO procedure in UFA HIP realizes proactive handover. This means that the contexts for data link, network and HIP-layer are established and updated by the control plane in the UE, GWs and the UE's peers before the UE is physically reattached to the next GW. In the first phase of HO (I) the target GW requests the source GW to establish HA with the UEs peers using UPDATEw/CERT procedures. At the end of the phase, the security contexts are transferred from the source GW to the target GW. In the second phase (II) the target GW updates the traffic forwarding policies for the UE at the UE's peers, in the RVS and within the UE itself using UPDATEw/CERT procedures. Therefore the UE's traffic is redirected and passing through the target GW. After that the UE physically reattaches to the new access network. |

than a constant parameter $T$. $T$ is the length of the unused association lifetime as introduced in Table 1. $G$ denotes the SA period, while $\overline{G}$ indicates the period where there is no SA pair established between two nodes.

Fig. 7 illustrates the behavior of SA periods. An SA period is composed of a random number of busy-idle period pairs $(X + \check{Y})$, as long as the last busy period is not followed by an idle period longer than $T$, i.e., $\widehat{Y}$. $G$ finishes $T$ time after the end of the last busy period $X$. Let $p$ be the probability that $Y < T$. $p$ can be calculated as

$$p = 1 - F_Y(T) = \overline{F}_Y(T) = e^{-\lambda T}. \tag{1}$$

Table 2. HIP control functions in E-E and UFA HIP (second part).

| Abbrev. | Description |
|---|---|
| RV | RVS update (RV) means registration of the fresh locator of a HIP-enabled host at the rendezvous service. The registration lifetime denoted by the symbol $T_{\mathrm{RVS}}$ determines the lifetime of an address entry in the database of the RVS. |
| in E-E | An UE registers its IP address at the RVS right after TA [14]. Further registrations are triggered due to two factors. First, the registration lifetime $T_{RVS}$ configured at the RVS server determines the minimum frequency of periodic location updates that should be initiated by the UE. Second, during handovers the UE notifies the RVS about its new IP address. The UPDATE procedure is used in both cases. |
| in UFA | Mandated registrations of the UEs' addresses are triggered during every HO and at every $T_{\mathrm{RVS}}$ time by the UFA GWs. The applied procedure is UPDATEw/CERT including the registration request and reply fields. |
| RK | Rekeying (RK) aims to create fresh keys for a given HA and the related SA pair, using UPDATEw/DH procedure between the peers. $T_{KEY}$ denotes the length of the rekeying period between HIP-enabled hosts. |
| in E-E | RK may occur between the UEs and between the UE and the RVS. |
| in UFA | RK may occur between the UEs and their actual serving GW, between GWs and between the RVS and GWs. |
| DR | Delegation of rights (DR) is present only in UFA HIP. It uses the UPDATEw/CERT procedure involving registration request and reply fields. |
| in UFA | It occurs in three main cases: first, following the TA, second, when the lifetime of delegation authorization expires, third, during the handovers. In the first two cases the UE registers to the delegation service of its access GW. It generates a temporary public-key certificate for the GW for a certain time called delegation lifetime, denoted by the symbol $T_{\mathrm{DEL}}$ in the following. Hence, the GW will be able to sign control messages until the expiration of $T_{\mathrm{DEL}}$ in the name of the UE by attaching the certificate of the UE to its signed messages. The third case of DR happens during HO. During HO, either the UE or the previous GW delegates the UE's signaling rights to the next GW. Let denote with the symbol $L$ the maximum certificate-chain length. If the actual length of the certificate-chain is smaller than $L$ before the HO, then the previous GW will propagate the UE's signaling right to the target GW by registering to the delegation service of the target GW, and authorizing it to proceed in the name of the UE. That means that the previous GW appends his certificate to the certificate-chain and conveys that in the CERT field to the target GW. If the length of the certificate-chain is $L$ before HO, then the UE is responsible for the re-delegation of its rights. Hence, the UE will send a new certificate in the CERT field for the target GW with a new lifetime, which will authorize the target GW to proceed in his name. |

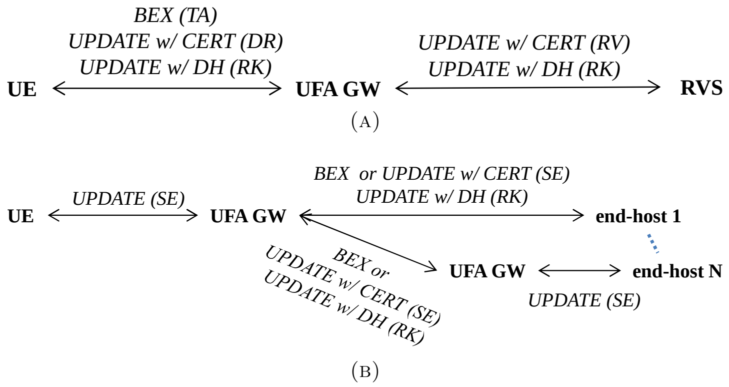
BEX procedure is triggered at the initiation of SA period. $\widehat{\lambda}$ denotes the mean rate of BEX procedures and $1/\widehat{\lambda}$ is the mean time between BEX procedures of a pair of hosts.

The following notations are used throughout the paper. The CDF and CCDF of a random variable $X$ is denoted by $F_X(x)$ and $\overline{F}_X(x)$ and defined by $F_X(x) = Pr(X < x)$ and $\overline{F}_X(x) = 1 - F_X(x)$, respectively. $f_X(x)$ denotes the probability density function (PDF) of $X$. $f_X(x) = dF_X(x)/dx$. $X^*(s)$ stands for the LT of the non-negative random variable $X$, and is defined as

$$X^*(s) = E(e^{-sX}) = \int_{x=0}^{\infty} e^{-sx} f_X(x) dx = \int_{x=0}^{\infty} e^{-sx} dF_X(x). \qquad (2)$$

The last expression is also referred to as the Laplace-Stieltjes Transform (LST) of $F_X(x)$.

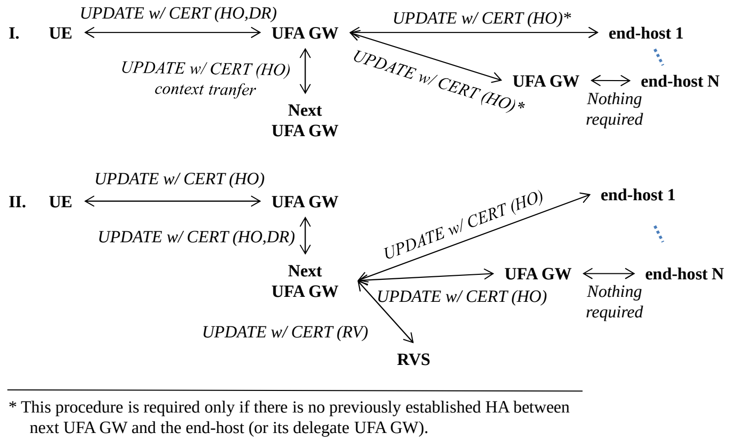
The distribution of the sum of two positive random variables ($Z = X + Y$) can be calculated, both, in time and in LT domain. In time domain PDF is obtained



(A)

(B)

Figure 5. (A) Terminal attachment and (B) session establishment in UFA HIP based network



* This procedure is required only if there is no previously established HA between next UFA GW and the end-host (or its delegate UFA GW).

Figure 6. Handover procedure in UFA HIP based network

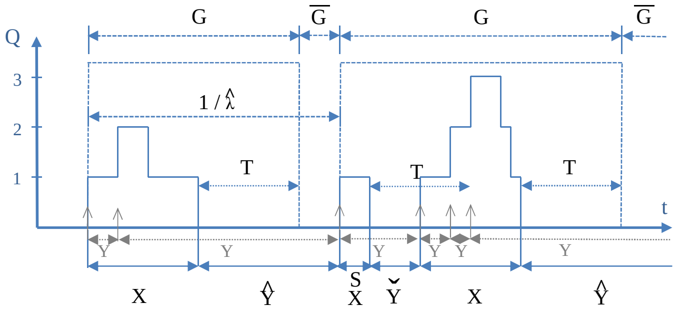


Figure 7. Behavior of SA periods over an M/G/∞ system of session periods.

by convolution

$$f_Z(t) = f_X(t) \circledast f_Y(t) = \int_{u=0}^{t} f_X(u) f_Y(t-u) du, \tag{3}$$

where the convolution operator is denoted by $\circledast$. In Laplace domain the convolution is replaced by multiplication, i.e.,

$$Z^*(s) = X^*(s)Y^*(s). \tag{4}$$

The $n^{th}$ moment of random variable $X$ can be calculated from the LT domain description as follows

$$E(X^n) = (-1)^n \frac{d^n X^*(s)}{ds^n}\bigg|_{s=0}. \tag{5}$$

Next, we compute the main properties of $T$, $\check{Y}$ and $\widehat{Y}$. Since $T$ is a constant (deterministic) quantity we have

$$T^*(s) = E(e^{-sT}) = e^{-sT}. \tag{6}$$

The CDF of $\check{Y}$ is

$$\begin{aligned} F_{\check{Y}}(t) = Pr\{Y < t | Y < T\} &= \begin{cases} \frac{Pr\{Y<t, Y<T\}}{Pr\{Y<T\}}, & \text{if} \quad t < T, \\ 1, & \text{if} \quad t > T \end{cases} = \\ &= \begin{cases} \frac{1-e^{-\lambda t}}{1-e^{-\lambda T}}, & \text{if} \quad t < T, \\ 1, & \text{if} \quad t > T. \end{cases} \end{aligned} \tag{7}$$

From that, the PDF of $\check{Y}$ is

$$f_{\check{Y}}(t) = \frac{d}{dt} F_{\check{Y}}(t) = \begin{cases} \frac{\lambda e^{-\lambda t}}{1-e^{-\lambda T}}, & \text{if} \quad t < T, \\ 0, & \text{if} \quad t > T, \end{cases} \tag{8}$$

and finally the LT, $\check{Y}^*(s)$, is

$$\check{Y}^*(s) = \int_0^\infty f_{\check{Y}}(t) e^{-st} dt = \int_0^T \frac{\lambda e^{-\lambda t}}{1-e^{-\lambda T}} e^{-st} dt = \frac{\lambda}{1-e^{-\lambda T}} \frac{1-e^{-(s+\lambda)T}}{\lambda + s}. \tag{9}$$

The first moment of $\check{Y}$ can be calculated in two ways: by integration of $f_{\check{Y}}(t)$ and by (5) using (9). Both methods give the same result, that is,

$$\begin{aligned} E(\check{Y}) = \int_0^\infty t f_{\check{Y}}(t) dt = -\frac{d\check{Y}^*(s)}{ds}\bigg|_{s=0} &= \frac{1}{\lambda} - \frac{T}{e^{\lambda T} - 1} = \\ &= \frac{1}{\lambda} - \frac{1}{\sum_{i=1}^{\infty} \frac{\lambda^i T^{i-1}}{i!}}, \end{aligned} \tag{10}$$

where the last expression is introduced to characterize the limiting behavior of $E(\check{Y})$. The correctness of $E(\check{Y})$ can be checked by its limiting values. As $T \to \infty$, i.e., the upper constraint for the exponentially distributed inter arrival period disappears, we have

$$\lim_{T\to\infty} E(\check{Y}) = \frac{1}{\lambda} = E(Y). \tag{11}$$

On the other hand, as $T \to 0$, i.e., the upper constraint for the exponential period tends to zero we have

$$\lim_{T\to 0} E(\check{Y}) = 0. \tag{12}$$

The CDF of $\widehat{Y}$ is

$$\begin{aligned} F_{\widehat{Y}}(t) = Pr\{Y<t|Y>T\} = &\left\{ \begin{array}{ll} \frac{Pr\{Y<t,Y>T\}}{Pr\{Y>T\}}, & \text{if} \quad t>T, \\ 0, & \text{if} \quad t<T \end{array} \right. = \\ = &\left\{ \begin{array}{ll} 1-\frac{e^{-\lambda t}}{e^{-\lambda T}}, & \text{if} \quad t>T, \\ 0, & \text{if} \quad t<T. \end{array} \right. \end{aligned} \tag{13}$$

From that, the PDF of $\widehat{Y}$ is

$$f_{\widehat{Y}}(t) = \frac{d}{dt} F_{\widehat{Y}}(t) = \left\{ \begin{array}{ll} \frac{\lambda}{e^{-\lambda T}} e^{-\lambda t}, & \text{if} \quad t>T, \\ 0, & \text{if} \quad t<T. \end{array} \right. \tag{14}$$

Hence, the LT, $\widehat{Y}^*(s)$, is

$$\begin{aligned} \widehat{Y}^*(s) = &\int_0^\infty f_{\widehat{Y}}(t) e^{-st} dt = \frac{\lambda}{e^{-\lambda T}} \int_T^\infty e^{-(\lambda+s)t} dt = \\ = &\frac{\lambda}{e^{-\lambda T}} \left[ -\frac{1}{\lambda+s} e^{-(\lambda+s)t} \right]_{t=T}^{\infty} = -\frac{\lambda}{\lambda+s} e^{-sT}. \end{aligned} \tag{15}$$

The first moment of $\widehat{Y}$ is

$$E(\widehat{Y}) = -\left.\frac{d\widehat{Y}^*(s)}{ds}\right|_{s=0} = \left.\left( Te^{-sT} \frac{\lambda}{\lambda+s} + \frac{\lambda}{(\lambda+s)^2} e^{-sT} \right)\right|_{s=0} = T + \frac{1}{\lambda}. \tag{16}$$

3.2. **Modeling assumptions.** Fig. 8 illustrates our network model and parameters. To keep simple the analytic description of the signaling behavior, we made simple assumptions regarding the attachment, session establishment and mobility behavior of the UEs, without entering too much into the details of their relations. UEs form a very large population of size $N$. An UE attaches to the network with rate $\alpha$ and remains attached for a generally distributed time with mean $1/\omega$. The number of attached UEs is not limited. Hence the number of attached UEs is described by an M/G/$\infty$ queue. The mean number of attached users is $N_{\text{UE}} = \frac{N\alpha}{\omega}$

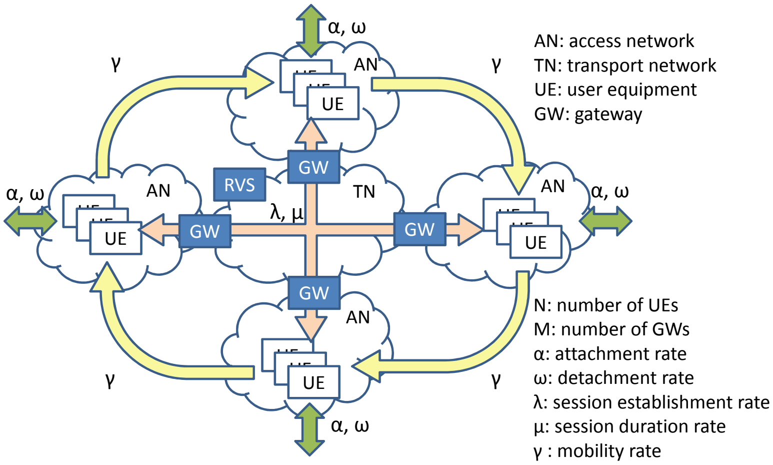


FIGURE 8. Network model.

and the mean number of destination peers of a UE is $N_{\mathrm{UE}} - 1$. The UEs are uniformly distributed in the access networks (ANs). There are $M$ GWs, hence each GW provides access to $\frac{N_{\mathrm{UE}}}{M}$ UEs on average.

The number of data sessions between an UE and all the other attached UEs is assumed to be an M/G/$\infty$ process, which means that data sessions are established according to a Poisson process with rate $\lambda$, the session duration is generally distributed with mean $E(S) = 1/\mu$ and the mean number of sessions in the network initiated by a UE is $N_{\mathrm{SE}} = \frac{\lambda}{\mu}$.

With respect to the mobility of UEs we assume that a UE is associated with GWs one after the other for exponentially distributed periods with parameter $\gamma$. Consequently, the end of the visit time at the $k$th GW ($V_k$) is the sum of $k$ independent exponentially distributed random variables, and $V_k^*(s) = (\frac{\gamma}{s+\gamma})^k$.

**Corollary 1.** *The session establishment rates between a pair of UEs (denoted by $\lambda_A$), between an UE and its access GW ($\lambda_B$) and between a pair of GWs ($\lambda_B$) can be calculated as*

$$\lambda_A = \frac{2\lambda}{N_{UE} - 1}, \tag{17}$$

$$\lambda_B = 2\lambda, \tag{18}$$

$$\lambda_C = \frac{2N_{UE}\lambda}{M^2}. \tag{19}$$

*Proof.* Let focus on an UE pair, where $A$ and $B$ denotes the two UEs. $A$ initiates sessions with rate $\lambda$ towards $N_{\mathrm{UE}}-1$ destinations that are uniformly selected. Hence the rate of sessions initiated by $A$ toward $B$ is $\frac{\lambda}{N_{\mathrm{UE}}-1}$. Additionally, the $B$ initiates sessions towards every UE, hence $A$ is the destination with rate $\frac{\lambda}{N_{\mathrm{UE}}-1}$. $\lambda_A$ is the sum of the session establishment rates initiated from $A$ and $B$.

The network segment between the UE and the GW conveys the data sessions between the UE and every peers of the UE. Hence, $\lambda_B = (N_{\mathrm{UE}} - 1)\lambda_A = 2\lambda$.

The network segment between two GWs transfers data for all UE pairs attached to this GW pair. The overall session establishment rate in the network is $N_{\mathrm{UE}}\lambda$. $\frac{M-1}{M}$ is the proportion of session establishments where $B$ is in a separate AN than $A$. $\frac{M(M-1)}{2}$ is the number of GW pairs. The session establishment rate between two GWs is hence

$$\lambda_C = \frac{M-1}{M}\frac{1}{\frac{M(M-1)}{2}}N_{\mathrm{UE}}\lambda = \frac{2N_{\mathrm{UE}}\lambda}{M^2} \tag{20}$$

□

Furthermore, $\widehat{\lambda}_A$, $\widehat{\lambda}_B$, $\widehat{\lambda}_C$ and $E(G_A)$, $E(G_B)$, $E(G_C)$ denote the BEX rates and the mean lengths of the SA period between the node pairs defined in Corollary 1. The relation of the session arrival rate, the mean length of the SA period and the BEX rate is characterized by

$$1/\widehat{\lambda}_i = E(G_i) + E(\overline{G}_i) = E(G_i) + 1/\lambda_i, \tag{21}$$

which is depicted in Fig. 7 without subscripts. Due to the Poisson session arrival with rate $\lambda_i$ the mean time between consecutive SA periods is $E(\overline{G}_i) = 1/\lambda_i$.

3.3. **Analysis of the security association period.** In this section we derive the CDF of the SA period ($G$) both in Laplace and time domain and present numerical methods for computing $\overline{F}_G(t)$ and $f_G(t)$.

**Theorem 3.1.** *Assuming that the number of sessions between two HIP-enabled hosts can be described with an M/G/$\infty$-type queuing service, with session establishment rate $\lambda$, generally distributed session holding time and UAL parameter $T$, the LT of $G$ satisfies*

$$G^*(s) = \frac{pX^*(s)T^*(s)}{1-(1-p)X^*(s)\check{Y}^*(s)}. \tag{22}$$

*Proof.* Using the notations given in Sec. 3.1 the probability of having $j+1$ cycles of busy-idle periods in $G$ is $p(1-p)^j$. Therefore,

$$G = \begin{cases} X+T, & \text{with probability} \quad p, \\ X+\check{Y}+X+T, & \text{with probability} \quad p(1-p), \\ \vdots & \qquad\qquad \vdots \\ i(X+\check{Y})+X+T, & \text{with probability} \quad p(1-p)^i, \\ \vdots & \qquad\qquad \vdots \end{cases} \tag{23}$$

Alternatively, we can utilize the fact that the process renews at each starting point of busy and idle period. If $Y > T$ after the first busy period ($X_1$) of $G$, then the SA period finishes after time $T$. However, if $Y < T$, then the underlying session establishment process renews and the remaining time of $G$ has the same distribution as $G$.

$$G = X + \begin{cases} T, & \text{with probability} \quad p, \\ \check{Y}+G, & \text{with probability} \quad 1-p. \end{cases} \tag{24}$$

Using (23) the PDF of the SA period can be written as

$$f_G(t) = \sum_{j=0}^{\infty} \underbrace{(1-p)^j p}_{\Pr\{\text{(1+j) busy+idle}\}} \underbrace{\{f_X(t) \circledast f_{\check{Y}}(t)\}^j}_{j\cdot(\text{busy+(idle}<T))} \underbrace{\circledast f_X(t) \circledast f_T(t)}_{\text{last busy+}T}, \tag{25}$$

The convolution is transformed to simple multiplication in the Laplace domain. The summation and multiplication with a coefficient in the time domain remain the same operations in the Laplace domain. The Laplace transform (LT) $G^*(s)$ of the PDF $f_G(t)$ is

$$\begin{aligned} G^*(s) = &\int_{t=0}^{\infty} e^{-st} f_G(t) dt = \\ = &\sum_{j=0}^{\infty} \underbrace{(1-p)^j p}_{\Pr\{\text{(1+j) busy+idle}\}} \underbrace{\Big(X^*(s)\check{Y}^*(s)\Big)^j}_{j\cdot(\text{busy+(idle}<T))} \underbrace{X^*(s)T^*(s)}_{\text{last busy+}T} = \\ = &pX^*(s)T^*(s) \sum_{j=0}^{\infty} ((1-p)X^*(s)\check{Y}^*(s))^j = \\ = &\frac{pX^*(s)T^*(s)}{1-(1-p)X^*(s)\check{Y}^*(s)}. \end{aligned} \tag{26}$$

Another way to calculate $G^*(s)$ is based the recursive relation in (24). The Laplace transform of (24) gives

$$G^*(s) = pX^*(s)T^*(s) + (1-p)X^*(s)\check{Y}^*(s)G^*(s) \tag{27}$$

$$G^*(s)(1-(1-p)X^*(s)\check{Y}^*(s)) = pX^*(s)T^*(s) \tag{28}$$

$$G^*(s) = \frac{pX^*(s)T^*(s)}{1-(1-p)X^*(s)\check{Y}^*(s)}. \tag{29}$$

□

**Theorem 3.2.** *When the conditions of Theorem 3.1 hold the CCDF of SA period (G) is*

$$\overline{F}_G(x) = \overline{F}_X(x-T) - \int_{z=0}^{x-T}\int_{y=0}^{T} \lambda e^{\lambda y}\overline{F}_G(x-z-y)dyd\overline{F}_X(z). \tag{30}$$

*Proof.* We evaluate $\overline{F}_G(x) = \Pr\{G > x\}$ conditioned on the length of the first busy period, $X_1$. Fig. 9 illustrates the recursive behavior of the SA period.

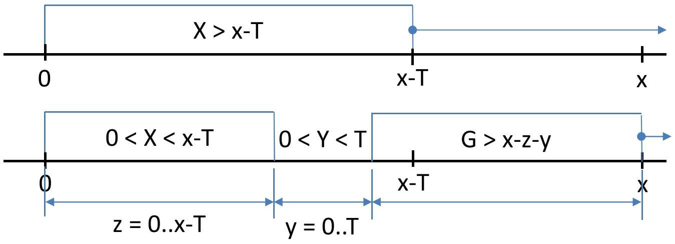


FIGURE 9. Recursive behavior of the SA period.

If $X_1 > x-T$, then $G > x$ is true. This is shown in the upper part of Fig. 9. Since $\Pr\{X_1 > x-T\} = \overline{F}_X(x-T)$, the first member of (30) is $\overline{F}_X(x-T)$.

If $X_1 < x-T$, then $G > x$ holds only if it contains at least two busy periods as presented in the lower part of Fig. 9. The remaining time of $G$ (after the first busy-idle period) must be greater than $x - X_1 - \check{Y}_1$. The distribution of the remaining time of $G$, however, is the same as for $G$ due to the fact that the process renews when it starts a new busy period. This case is reflected in the second term of (30). $\lambda e^{\lambda y}$ and $-d\overline{F}_X(z) = -\frac{d}{dz}\overline{F}_X(z)$ give the probability densities that the first idle and busy periods take $y$ and $z$ time, respectively. □

The calculation of the moments of $G$ based on its CCDF is rather inefficient. Fortunately, due to (30), the CDF and PDF of $G$ can be calculated numerically using Algorithm 1 and 2. As $\Delta t$ tends to 0, the result is longer to compute and more accurate. The time and memory requirement of the calculation is proportional with $x/\Delta t$.

ALGORITHM 1. Numerical calculation of $\overline{F}_G(x)$

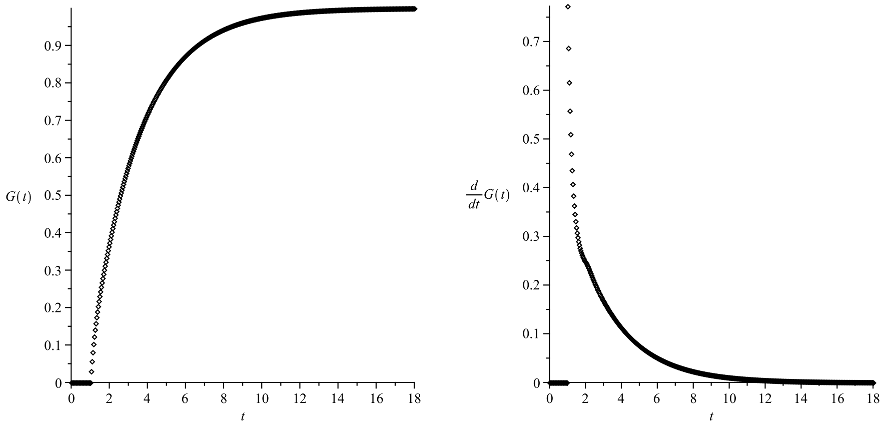
```
$\overline{F}_G(x)$ := proc (x, $\Delta t$)
    $\overline{F}_G(0) = 1$;
    for $u = \Delta t$ to $\lfloor x/\Delta t \rfloor$ stepBy $\Delta t$
        $K = \lfloor u/\Delta t \rfloor$; $N = \lfloor (u-T)/\Delta t \rfloor$; $M = \lfloor T/\Delta t \rfloor$;
        $\overline{F}_G(u) = \overline{F}_X(u-T) -$
```

```
      $-\sum_{n=1}^{N}\sum_{m=1}^{M}\lambda e^{-\lambda m\Delta t}\overline{F}_G((K-n-m)\Delta t)d\overline{F}_X(n\Delta t)(\Delta t)^2$
    end
    return $\overline{F}_G(u)$;
  end proc
```

ALGORITHM 2. Numerical calculation of $f_G(x)$

```
  $f_G(x)$ := proc (x, $\Delta t$)
      $K = \lfloor x/\Delta t \rfloor$;
      $f_G(x) = \frac{\overline{F}_G((K-1)\Delta t)-\overline{F}_G(K\Delta t)}{\Delta t}$
  end proc
```

Fig. 10 illustrates the CDF and PDF of $G$ with parameters $\lambda = 2$ $\mu = 1$, $T = 1$, and $\Delta t = 0.005$.



(A) CDF

(B) PDF

FIGURE 10. The CDF and PDF of the SA period with $\lambda = 2$ $\mu = 1$, $T = 1$, and $\Delta t = 0.005$.

3.4. **The BEX rate and the moments of the SA period.** This section presents an efficient calculation of $\widehat{\lambda}$. From now on we implicitly assume that the modeling assumptions of Sec. 3.2 hold.

**Theorem 3.3.** *The BEX rate between nodes is*

$$\widehat{\lambda} = e^{-\lambda E(S)} e^{-\lambda \cdot T} \lambda, \tag{31}$$

*where $E(S)$ is the mean of session length.*

*Proof.* The steady state probability of having 0 active sessions equals to the limiting probability of having 0 jobs in an M/G/$\infty$ service [8]. That is

$$\Pr\{Q = 0\} = e^{-\lambda E(S)}. \tag{32}$$

$\widehat{\lambda}$ corresponds to the mean rate of transitions from $Q = 0$ to $Q = 1$, which will take longer than time $T$. That can be calculated by multiplying $\lambda$ session inter arrival rate with (32) and the probability that the transition takes longer than $T$. The last term equals with $\Pr\{Y > T\}$ given in (1). I.e.,

$$\widehat{\lambda} = \Pr\{Q = 0\} \cdot \Pr\{Y > T\} \cdot \lambda = e^{-\lambda E(S)} e^{-\lambda \cdot T} \lambda. \tag{33}$$

□

From the Laplace domain description in (22) any moment of $G$ can be calculated by (5). For computing the first and second moments of $G$ we need the first and second moments of $T$, $X$ and $\check{Y}$. For $T$ we have, $E(T) = T$ and $E(T^2) = T^2$. $E(X)$ and $E(X^2)$ are given in (97) and (98) for generally distributed session time distribution. For exponentially distributed session time distribution with rate $\mu$, these equations simplify to (102) and (103). $E(\check{Y})$ is given in (10).

**Corollary 2.** *The second moment of $\check{Y}$ is*

$$E(\check{Y}^2) = \frac{2e^{\lambda T} - 2\lambda T - 2 - T^2\lambda^2}{(e^{\lambda T} - 1)\lambda^2}. \tag{34}$$

*Proof.*

$$\begin{aligned} E(\check{Y}^2) &= \int_{t=0}^{\infty} t^2 f_{\check{Y}}(t) dt = \int_{t=0}^{\infty} 2 \int_{x=0}^{t} x dx f_{\check{Y}}(t) dt = \\ &= \int_{x=0}^{\infty} \int_{t=x}^{\infty} f_{\check{Y}}(t) dt 2x dx = \int_{x=0}^{\infty} (1 - F_{\check{Y}}(x)) 2x dx = \\ &= 2 \int_{x=0}^{T} x \frac{e^{-\lambda x} - e^{-\lambda T}}{1 - e^{-\lambda T}} dx = \frac{2e^{\lambda T} - 2\lambda T - 2 - T^2\lambda^2}{(e^{\lambda T} - 1)\lambda^2}. \end{aligned} \tag{35}$$

$t^2$ has been substituted with $2\int_{x=0}^{t} x dx$, then the order of integration was reversed in order to get $1 - F_{\check{Y}}(x)$ for the inner integral. $F_{\check{Y}}(x)$ is given by (7). □

Let $Z$ denote the second term on the right hand side of (24), i.e.,

$$Z = \begin{cases} T, & \text{with probability} \quad p, \\ \check{Y} + G, & \text{with probability} \quad 1 - p. \end{cases} \tag{36}$$

$E(Z)$ and $E(Z^2)$ can be written as

$$E(Z) = pE(T) + (1 - p)(E(\check{Y}) + E(G)) \tag{37}$$

and

$$E(Z^2) = pE(T^2) + (1 - p)(E(\check{Y}^2) + E(G)^2 + 2E(\check{Y})E(G)). \tag{38}$$

The last term $(2E(\check{Y})E(G))$ equals to $2E(\check{Y}G)$ because the lengths of the first idle period is independent of the length of the remaining busy and idle periods.

**Corollary 3.** *Assuming that the number of sessions between two HIP-enabled hosts is an $M/G/\infty$ process and the UAL is set to $T$, the first moment of $G$ is*

$$E(G) = T + \frac{E(X)}{p} + \frac{(1 - p)E(\check{Y})}{p}. \tag{39}$$

*Proof.* Using (23), $E(G)$ can be expressed as

$$\begin{aligned} E(G) &= T + \sum_{i=0}^{\infty} p(1 - p)^i (E(X) + i(E(X) + E(\check{Y}))) = \\ &= T + \frac{E(X)}{p} + \frac{(1 - p)E(\check{Y})}{p}, \end{aligned} \tag{40}$$

leading to the same result as (39). The recursive relation, (24), can be applied to calculate the first moment of $G$, as well. That is

$$\begin{aligned} E(G) =& T + pE(X) + (1-p)(E(X) + E(\check{Y}) + E(G)) = \\ =& T + E(X) + (1-p)(E(\check{Y}) + E(G)), \end{aligned} \tag{41}$$

from which

$$E(G) = T + \frac{E(X)}{p} + \frac{(1-p)E(\check{Y})}{p}. \tag{42}$$

□

**Corollary 4.** *Under the conditions of Corollary 3 the second moment of $G$ is*

$$E(G^2) = E(X^2) + E(Z^2) + 2E(X)E(Z), \tag{43}$$

*where $Z$ is defined in (36).*

*Proof.*

$$\begin{aligned} E(G^2) =& E((X+Z)^2) = E(X^2 + Z^2 + 2XZ) = \\ =& E(X^2) + E(Z^2) + 2E(X)E(Z), \end{aligned} \tag{44}$$

where we utilized the independence of $X$ and $Z$, that is, $E(XZ) = E(X)E(Z)$. □

The variance of the SA period can be calculated from $Var(G) = E(G^2) - E(G)^2$. In the special case of exponentially distributed session times with parameter $\mu$, substituting (102) and (103) into (39) and (43) gives

$$E(G) = \frac{e^{\lambda(T+\frac{1}{\mu})} - 1}{\lambda}, \tag{45}$$

and

$$\begin{aligned} E(G^2) = -\frac{2}{\lambda^2}\Bigg( & e^{\frac{\lambda(T\mu+1)}{\mu}} T\lambda + e^{\frac{\lambda(T\mu+2)}{\mu}} - e^{\frac{2\lambda(T\mu+1)}{\mu}} - \\ & - e^{\frac{\lambda(T\mu+2)}{\mu}} \lambda \int_{t=0}^{\infty} t\lambda e^{\frac{\lambda e^{\mu t} - \mu^2 t - \lambda}{\mu}} dt \Bigg). \end{aligned} \tag{46}$$

The expressions with general session time distributions are obtained by substituting (97) and (98) into (39) and (43), but are too cumbersome to present here.

3.5. **The mean rates of update procedures.** In UFA HIP, BEX or UPDATEw/CERT is triggered between the GWs of a pair of UEs, as long as there is no HA between the UEs, as presented in Table 1 and depicted by Fig. 5b. A practically important performance measure is the rate of these procedures. The BEX rate between a pair of GWs due to SE is denoted by $\widehat{\lambda}_C$. Its relation to the session arrival rate and the mean length of the SA period is characterized by (21). It can be calculated with Theorem 3.3 using $\lambda_C$, $\mu$ and $T$ as input parameters.

The rate of UPDATEw/CERT procedures between a pair of GWs due to session establishment is indicated by $\lambda_{\mathrm{SE,U}}$, and can be calculated in the following way.

**Theorem 3.4.** *Assuming an $M/G/\infty$-type session establishment process between UEs in a synthetic network model described in Sec. 3.2, the UPDATEw/CERT rate between a pair of GWs due to session establishments in UFA HIP is*

$$\lambda_{SE,U} = \frac{N_{UE}(N_{UE}-1)}{M^2}\widehat{\lambda}_A - \widehat{\lambda}_C \tag{47}$$

*Proof.* In UFA HIP, an SA pair is maintained between a pair of GWs (and between an UE and its GW) until at least one host association (HA) uses it. Fig. 11 illustrates a simple case when the SA is shared between two pairs of UEs. $\lambda_1$ and $\lambda_2$ denote the arrival rates of session establishments under each pair of UEs. HAs are established for $G_1$ and $G_2$ periods with rates $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$, respectively. The SA pair between the GWs is utilized by these two HAs. Within an SA period the first HA establishment triggers BEX, the other HA establishments trigger UPDATEw/CERT.

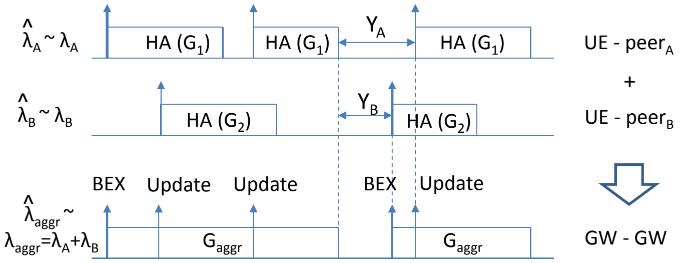


FIGURE 11. Occurrences of BEX and UPDATEw/CERT during session establishment in UFA HIP between a pair of GWs.

Due to the Poisson arrival of session establishments, the remaining time of the inter-arrival periods, denoted by $Y_1$ and $Y_2$ in Fig. 11, are exponentially distributed with parameter $\lambda_1$ and $\lambda_2$. Another consequence of the Poisson arrival of session establishments is the following. When two HAs share one SA pair, as it is in Fig. 11, the probability that the HA establishment of the first UE pair initiates a BEX of the aggregate process (3rd line in Fig. 11) is $\Pr\{Y_1 < Y_2\} = \frac{\lambda_1}{\lambda_1+\lambda_2}$. In general, when $K$ HAs share one SA pair, the probability that the $i^{\text{th}}$ HA initiates a BEX of the aggregate process is $\Pr\{Y_i = \min(\{Y_1..Y_K\})\} = \frac{\lambda_i}{\sum_{k=1}^{K}\lambda_k}$.

Therefore the BEX rate caused by the $i^{\text{th}}$ UE pair is $\widehat{\lambda}_{\text{aggr}}\frac{\lambda_i}{\sum_{k=1}^{K}\lambda_k}$. The UPDATEw/CERT rate caused by the $i^{\text{th}}$ UE pair is $\widehat{\lambda}_i - \widehat{\lambda}_{\text{aggr}}\frac{\lambda_i}{\sum_{k=1}^{K}\lambda_k}$.

In our network model, the proportion of session establishments, which appear on the transport network is $\frac{M}{M-1}$. The number of UE and GW pairs in the network is $\frac{N_{\text{UE}}(N_{\text{UE}}-1)}{2}$ and $\frac{M(M-1)}{2}$, respectively. Hence, the number of UE pairs that occur between a pair of GWs is $K = \frac{\frac{M}{M-1}\frac{N_{\text{UE}}(N_{\text{UE}}-1)}{2}}{\frac{M(M-1)}{2}} = \frac{N_{\text{UE}}(N_{\text{UE}}-1)}{M^2}$.

Therefore, the overall rate of UPDATEw/CERT procedures in UFA HIP between a pair of GWs is

$$\lambda_{\text{SE,U}} = \sum_{i=1}^{K}\left(\widehat{\lambda}_i - \widehat{\lambda}_{\text{aggr}}\frac{\lambda_i}{\sum_{k=1}^{K}\lambda_k}\right). \tag{48}$$

That simplifies to (47), if for $i = 1..K$, $\lambda_i = \lambda_A$ and $\widehat{\lambda}_i = \widehat{\lambda}_A$, and $\widehat{\lambda}_{\text{aggr}} = \widehat{\lambda}_C$. □

Rekeying (RK) procedures are present in both architectures, as illustrated in Figs. 4a and 5. They are triggered $\left\lfloor \frac{G}{T_{KEY}} \right\rfloor - 1$ times during an SA period as presented in Fig. 12. Let $N_1$ denote the number of rekeyings during a $G$ period of Fig. 12 and $\lambda_{\text{RK}}$ denote the rate of rekeying procedures between two HIP-enabled hosts.

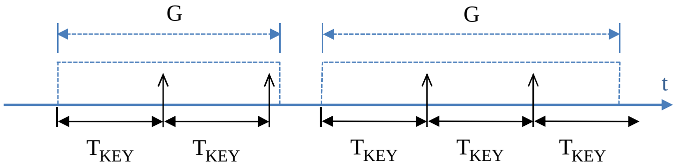


FIGURE 12. Occurences of rekeying in SAs.

**Theorem 3.5.** *Given the CCDF of SA periods between two HIP-enabled hosts, the mean rekeying rate is*

$$\lambda_{RK} = \widehat{\lambda}\sum_{i=1}^{\infty}\overline{F}_G(iT_{KEY}). \tag{49}$$

*Proof.* The probability that there are $i$ updates with DH under an SA period is $\Pr\{N_1 = i\} = \Pr\{iT_{KEY} < G < (i+1)T_{KEY}\}$, from which

$$\begin{aligned} E(N_1) &= \sum_{i=1}^{\infty} i\Pr\{iT_{KEY} < G < (i+1)T_{KEY}\} = \\ &= \sum_{i=1}^{\infty} i(\overline{F}_G((i+1)T_{KEY}) - \overline{F}_G(iT_{KEY})) = \\ &= \sum_{i=1}^{\infty} \overline{F}_G(iT_{KEY}). \end{aligned} \tag{50}$$

$\lambda_{RK} = \frac{E(N_1)}{E(G+\overline{G})} = E(N_1)\widehat{\lambda}$ gives the mean rekeying rate. □

We remark that if $\rho = \widehat{\lambda}E(G) \rightarrow 1$ (e.g., $\rho > 0.95$) then, due to the inefficiency of numerical calculation of $\overline{F}_G(x)$, the following approximation can be applied: $\lambda_{\mathrm{RK}} = \widehat{\lambda}E(G)\frac{1}{T_{\mathrm{KEY}}}$

The RV function introduced in Table 2 is responsible for the following update procedures. In E-E HIP, RVS updates are triggered between an UE and the RVS per each inter-GW handover event. Additionally, the RVS update period is upper constrained by the registration lifetime to RVS, $T_{\mathrm{RVS}}$, as illustrated in Fig. 13.

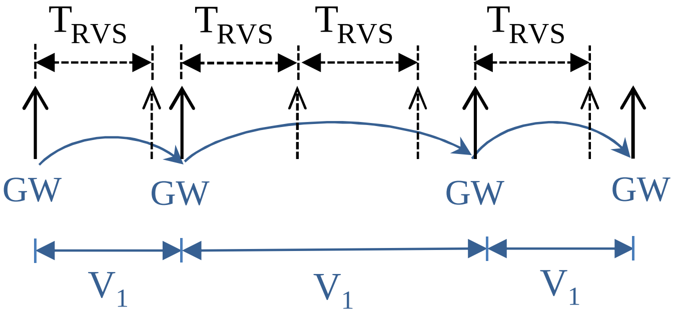


FIGURE 13. Occurrences of RVS updates of an UE in E-E HIP.

Let $N_2$ denote the number of RVS UPDATEs during the visit period of a GW (denoted by $V_1$ in Fig. 13) and $\lambda_{\mathrm{RV}}$ the update rate.

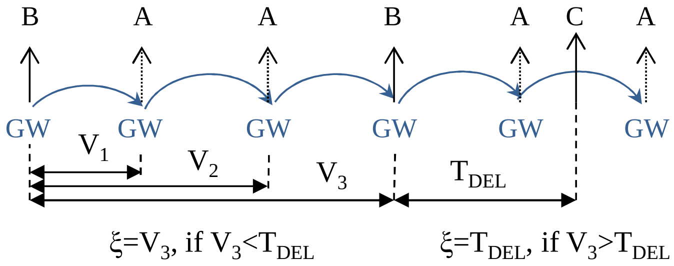


FIGURE 14. Calculation model of the mean rate of update procedures with delegation.

**Corollary 5.** *In E-E HIP, the mean rate of UPDATE procedures under the RV function for an UE can be calculated as*

$$\lambda_{RV,E\text{-}E} = 1 + \sum_{i=1}^{\infty} \overline{F}_{V_1}(iT_{RVS})\gamma \tag{51}$$

*Proof.* By replacing $G$ with $V_1$ and $T_{\text{KEY}}$ with $T_{\text{RVS}}$ in (50), we get $\sum_{i=1}^{\infty} \overline{F}_{V_1}(iT_{\text{RVS}})$ for the mean number of RVS updates in $V_1$. Additionally, there is one update at the start of $V_1$. Hence,

$$E(N_2) = 1 + \sum_{i=1}^{\infty} \overline{F}_{V_1}(iT_{\text{RVS}}). \tag{52}$$

Multiplying $E(N_2)$ with the mean handover rate gives the mean rate of RVS updates from an UE to the RVS. □

In case of UFA HIP, the target GW within an inter-GW handover will initiate the update procedure to the RVS in the name of the UE.

**Lemma 3.6.** *In UFA HIP, the mean rate of UPDATEw/CERT procedures under the RV function for a GW is*

$$\lambda_{RV,UFA} = \frac{\gamma}{M} \cdot \frac{N_{UE}}{M} + \frac{1}{T_{RVS}} \tag{53}$$

*Proof.* Given that $V_1$ is exponentially distributed with parameter $\gamma$ under every GW for all UEs, $\frac{\gamma}{M}$ is the probability that the UE is handed off to a specific GW. $\frac{N_{UEs}}{M}$ is the mean number of UEs served by the GW. Their product results in the mean RVS update rate for every UE attached to the GW. Additionally, a GW re-registers at every $T_{RVS}$ period its IP address at the egress interface and its HIT together with the HIT of all of the delegated UEs. The sum of these two rates results in (53). □

The DR control functions, introduced in Table 2, cause the following update procedures in UFA HIP. There are three different cases for signaling delegation illustrated by Fig. 14. In case A, the source GW propagates signaling rights to the target GW. In this case the source GW generates a new certificate, which extends the signaling delegation chain of the UE. In case B the UE re-delegates signaling rights because the delegation certificate-chain has reached its maximum length ($L$) at the source GW. E.g., $L = 3$ in Fig. 14. In this case the UE creates a new certificate for the target GW and the old delegation chain becomes invalid. Both cases happen at inter-GW handovers.

There is a third case, case C, that is independent of handovers. In this case the UE re-delegates rights to its serving GW by generating and sending a new authorization certificate. This event occurs when the lifetime of the initial authorization by the UE, $T_{\mathrm{DEL}}$, has expired.

A practically important performance measure is the rate at which an UE generates new signaling delegation certificate. This happens in case of events B and C and we denote their rates by $\nu_B$ and $\nu_C$. The period between consecutive certificate generation by an UE is denoted by $\xi$ and depicted in Fig. 14. $\xi_1$ illustrates case B, when the certificate-chain has reached its maximum length before the expiration of $T_{\mathrm{DEL}}$, therefore $\xi_1 = V_L$. $\xi_2$ illustrates case C, when $T_{\mathrm{DEL}}$ lifetime of the delegation certificate-chain expires, hence, $\xi_2 = T_{\mathrm{DEL}}$. The CCDF of $\xi$ is

$$\overline{F}_\xi(t) = \begin{cases} \overline{F}_{V_L}(t), & \text{if} \quad t < T_{\mathrm{DEL}}, \\ 0, & \text{if} \quad t > T_{\mathrm{DEL}}. \end{cases} \tag{54}$$

The mean value of $\xi$ is

$$E(\xi) = \int_{x=0}^{\infty} \overline{F}_\xi(x)dx = \int_{x=0}^{T_{\mathrm{DEL}}} \overline{F}_{V_L}(x)dx. \tag{55}$$

Let $N_A$, $N_B$ and $N_C$ denote the number of events of type A, B and C in $\xi$.

**Corollary 6.** *Considering a UE, in UFA HIP, the mean rates of events where the UE re-delegates a new certificate due to the fact that the maximum delegation chain length had been reached (i.e., case B) or due to the expiration of delegation lifetime ($T_{DEL}$) (i.e., case C) are*

$$\nu_B = \frac{\overline{F}_{V_L}(T_{DEL})}{E(\xi)} \quad \textit{and} \tag{56}$$

$$\nu_C = \frac{F_{V_L}(T_{DEL})}{E(\xi)}, \tag{57}$$

*respectively.*

*Proof.* The probabilities of the occurrence of event B and C under in period $\xi$ can be computed as $p = \Pr\{N_B = 1\} = \Pr\{N_C = 0\} = \Pr\{V_L > T_{DEL}\}$ for event B, and $1 - p = \Pr\{N_C = 1\} = \Pr\{N_B = 0\} = \Pr\{V_L < T_{DEL}\}$ for event C. The mean rates of these events are $\nu_B = E(N_B)/E(\xi) = p/E(\xi)$ and $\nu_C = (1 - p)/E(\xi)$. □

**Theorem 3.7.** *Considering a UE, in UFA HIP, the mean rate of events where the previous GW propagates the UE's signaling delegation authorization to the next GW during HO (i.e., case A) is $\nu_A = E(N_A)/E(\xi)$. The mean number of events A in periods $\xi$ can be calculated as*

$$E(N_A) = (L-1)\overline{F}_{V_{L-1}}(T_{DEL}) + \sum_{k=1}^{L-2} k \int_{x=0}^{T_{DEL}} e^{-\gamma(T_{DEL}-x)} f_{V_k}(x)dx. \tag{58}$$

*Proof.* The calculation requires the specification of probabilities of different number of occurrences of event A under a period $\xi$, i.e.,

$$\Pr\{\xi < T_{\mathrm{DEL}}, N_A = k\} = \Pr\{V_L < T_{\mathrm{DEL}}\}, \qquad \text{if } k = L-1, \tag{59}$$

$$\Pr\{\xi < T_{\mathrm{DEL}}, N_A = k\} = 0, \qquad \text{if } k = 0..L-2, \tag{60}$$

$$\Pr\{\xi = T_{\mathrm{DEL}}, N_A = k\} = \Pr\{V_1 > T_{\mathrm{DEL}}\}, \qquad \text{if } k = 0, \tag{61}$$

$$\Pr\{\xi = T_{\mathrm{DEL}}, N_A = k\} = \Pr\{V_{k+1} > T_{\mathrm{DEL}} > V_k\}, \qquad \text{if } k = 1..L-1, \tag{62}$$

The events can be divided to two categories. When $\xi < T_{\text{DEL}}$, the UE visits $L$ GWs before the expiration of $T_{\text{DEL}}$, hence the number of events A is $L-1$. When $\xi = T_{\text{DEL}}$, the expiration of $T_{\text{DEL}}$ prevents the visit of $L$ GWs with the same certificate-chain. The number of events A depends on the number of visited GWs before re-delegation of rights.

From this, the mean number of events A under a period can be calculated as

$$E(N_A) = (L-1)\Pr\{T_{\text{DEL}} > V_{L-1}\} + \sum_{k=1}^{L-2} k\Pr\{V_{k+1} > T_{\text{DEL}} > V_k\}. \tag{63}$$

The first and second terms of the right hand side of (63) cover the categories when $\xi < T_{\text{DEL}}$ and when $\xi = T_{\text{DEL}}$, respectively. In the first term $\Pr\{T_{\text{DEL}} > V_{L-1}\}$ equals to $F_{V_{L-1}}(T_{\text{DEL}})$. $V_{k+1}$ and $V_k$ are not independent random variables, hence the calculation of $\Pr\{V_{k+1} > T_{\text{DEL}} > V_k\}$ in the second member is the following.

Let introduce the running variable $x$, which marks the current age of the delegation chain. $f_{V_k}(x)$ gives the probability density of being under the $k$th GW at age $x$ of the delegation chain.

Let introduce $u$ for the remaining time under the $k$th GW. The probability of the occurrence of the expiration of the delegation chain (i.e., $x + u > T_{\text{DEL}}$) under that GW is $\Pr\{u > T_{\text{DEL}} - x\}$. Due to the memoryless property of exponential distribution, the remaining time, $u$, has the same distribution as $V_1$. That is, $\Pr\{u > T_{\text{DEL}} - x\} = F_u(T_{\text{DEL}} - x) = F_{V_1}(T_{\text{DEL}} - x) = e^{-\gamma(T_{\text{DEL}}-x)}$.

$f_{V_k}(x)e^{-\gamma(T_{\text{DEL}}-x)}$ is the probability density of having a delegation chain of length $k$ at age $x$ and knowing that it will expire before visiting the $k+1$th GW. Thus, by running $x$ from 0 to $T_{\text{DEL}}$, we can compute the probabilities for reaching different delegation-chain lengths. Therefore, (63) becomes (58). □

Independently of whether the UE or the source GW generates the delegation certificate, it is always the source GW, which conveys the delegation certificate-chain to the target GW in cases of event A and B. This occurs in the second phase of HO in the UPDATEw/CERT procedure between the previous and next GWs, depicted in Fig. 6. The signaling load of this update procedure depends on the length of the certificate-chain. Therefore a practically important question is the mean length of the certificate-chain.

In the following, we introduce index $i$ for indicating the position of the source GW (i.e., the one from which the UE is handed off) in the delegation chain. In Fig. 14 event A occurs when $i \in \{1..L-1\}$, and event B happens when $i = L$.

**Theorem 3.8.** *Under the conditions of Section 3.2 with respect to the mobility behaviour of the UEs, the probability of having k certificates in the certificate-chain, which is sent from the source GW to the target GW in the UPDATEw/CERT procedure in the second phase of the HO for the delegation of signaling rights of the UE to the target GW is*

$$Pr\{k\ cert.\} = \begin{cases} \frac{F_{V_L}(T_{DEL})}{\sum_{l=1}^{L} F_{V_l}(T_{DEL})}, & if \quad k = 1, \\ \frac{F_{V_{k-1}}(T_{DEL})}{\sum_{l=1}^{L} F_{V_l}(T_{DEL})}, & if \quad k = 2..L, \end{cases} \tag{64}$$

*and $\sum_{k=1}^{L} kPr\{k\ cert.\}$ gives the average length of certificate-chain.*

*Proof.* Let $O_{\text{i,i+1}}$ denote the number of handovers from $\text{GW}_i$ to $\text{GW}_{i+1}$ in a $\xi$ period and let $\eta_{\text{i,i+1}}$ indicate the average rate of handovers from from $\text{GW}_i$ to $\text{GW}_{i+1}$. Since

$O_{\mathrm{i,i+1}}$ is a binary random variable, for the mean number of handovers from GW$_i$ to GW$_{i+1}$ in a $\xi$ period we have

$$E(N_{\mathrm{i,i+1}}) = \Pr\{O_{\mathrm{i,i+1}} = 1\} = \Pr\{V_i < T_{\mathrm{DEL}}\} = F_{V_i}(T_{\mathrm{DEL}}), \tag{65}$$

where $\Pr\{O_{\mathrm{i,i+1}} = 1\}$ is the probability that the handover happens, i.e., $V_i$ illustrated in Fig. 14 is smaller than $T_{\mathrm{DEL}}$.

The average rate of handovers from GW$_i$ to GW$_{i+1}$ is

$$\eta_{\mathrm{i,i+1}} = \frac{E(O_{\mathrm{i,i+1}})}{E(\xi)}, \tag{66}$$

and the probability that a handover is from GW$_i$ to GW$_{i+1}$ is

$$\Pr\{\mathrm{i,i+1}\} = \frac{\eta_{\mathrm{i,i+1}}}{\sum_{j=1}^{L} \eta_{\mathrm{j,j+1}}}. \tag{67}$$

Now, substituting (65) into and than (66) into (67) results in (64). □

In UFA HIP, a GW conveys the delegation certificate together with public key signatures, when acting in the name of UE. The length of the delegation certificate-chain is an important parameter, which has influence on the number and length of update messages in mandated update procedures.

Let $E(N_{\mathrm{certs}})$ denote the mean certificate-chain length of a UE at a delegate GW over time and $A_i$ ($i = 1..L$) the lifetimes of certificate-chains with length $i$. $A_1$ corresponds to the visit time of the GW, which got the certificate directly from the UE. The visit time however is upper-bounded by $T_{\mathrm{DEL}}$. $A_2$ is the visit time of the UE under the second GW. $E(A_i)$ is the general form of the mean visit time of a GW by the UE, at which GW the certificate-chain has length $i$ ($i \leq L$). For $A_i$ we have

$$E(A_1) = \int_{x_1=0}^{T_{\mathrm{DEL}}} x_1 \gamma e^{-\gamma x_1} dx_1 + T_{\mathrm{DEL}} e^{-\gamma T_{\mathrm{DEL}}}, \tag{68}$$

$$E(A_2) = \int_{x_1=0}^{T_{\mathrm{DEL}}} \gamma e^{-\gamma x_1} \left( \int_{x_2=0}^{T_{\mathrm{DEL}}-x_1} x_2 \gamma e^{-\gamma x_2} dx_2 + (T_{\mathrm{DEL}} - x_1) e^{-\gamma T_{\mathrm{DEL}} - x_1} \right) dx_1, \tag{69}$$

$$\begin{aligned} E(A_i) = \int_{x_1=0}^{T_{\mathrm{DEL}}} & \gamma e^{-\gamma x_1} \left( \int_{x_2=0}^{T_{\mathrm{DEL}}-x_1} \gamma e^{-\gamma x_2} \cdots \left( \int_{x_{i-1}=0}^{T_{\mathrm{DEL}}-x_1-x_2-\cdots-x_{i-2}} \gamma e^{-\gamma x_{i-1}} \right.\right. \\ & \left( \int_{x_i=0}^{T_{\mathrm{DEL}}-x_1-x_2-\cdots-x_{i-1}} x_i \gamma e^{-\gamma x_i} dx_i + \right. \\ & \left. + (T_{\mathrm{DEL}} - x_1 - x_2 - \cdots - x_{i-1}) e^{-\gamma T_{\mathrm{DEL}} - x_1 - x_2 - \cdots - x_{i-1}} \right) \\ & \left.\left. dx_{i-1} \cdots \right) dx_2 \right) dx_1, \end{aligned} \tag{70}$$

and in accordance with the definition of $A_i$ the integrals satisfy

$$E(\xi) = \sum_{i=1}^{L} E(A_i). \tag{71}$$

**Theorem 3.9.** *The average length of certificate-chain that a delegate GW conveys in mandated update procedures can be calculated as*

$$E(N_{certs}) = \sum_{i=1}^{L} i \cdot \frac{E(A_i)}{\sum_{j=1}^{L} E(A_j)}. \tag{72}$$

*Proof.* The stationary mean certificate-chain length can be calculated based on the mean time with certificate-chains of length $i$, i.e.,

$$\Pr\{\text{stationary certificate-chain length} = i\} = \frac{E(A_i)}{\sum_{j=1}^{L} E(A_j)}, \tag{73}$$

from which the mean length of certificate-chain over time is (72). □

3.6. **Stationary number of host and security associations.** A suitable descriptor for the memory consumption of E-E and UFA HIP is the mean number of HAs and SAs in the UE, GW and RVS. Let $C_{i,j}$ and $D_{i,j}$ denote the mean number of HA and SA database entries in a network element, respectively, such that index $i$ indicates the relation and index $j$ the HIP type as follows. $i = 1, 2, 3$ stands for UE, GW, RVS and $j = 1, 2$ stands for E-E and UFA HIP, respectively. Based on the BEX rate and the mean SA period between two peers, $C$ can be calculated using Little's rule. In E-E HIP,

$$C_{1,1} = (N_{\text{UE}} - 1)\widehat{\lambda}_A E(G_A) \qquad \text{and} \qquad D_{1,1} = 2C_{1,1}. \tag{74}$$

In UFA HIP, $C_{1,2} = C_{1,1}$, i.e., the number of HAs equals to that of the E-E HIP due to the same traffic demands. However, only one SA pair is required from the UE to the serving GW. This SA pair is kept active while the considered UE establish sessions with a peer UE, following the behavior demonstrated in Fig. 11. Therefore,

$$D_{1,2} = 2\widehat{\lambda}_B E(G_B). \tag{75}$$

In E-E HIP every UE has HA with the RVS, hence $C_{2,1} = N_{\text{UE}}$ and $D_{2,1} = 2C_{2,1}$. In case of UFA HIP, the GWs establish HAs with the RVS as well. Therefore, $C_{2,2} = N_{\text{UE}} + M$. However, only the GWs establish SA pairs with the RVS, hence $D_{2,2} = 2M$.

In UFA HIP, on the GW side, the SA entries include the SA pairs maintained together with the UEs in the AN, with the GWs in the TN and with the RVS. Therefore,

$$D_{3,2} = \frac{N_{UEs}}{M} 2\widehat{\lambda}_B E(G_B) + (M-1)2\widehat{\lambda}_C E(G_C) + 2. \tag{76}$$

On the other hand the HAs include one HA for each SA pair and one for each pair of delegated UEs and their peers. Hence

$$C_{3,2} = \frac{D_{3,2}}{2} + \frac{N_{\text{UE}}}{M}(N_{\text{UE}} - 1)\widehat{\lambda}_A E(G_A). \tag{77}$$

4. **Analysis of signaling overhead in E-E and UFA HIP based mobile networks.** The objective of this section is twofold. First, we show that the introduced analytical model is applicable for the evaluation of the overhead of HIP control in E-E and UFA HIP. Second, some conclusions on the performance benefits and drawbacks of the signaling schemes can be stated based on the results.

Table 3 includes the selected input parameters.

We defined the subcategories "source" AN and "destination" AN for the differentiation of HIP procedures at the UE's and the remote UE's side, respectively.

Table 3. Input parameters.

| Parameter | Value | Parameter | Value | Parameter | Value |
|---|---|---|---|---|---|
| $N$ | 1E+06 | $\lambda$ | 1/10 $\text{min}^{-1}$ | $T$ | 15 min |
| $M$ | 1E+04 | $\mu$ | 1/30 $\text{min}^{-1}$ | $T_{\text{KEY}}$ | 6 h |
| $\alpha$ | 1 $\text{day}^{-1}$ | $\gamma$ | $M$/1E+06 $\text{min}^{-1}$ | $T_{\text{DEL}}$ | 1 h |
| $\omega$ | 1 $\text{day}^{-1}$ | $L$ | 3 | $T_{\text{RVS}}$ | 1 h |

Physically, these overheads are aggregated in the same set of ANs. We introduced two subcategories with respect to the TN. The first contains the signaling between the GWs and the RVS, while the second subcategory refers to the signaling between GW pairs.

Fig. 15 illustrates the signaling overhead at different network segments grouped by control functions (defined in Tables 1 and 2) and HIP procedure types. For each control function, the upper horizontal bar represents the load in UFA HIP, while the lower bar shows the load in E-E HIP. The used performance metric is the mean rate of the procedures.

TA induces $N_{\text{UE}}\alpha$ process rate in both architectures, however, it appears between the UE and RVS in E-E HIP and between the UE and GW in UFA HIP. This can be seen in Figs. 15a and 15c under TA.

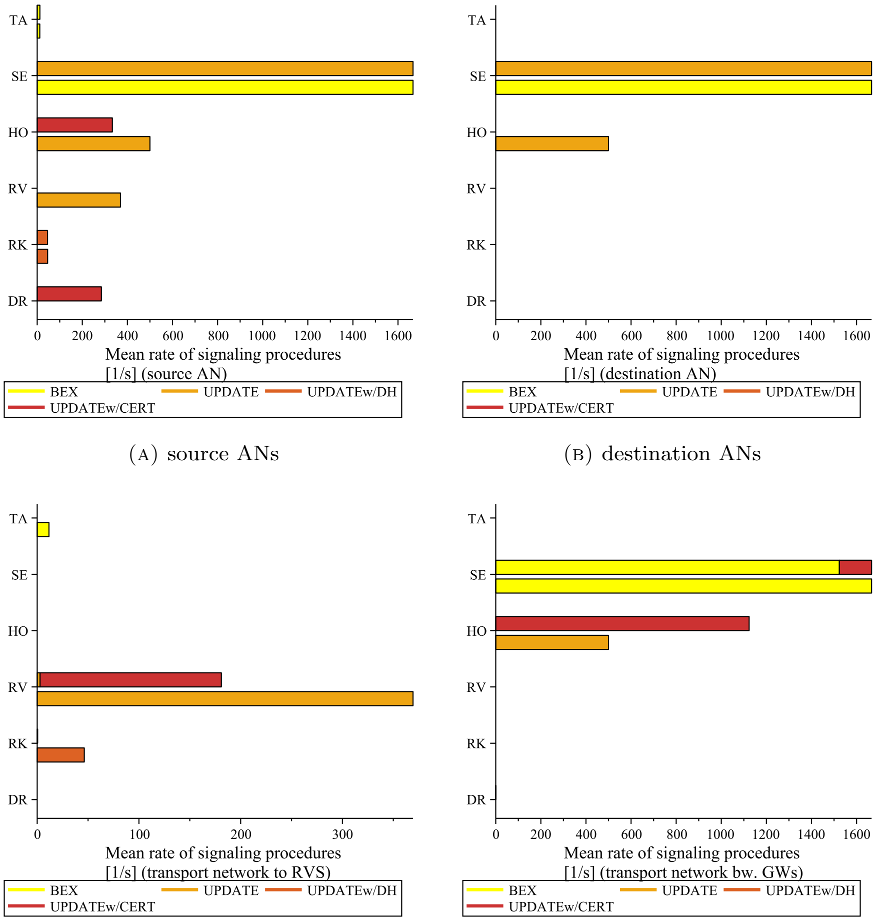
SE induces BEX and UPDATE procedures in E-E HIP and UFA HIP, respectively, with the same rate, $N_{\text{UE}}(N_{\text{UE}}-1)/2\widehat{\lambda}_A$, in the source and destination ANs as presented in Figs. 15a and 15b. In E-E HIP, $\frac{M-1}{M}$ ratio appears from that signaling overhead between the GWs (depicted in Fig. 15d under SE). In UFA HIP the same overhead appears but it is divided between BEX and UPDATEw/CERT as given by $\widehat{\lambda}_C$ and (47), respectively.

The UPDATE rate due to HOs is $N_{\text{UE}}N_{\text{SE}}\gamma$ in E-E HIP as depicted in Fig. 4b. In UFA HIP, one HO event induces two UPDATEw/CERT procedures in the source AN, as illustrated in Fig. 6. HO does not induce signaling in the destination AN, because the delegate GW of the remote UE handles location update in the name of the remote UE. Hence, the overall rate in the ANs is $2N_{\text{UE}}\gamma$. These can be seen in Figs. 15a and 15b under HO. The ratio of HO process rates in UFA compared to E-E HIP in the transport network (TN) between the GWs can be explained by watching Figs. 6 and 4b. In UFA HIP, the probability that an UE finds an established SA between the GW and its next GW at HA initiation is proportional to the time averages of $G_C$ and $\overline{G}_C$. That is,

$$Pr\{\text{established SA}\} = \frac{E(G_C)}{E(G_C) + E(\overline{G}_C)}, \tag{78}$$

where we use the PASTA (Poisson Arrival Sees Time Average) property (see e.g., [8]). In the TN, one HO event causes 2 updates between the source and target GW, $(1 - \Pr\{\text{SA established}\})N_{SE}$ updates toward the UE's peers in phase one of HO (due to the need for SA pair creation between the target GW and the UE's peer), and $N_{SE}$ in the second phase of HO (due to location updates at the UE's peers). Therefore, the handover of one UE triggers $2 + N_{SE}(1 + (1 - \Pr\{\text{SA established}\}))$ update procedures in UFA HIP. On the other hand, there are $N_{\text{SE}}$ update procedures in E-E HIP between the GWs. These rates appear in Fig. 15d under HO.

RV induces signaling overhead given by the rates in (51) and (53). RK causes a signaling overhead with rate given by (49). DR, in Fig. 15a, triggers signaling



(A) source ANs

(B) destination ANs

(C) core transport, GW-to-RVS

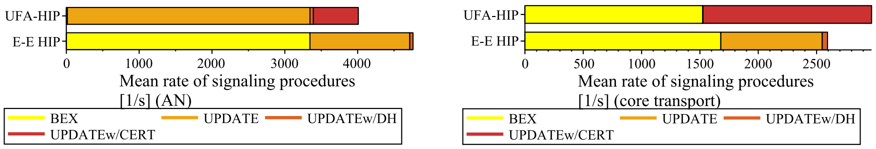
(D) core transport, GW-to-GW

FIGURE 15. Signaling process rates in different network segments grouped by control functions and HIP procedure types.

procedures with rate $N_{\mathrm{UE}}\lambda_C + N\alpha$. That includes signaling right delegation in the terminal attachment phase and due to the expiration of $T_{\mathrm{DEL}}$ (i.e., case C in Fig. 14). The other cases (A and B) of right delegations are included in the overhead of signaling due to HO.

Fig. 16 summarizes the overall signaling load in the AN and TN. We can state that in our network model specified in Sec. 3.2, using the input parameters given in Table 3, UFA HIP performs better in ANs and worse in the TN than E-E HIP. The different HIP procedure types, however, cause different job-sizes for CPU, memory and network processing, which will lead to different proportions in terms of required CPU time or network throughput.

In [16] we compared the performances of E-E HIP and UFA HIP architectures using the presented analytical model. In the analysis, the performance costs of different signaling procedures in terms of number and size of messages and CPU time were derived from real testbed-based measurements. Table 4 shows the achieved



(A) ANs, process rate

(B) core transport, process rate

FIGURE 16. Overall signaling overhead in the ANs and TN.

average resource utilization gains at different parts of the networks, i.e., the UE, RVS, all ANs and all core transport network. The input parameters of the 'low mobility' scenario (Scen. 1) were the same as given in Table 3. In case of the 'high mobility' scenario (Scen. 2) the mobility rate, $\gamma$, of the UEs was ten times higher. In the 'high lifetimes' scenario (Scen. 3) $T_{\text{UAL}}$, $T_{\text{KEY}}$ and $T_{\text{DEL}}$ were set to high values, i.e., 1 day, 1 week and 1 day, respectively. In the 'low lifetimes' scenario (Scen. 4) these lifetime parameters were set to 0 second, 1 hour and 15 minutes, respectively.

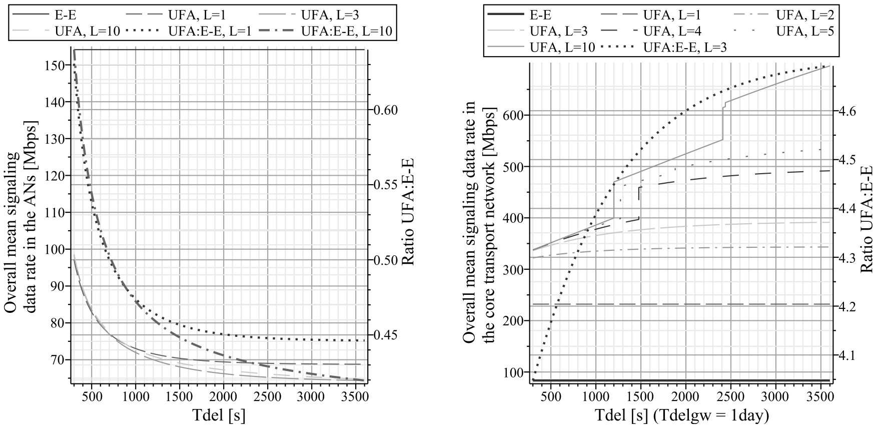
TABLE 4. Gains of UFA HIP compared to E-E HIP. (Gain = $(1 - \frac{P_{\text{UFA}}}{P_{\text{E-E}}}) \cdot 100\%$.)

| | Scen. 1<br>Low mobility | Scen. 2<br>High mobility | Scen. 3<br>High lifetimes | Scen. 4<br>Low lifetimes |
|---|---|---|---|---|
| CPU utilization at the UE | 62% | 71% | 67% | 47% |
| CPU utilization at the RVS | 74% | 15% | 58% | 91% |
| signaling data rate in the ANs | 59% | 56% | 64% | 50% |
| signaling data rate in the core transport | −62% | −350% | −74% | −32% |

We are now able to see the consequences of the introduction of signaling delegation service in HIP. Regarding the influence of maximum delegation chain length, the greater is the value of $L$, slightly higher is the utilization of the network in all parts. Consequently, assuming 1024-bit RSA-based signaling delegation certificates, 1260 byte MTU (minimum requirement) and the application of fragmentation method recommended for HIP CERT fields [13] , the best performance can be achieved if the UE re-delegates the right at each handover to the GW, i.e., set $L = 1$.

Fig. 17 illustrates the mean signaling data rates in the access network and core transport network in case of 'high mobility' scenario (Scen. 2). As $T_{\text{DEL}}$ increases, it enables the development of longer delegation chains, which means more update messages within the delegated update procedures due to HIP packet fragmentation. In Fig. 17b packet fragmentation causes the steps in the signaling load of the core transport network in case of $L = \{4, 5, 10\}$.

5. **Conclusions.** This paper presented the cumulative distribution function and moments of the SA periods in secure mobile networks. Many protocols use lifetimes with similar role as the unused association lifetime, where the underlying renewal process consists of the busy-idle periods of a communication channel. Thus, the results for the CDF and moments of SA perios may be re-utilized in the performance analysis of other protocols as well. The CDF of SA period has been utilized in the calculation of other parameters related to update procedure rates.



(A) signaling data rate in the AN.

(B) signaling data rate in the core transport.

FIGURE 17. Resource utilization in function of the maximum length of delegation chain ($L$) and $T_{\mathrm{DEL}}$ in the 'high mobility' scenario.

We established a detailed analytical model for the comparison of the signaling overheads of E-E HIP and UFA HIP architectures in a synthetic network model. We showed that the model can be utilized to calculate signaling process rates. We notice that other advanced mobility management solutions in HIP layer, such as HIP micromobility or network mobility schemes [17], also require delegation of signaling rights. Our analytical results could also be adapted for the performance analysis of those solutions.

UFA HIP proved to come up to the expectations regarding the reduction of network overheads due to signaling delegation. It performs better than E-E HIP at the access networks and towards the RVS, but performs similarly or worse in the core transport network.

As a continuation of this paper, we investigated several engineering questions in [16]. E.g., what are optimal configurations of lifetime parameters and the optimal number of GWs that will keep minimal the signaling overhead in these architectures. What is the tradeoff between increased number of states and longer SA periods, but at the same time higher rekeying rate, due to setting higher the unused association lifetimes. With respect to signaling delegation, how many times should the propagation of delegation rights from delegate to delegate enabled, and how high should the delegation lifetime be set. These influence the average length of delegation-chains in the network, hence the size of update packets conveying certificates, but also the load of the UEs, GWs and RVS.

**Acknowledgments.** The work of the first author has been carried out in the framework of CELTIC-Plus project CP2012/2-5 SIGMONA and supported by the "Virtualization-based Mobile Network Optimization" project of the Hungarian National Development Agency (EUREKA HU 12-1-2012-0054). The second author acknowledges the support of the OTKA K101150 project.

**Appendix** A. **Moments of the busy period in infinite server queues.** One of our basic assumptions is that the behavior of the session process, that is the number of active sessions in Fig. 7, can be modeled using an M/G/$\infty$ queue. This section presents the calculation of the busy period of M/G/$\infty$ queue following and also correcting the typos of the Kulkarni [8, p. 425].

Let $H(t)$ denote the probability that the system is idle at time $t$ assuming it is idle at $t = 0$, i.e.,

$$H(t) =\Pr\{Q(t) = 0|Q(0) = 0\}. \tag{79}$$

**Corollary 7.** *$H(t)$ can be expressed as*

$$H(t) =e^{-\lambda \int_{u=0}^{t}(1-F_S(u))du}, \tag{80}$$

*where $F_S(u)$ is the CDF of the service time.*

*Proof.* $H(t)$ can be obtained by dividing the $(0, t)$ interval into small intervals of length $\Delta$. The $i$th of such interval is $(i\Delta,(i + 1)\Delta)$. In any of these intervals the probability of 0, 1, and more than 1 arrivals are $1 - \lambda\Delta + o(\Delta)$, $\lambda\Delta + o(\Delta)$ and $o(\Delta)$, respectively, where $o(\Delta)$ is such that $\lim_{\Delta\to 0} \frac{o(\Delta)}{\Delta} = 0$. The system is idle at time $t$, if for all small intervals either there is no arrival or there is arrival and its serviced before time $t$. That is

$$H(t) =\prod_{i=1}^{N}\Bigg(\underbrace{1-\underbrace{\lambda\Delta}_{\text{prob. of arrival}}\Big(\underbrace{1-F_S(t-i\Delta)}_{\text{prob. of service incompletion}}\Big)}_{\text{prob. of service completion}}\Bigg). \tag{81}$$

For time $t + \Delta$, we have

$$\begin{aligned} H(t+\Delta) &= \prod_{i=1}^{N+1}\Big(1-\lambda\Delta(1-F_S(t+\Delta-i\Delta))\Big) = \\ &= \prod_{i=1}^{N+1}\Big(1-\lambda\Delta(1-F_S(t-(i-1)\Delta))\Big), \end{aligned} \tag{82}$$

from which

$$\frac{H(t+\Delta)}{H(t)} =1-\lambda\Delta(1-F_S(t). \tag{83}$$

Substituting $H(t+\Delta)$ with $H(t)+H'(t)\Delta+o(\Delta)$ and multiplying both sides with $H(t)$ gives

$$H'(t)\Delta + o(\Delta) = -H(t)\lambda\Delta(1 - F_S(t)). \tag{84}$$

Dividing both sides by $\Delta$ and making the $\Delta \to 0$ limit result in the following differentiation equation

$$H'(t) = -H(t)\lambda(1 - F_S(t)), \tag{85}$$

whose initial condition is

$$H(0) =1. \tag{86}$$

The solution of $H(t)$ is given in (80). □

**Corollary 8.** *In the special case of exp$\{\mu\}$ distributed service times (i.e., $F_S(t) = 1 - e^{-\mu t}$), $H(t)$ simplifies to*

$$H(t) = e^{-\lambda(1-e^{-\mu t})/\mu}. \tag{87}$$

The LST of $H(t)$ can be calculated as

$$\widetilde{H}(s) = \int_{t=0-}^{\infty} e^{-st} dH(t) = 1 + \int_{t=0+}^{\infty} e^{-st} dH(t), \tag{88}$$

because $H(0-) = 0$ and $H(0+) = 1$.

**Corollary 9.** *For general ($F_S(t)$) and exp$\{\mu\}$ distributed service time we have*

$$\widetilde{H}(s) =1 - \int_{t=0+}^{\infty} e^{-st} \lambda (1 - F_S(t)) e^{-\lambda \int_{u=0}^{t}(1-F_S(u))du} dt, \tag{89}$$

*and*

$$\widetilde{H}(s) =1 - \int_{t=0+}^{\infty} e^{-st} \lambda e^{-\mu t} e^{-\frac{\lambda(1-e^{-\mu t})}{\mu}} dt, \tag{90}$$

*respectively.*

The relation of $H(t)$ and the distribution of the busy period of the M/G/$\infty$ queue is presented in the following theorem.

**Theorem A.1.** *The LST of the busy period of a M/G/$\infty$ queue, $X$, is*

$$\widetilde{X}(s) = E(e^{-sX}) =1 + \frac{s}{\lambda} \cdot \frac{\widetilde{H}(s) - 1}{\widetilde{H}(s)}. \tag{91}$$

*Proof.* We evaluate $H(t)$ conditioned on the end of the first idle-busy cycle of the M/G/$\infty$ queue $P_1$. The conditional probability that the queue is idle at time $t$ is

$$\begin{aligned} H(t|P_1 = x) &= \Pr\{Q(t) = 0 \mid Q(t) = 0, P_1 = x\} \\ &= \begin{cases} H(t-x), & \text{if} \quad 0 \le x \le t, \\ \Pr\{Y_1 > t | P_1 = x\}, & \text{if} \quad\quad x > t, \end{cases} \end{aligned} \tag{92}$$

where $Y_1$ denotes the first arrival instance. In the first case, when $x < t$, the queue renews at $x$ and the probability of being idle at time $t$ is $H(t-x)$. In the second case, when $t < x$, the system is still in the first idle-busy cycle at time $t$. In this case the probability of being idle at time $t$ is identical with the probability that the first customer arrives after time $t$, i.e., $Y_1 > t$. $H(t)$ can be obtained from $H(t|P_1 = x)$ based on the law of total probability. When $F_P(x)$ is the CDF of the length of a idle-busy cycle ($P = Y + X$), then

$$\begin{aligned} H(t) &= \int_{t}^{\infty} \Pr\{Y > t | P_1 = x\} dF_P(x) + \int_{0}^{t} H(t-x) dF_P(x) = \\ &= \int_{0}^{\infty} \Pr\{Y > t | P_1 = x\} dF_P(x) + \int_{0}^{t} H(t-x) dF_P(x). \end{aligned} \tag{93}$$

If $x < t$, then the probability that $Y > t$ is 0, i.e., $\Pr\{Y > t | P_1 = x\} = 0$. Hence, the integration interval of the first part can be extended from $[t, \infty]$ to $[0, \infty]$. The first part in the summation is the CCDF of the arrival time ($\Pr\{Y > t\} = 1 - F_Y(t) = e^{-\lambda t}$).

The equation hence leads to

$$H(t) = e^{-\lambda t} + \int_{x=0}^{t} H(t-x) dF_P(x). \tag{94}$$

Taking Laplace-Stieltjes Transform on both sides, we have

$$\begin{aligned} \widetilde{H}(s) =& \frac{s}{s+\lambda} + \widetilde{H}(s)\widetilde{P}(s) = \\ =& \frac{s}{s+\lambda} + \widetilde{H}(s)\widetilde{X}(s)\frac{\lambda}{s+\lambda}, \end{aligned} \tag{95}$$

where $\widetilde{P}(s) = \widetilde{X}(s)\frac{\lambda}{s+\lambda}$, i.e., one renewal period consists of an idle ($\exp\{\lambda\}$) and the searched busy period.

The LST of the CDF of the busy period is hence

$$\widetilde{X}(s) = 1 + \frac{s}{\lambda} \cdot \frac{\widetilde{H}(s) - 1}{\widetilde{H}(s)}. \tag{96}$$

□

**Corollary 10.** *The mean value of the busy period is*

$$\begin{aligned} E(X) =& \frac{1}{\lambda} \cdot \frac{1-\widetilde{H}(s)}{\widetilde{H}(s)} + \frac{s}{\lambda} \cdot \frac{d}{ds}\frac{1-\widetilde{H}(s)}{\widetilde{H}(s)} = \\ =& -\frac{1}{\lambda} + \frac{1}{\lambda\widetilde{H}(0)}. \end{aligned} \tag{97}$$

*The second moment of the busy period is*

$$E(X^2) = \frac{2}{\lambda\widetilde{H}(0)^2} \cdot \frac{d}{ds}\widetilde{H}(s)\bigg|_{s=0}. \tag{98}$$

*Proof.* Since the relation between LST and LT, (2) , is true for $X$, (5) can be used to calculate the moments of $X$.

$$E(X) = -\frac{d}{ds}\widetilde{X}(s)\bigg|_{s=0} \tag{99}$$

$$E(X^2) = \frac{d^2}{ds^2}\widetilde{X}(s)\bigg|_{s=0} \tag{100}$$

We note that in order to calculate the first moment of busy period it is not necessary to know the exact analytic expression for $\widetilde{H}(s)$. Only $\lim_{s\to 0}\widetilde{H}(s)$ given by (101) is needed. For this we know that

$$\begin{aligned} \lim_{s\to 0}\widetilde{H}(s) = & \int_{t=0-}^{\infty} e^{-st} dH(t)\bigg|_{s=0} = \int_{t=0-}^{\infty} dH(t) = \\ =& H(\infty) - H(0-) = H(\infty). \end{aligned} \tag{101}$$

$H(0-) = 0$, because $\Pr\{Q(t) = 0 | Q(0) = 0\} = 0$ at $t < 0$. On the other hand, $H(\infty) = e^{-\lambda E(S)}$, where $S$ is the service time and $E(S)$ is its mean value. This is the stationary probability of the idle state of the system. □

For the calculation of the second moment of the busy period the numerical computation of $\frac{d}{ds}H(s)\bigg|_{s=0}$ is required, which is a too cumbersome integral expression to present here. This integral can be computed with arbitrary precision with common numerical packages.

In the special case when the service time is exponentially distributed with parameter $\mu$, the M/G/$\infty$ queuing system simplifies to a M/M/$\infty$ queuing system for which the first and second moments of busy period simplify to

$$E(X) = \frac{e^{\lambda/\mu} - 1}{\lambda} \tag{102}$$

and

$$E(X^2) = \frac{1}{\lambda} 2e^{2\lambda/\mu} \int_0^\infty t\lambda e^{\frac{\lambda e^{-\mu t} - \mu^2 t - \lambda}{\mu}} dt, \tag{103}$$

respectively. For the calculation of (103) the numerical computation of

$$\frac{d}{ds} H(s)\bigg|_{s=0} = \int_{t=0}^\infty t\lambda e^{(\lambda e^{-\mu t} - \lambda - \mu^2 t)/\mu} dt \tag{104}$$

is required.

**Appendix** B. **CCDF of busy period in infinite server queues.** Daley provided an integral description of the CCDF of busy period ($\overline{F}_X(x)$) for M/G/$\infty$ system [9]

$$\overline{F}_X(x) = \overline{F}_S(x) + \int_{t=0}^x \overline{F}_X(x-t) e^{-\lambda \tilde{\overline{F}}_S(t)} \lambda \left[\overline{F}_S(t) - \overline{F}_S(t)\right] dt, \tag{105}$$

where $\tilde{\overline{F}}_S(t) = \int_{u=0}^t \overline{F}_S(u) du$. (105) is an implicit expression which can be calculated numerically, e.g., with Algorithm 3.

Algorithm 3. Numerical calculation of $\overline{F}_X(x)$

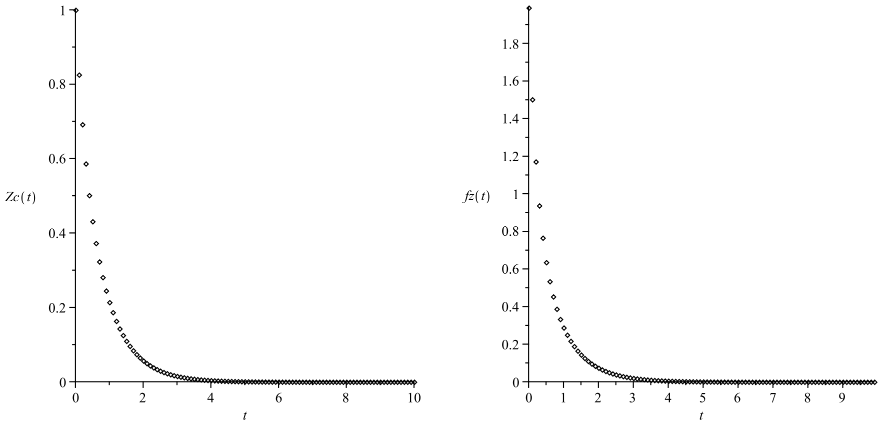
```
$\overline{F}_X(x)$ := proc (x, $\Delta t$ )
    $\overline{F}_X(0) = \overline{F}_S(0)$;
    for $u = \Delta t$ to $\lfloor x/\Delta t \rfloor$ stepBy $\Delta t$
        $N = \lfloor u/\Delta t \rfloor$;
        $\overline{F}_X(u) = \overline{F}_S(u) +$
            $+ \sum_{n=1}^N \overline{F}_X(u - n\Delta t) e^{-\lambda \tilde{\overline{F}}_S(n\Delta t)} \lambda \left[\overline{F}_S(n\Delta t) - \overline{F}_S(u)\right] \Delta t$
    end
    return $\overline{F}_X(u)$;
end proc
```

The PDF of $X$ can be calculated numerically from the CCDF, e.g., with Algorithm 4.

Algorithm 4. Numerical calculation of $f_X(x)$

```
$f_X(x)$ := proc (x, $\Delta t$)
    $K = \lfloor x/\Delta t \rfloor$;
    $f_X(x) = \frac{\overline{F}_X((K-1)\Delta t) - \overline{F}_X(K\Delta t)}{\Delta t}$
end proc
```

Fig. 18 illustrates the CDF and PDF of the busy period for an M/M/$\infty$ queue, using $\exp\{\lambda = 1\}$ arrival times, $\exp\{\mu = 2\}$ service times, and $\Delta t = 0.005$ step-size for the numerical calculation.



(A) CCDF

(B) PDF

FIGURE 18. CCDF and PDF of the busy period of M/M/$\infty$-type Markov chain ($\lambda = 1$, $\mu = 2$, $\Delta t = 0.005$).

## REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2013-2018," White Paper, Cisco, Feb 5, 2014. Available from: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf

[2] K. Daoud, P. Herbelin, and N. Crespi, *UFA: Ultra Flat Architecture for high bitrate services in mobile networks*, Proceedings of the IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2008), Cannes, France, Sep. 15–18, 2008, 1–6.

[3] Z. Faigl, L. Bokor, P. Neves, K. Daoud, and P. Herbelin, *Evaluation of Two Integrated Signalling Schemes for the Ultra Flat Architecture using SIP, IEEE 802.21, and HIP/PMIP Protocols*, Computer Networks, **55(7)** (May 2011), 1560–1575.

[4] Lászlo Bokor, Zoltán Faigl, and Sándor Imre, *A Delegation-based HIP Signaling Scheme for the Ultra Flat Architecture*, Proceedings of the 2nd International Workshop on Security and Communication Networks (IWSCN'10), Karlstad, Sweden, May 26–28, 2010, 9–16.

[5] R. Moskowitz et al, "Host Identity Protocol," RFC 5201, IETF, April 2008. Available from: http://tools.ietf.org/rfc/rfc5201.txt.

[6] A. Gurtov, M. Komu, and R. Moskowitz. *Host Identity Protocol (HIP): Identifier/Locator Split for Host Mobility and Multihoming*, Internet Protocol Journal, **12(1)** (2009), 27–32.

[7] Pekka Nikander and Jari Arkko, *Delegation of Signalling Rights*, in "Security Protocols, Lecture Notes in Computer Science" (eds. Bruce Christianson, Bruno Crispo, James A. Malcolm, and Michael Roe), **2845**, (2004), 575–586.

[8] Vidyadhar G. Kulkarni, "Modeling and analysis of stochastic systems," 2nd edition, Chapman & Hall, Ltd., London, UK, 2009.

[9] D. J. Daley, *The Busy Period of the M/GI/$\infty$ Queue*, Queueing Syst. Theory Appl., **38(2)** (June 2001), 195–204.

[10] E. Rescorla, "Diffie-Hellman Key Agreement Method," RFC 2631, IETF, June 1999. Available from: http://tools.ietf.org/rfc/rfc2631.txt.

[11] T. Kivinen and M. Kojo, "More Modular Exponential (MODP) Diffie-Hellman groups for Internet Key Exchange (IKE)," RFC 3526, IETF, May 2003. Available from: http://tools.ietf.org/rfc/rfc3526.txt.

[12] P. Jokela, R. Moskowitz, and P. Nikander. "Using the Encapsulating Security Payload (ESP) Transport Format with the Host Identity Protocol (HIP)," RFC 5202, IETF, April 2008. Available from: http://tools.ietf.org/rfc/rfc5202.txt.

[13] T. Heer and S. Varjonen, "Host Identity Protocol Certificates," RFC 6253, IETF, May 2011. Available from: http://tools.ietf.org/rfc/rfc6253.txt.

[14] J. Laganier, T. Koponen, and L. Eggert, "Host Identity Protocol (HIP) Registration Extension," RFC 5203, IETF, April 2008. Available from: http://tools.ietf.org/rfc/rfc5203.txt.

[15] P. Nikander, T. Henderson, C. Vogt, and J. Arkko, "End-Host Mobility and Multihoming with the Host Identity Protocol," RFC 5206, IETF, April 2008. Available from: http://tools.ietf.org/rfc/rfc5206.txt.

[16] Z. Faigl, *Performance Analysis of Signalling Overhead in Host Identity Protocol-based Secure Mobile Networks: Ultra Flat Architecture or End-to-End Signalling?,* Wireless Networks, *under minor revision, 2014*.

[17] L. Bokor, Z. Faigl, S. Imre, *Survey and Evaluation of Advanced Mobility Management Schemes in the Host Identity Layer*, International Journal of Wireless Networks and Broadband Technologies (IJWNBT), vol. 3, no. 1, pp. 34–59, 2014.

Received xxxx 20xx; revised xxxx 20xx.

*E-mail address*: zfaigl@mik.bme.hu
*E-mail address*: telek@hit.bme.hu