Evaluation of two integrated signalling schemes for the Ultra Flat Architecture using SIP, IEEE 802.21, and HIP/PMIP protocols

Zoltán Faigl a,⇑, László Bokor a, Pedro Miguel Neves b, Khadija Daoud c, Philippe Herbelin c

a Mobile Innovation Centre, Budapest University of Technology and Economics, H-1111 Budapest, Bertalan Lajos u. 2., Z building 301, Hungary
b Portugal Telecom Inovação, S.A, Rua Eng. José Ferreira Pinto Basto, 3810 – 106 Aveiro, Portugal
c France Telecom Orange Labs, 38-40 rue du General Leclerc, 92130 Issy Les Moulineaux, France

Article info
Article history:
Available online 18 February 2011
Keywords:
Ultra flat architecture
Internet interworking
SIP
HIP
PMIP
IEEE 802.21

Abstract
Telecommunication suppliers predict a huge mobile Internet traffic increase for the next decade. It seems to be technically challenging and expensive to adapt current mobile network architectures to the increasing traffic demand. Core network technology must scale to the demands under limited revenue growth. This work is to discuss a new, flat, fully distributed and convergent network architecture, called Ultra Flat Architecture. It is well scalable due to the distribution of IP and numerous control functions at intelligent gateways placed within or close to the base stations. This paper focuses on the detailed presentation of two relevant integrated signalling schemes of the architecture. These schemes extend the Ultra Flat Architecture to support legacy Internet applications including IMS compliant applications with all necessary network functions, such as security, mobility, and Quality of Service. Besides Session Initialization Protocol, layers 2 and 3 are taken into account as well, in the terminal attachment and handover procedures. IP mobility is provided by the Host Identity Protocol or the Proxy Mobile IP protocol in the two schemes. We analyse the suitability of the schemes using the Multiplicative Analytic Hierarchy Process. Both schemes have nearly the same performance, however the HIP-based scheme got slightly better scores under our evaluation criteria due to its stronger security and fewer functional modules to deploy.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Telecommunication suppliers predict huge and rapid mobile Internet traffic demand growth for the next decade, due to the increasing capacity of radio network technologies, the expected realization of Internet of things and ambient networks. There are several solutions to tackle capacity shortage problems in the radio access networks (RAN). Increasing spectrum resource is the most straightforward technique. Spectrum efficiency can be improved due to latest releases in the 3rd Generation Partnership Project (3GPP), i.e., High Speed Packet Access and Long Term Evolution (LTE). Cell densification and traffic offload through femtocells and non-3GPP access networks extend the capacity as well. The increasing traffic demands and the novel RAN traffic engineering techniques expose the core network to growing pressure. Hence, telecommunication suppliers are concerned with the scalability of the core networks. They need sustainable networks producing enough return on investment. Scalability problems of 3GPP core networks are mainly due to the existence of anchor-points in the network. Anchors are centralized functions in the data forwarding and control plane. Such anchors are the first IP gateways (e.g., Packet Data Network GW) enabling communication with the Internet, centralized mobility services (e.g., Mobile IP Home Agents, Mobility...
Management Entity), and converged services, such as the Proxy Call Session Control Function (P-CSCF) and the Serving CSCF (S-CSCF) in the IP Multimedia Subsystem (IMS). A discussion of the disadvantages of anchor-based and advantages of end-to-end mobility protocols regarding the architecture can be found in [1]. The scalability problems of IP-based services and the IMS service layer in 3GPP architectures are detailed in [2].

To enhance scalability of the core network, the Ultra Flat architecture (UFA) has been introduced by Daoud et al. [1]. UFA represents the ultimate step towards flattening IP-based core networks, e.g., the Evolved Packet Core (EPC) in 3GPP. The objective of the UFA design is to distribute core functions into single nodes at the edge of the network, e.g., the base stations. Certain control functions could remain in the core, e.g., to support 3GPP and IMS roaming or to centralize the subscriber information base. The intelligent nodes at the edge of the network are called UFA gateways. Daoud et al. focus on the handover procedure problems of UFA [1]. If the first IP router is located in the access network, mobility introduces frequent IP-level handovers, especially in dense areas. The authors have developed a Session Initialization Protocol (SIP) based handover procedure for UFA. It has been proven by analysis [1] and in a testbed [2] that seamless handovers can be guaranteed for SIP-based applications. SIP Back-to-Back User Agents (B2BUAs) in UFA GWs can prepare for fast handovers by communicating the necessary contexts, e.g., the new IP address before physical handover. This scheme supports both mobile node-(MN) and network-decided handovers.

The session establishment procedure and the integration of IMS and UFA have already been investigated [3]. In 3GPP, IMS facilitates service and network convergence by separating the service level from the access layer. This introduces a two-level session establishment procedure. First, the MN and the correspondent node (CN) negotiate the session parameters on the service level, then Policy and Charging Control ensures that the bearer established in the access layer uses the resources corresponding to the negotiated session. The problem is that the service level is not directly notified about access layer resource problems, and e.g., it is difficult to adapt different components of the same service to the available resources in the access layer. In UFA, access layer resource information is present in the close neighbourhood of UFA GWs. UFA GWs (B2BUAs) can influence the negotiated parameters during the SIP session establishment and update. Consequently, in addition to enhancing scalability, purely SIP-based UFA is entirely controlled by the operator, and integrates Quality of Service (QoS) in its establishment and mobility procedures.

Interworking with Internet applications is a major requirement for mobile operators. In converged networks that use SIP control, an important problem is that many applications preferred by users apply other protocols for session establishment, e.g., Hypertext Transfer Protocol (HTTP). We refer to those applications as non-SIP applications.

In this paper, we present two possible signalling schemes in detail, which extend UFA, i.e., non-SIP applications are supported at the same level as SIP application. The idea behind them is the realization of IP mobility management in the layers lower than the application-layer. The first scheme is based on the Host Identity Protocol (HIP) [4], the second scheme uses Proxy Mobile IPv6 (PMIP) [5] for IP-mobility control. After describing the main procedures of the schemes, they are evaluated under performance, security and deployment criteria using the Multiplicative Analytic Hierarchy Process (MAHP) [6].

We have previously elaborated and analysed a HIP, Mobile IP (MIP) [7] and SIP-based scheme for UFA in [8,9]. In these schemes handover preparation is performed by the IEEE 802.21 Media Independent Handover (MIH) [10] protocol. The conclusion was that the combination of the signalling schemes, i.e., (1) SIP and HIP or (2) SIP and MIP, can ameliorate UFA to support non-SIP applications [8]. These mixed signalling schemes have been developed and are presented in this paper. Both schemes have been extended with proactive context transfer using Context Transfer Protocol (CXTP) [11] and MIH for more efficient handover preparation. MIP has been changed to PMIP for network-based mobility control. This work deals with problems of UFA that were previously out of our focus. It specifies the terminal attachment, access authorization and handover procedure involving layer 2 (L2) and layer 3 (L3).

Both HIP- and PMIP-based mobility controls have been adapted to the Ultra flat architecture. In [12], a new HIP extension has been presented enabling signalling delegation. Moreover, a new cross-layer access authorization mechanism for L2 and HIP was introduced, to replace certificate-based access authorization in an operator-based environment. In the current paper, the application of these ideas is presented for the HIP-based scheme. Furthermore, we generalize cross-layer authorization, and apply it to the PMIP-based scheme as well.

We must note that the PMIP philosophy was not modified in this work. PMIP is an anchor-based mobility management protocol, but on the other hand it provides mobility independently from the application type. It is currently used in EPC. In this work we have integrated the PMIP protocol within the Ultra Flat Architecture, and assumed that multiple LMAs can exist when deploying the UFA.

The remainder of the paper is organized as follows. Section 2 provides an overview of related works on flattening network architectures. Section 3 presents the reference scenarios of HIP and PMIP-based UFA signalling schemes. Section 4 specifies the main procedures of the two signalling schemes. Section 5 describes the main steps of the suitability analysis of the schemes according to our requirements. Section 6 compares the signalling schemes based on our evaluation criteria. Finally, Section 7 concludes the paper and presents our future plans.

2. Related work

When triple play services took off, the fixed network architecture was modified. The first IP routing function has been pushed close to the subscribers, e.g., in the Digital Subscriber Line Access Multiplexer (DSLAM) for DSL
services [13]. This has solved scalability problems because core network nodes provide simple routing or switching, and the number of intelligent edge nodes is mostly influenced by the number of users physically linked to the network. This change reduced the investment costs for the fixed network.

The flat architecture concept for mobile cellular networks is not new. Bosch et al. [14] explained that a shift from hierarchical to flat architecture is needed using one special network entity to provide radio access network functionality, and standard IP-based network elements in the core.

The concept of a completely flat and distributed architecture has been proposed by Yan et al. in [15]. The authors describe Wireless IP/Internet Service Environment (WISE) that enhances the Evolved Packet System by integrating the RAN and core network functions in one node. Mobility management and impact of WISe on IMS are not detailed in the referred paper. Their paper concludes that mobility management and distributed computing are challenging without centralized anchors.

Bertin et al. [16] published another article which describes an IP mobility procedure, named Dynamic Mobility Anchor, for a flat architecture where IP anchors are distributed in the base stations. Authentication, security and IMS operation are not developed in those mobility procedures.

As the aforementioned work shows, flattening existing telecommunication architectures seems to be impossible without pushing IP routing close to the users. However, this will easily result in significant growth in the number of IP address changes during operation. Conventional IP architecture uses IP addresses as both node identifiers (IDs) and locators. This leads to limitations on supporting mobility, multihoming, fast endpoint renumbering, traffic engineering, and scalable routing. Therefore, ID/Locator separation is essential in future flat architectures. Several existing approaches can distinguish identities needed by applications from locators which are used for addressing. These solutions include the Locator/ID Separation Protocol (LISP) [17], Site Multihoming by IPv6 Intermediation (SHIM6) [18], Forwarding directive, Association, and Rendezvous Architecture (FARA) [19], and HIP. The Host Identity Indirection Infrastructure (Hi3) [20] is based on HIP, and introduces a new logical protocol layer as a distributed and well scalable overlay for translating locators to identifiers. In a wide sense SIP also offers ID/Locator separation, because it uses IP addresses to locate nodes and SIP URIs to identify users or applications. However, if we define ID/ Locator separation as a technique which allows the network layer to change locators without interfering with upper layer procedures, then SIP is not an ID/ Locator separation technique. The ID/Locator separation concept has been introduced in the standardization of future network architectures by ITU-T [21,22].

3. Architectural overview of the proposed signalling schemes

The HIP- and PMIP-based signalling schemes are designed with the following requirements in mind. UFA must support simultaneous execution of SIP and non-SIP applications on multi-interface devices. UFA must provide seamless L2 and L3 handovers over heterogeneous access networks. Establishment and handover procedures must take available resources in the access layer into account. Network-decided handovers must be provided by UFA GWs, but MNs should also be able to influence handover decisions. An important aspect is to support the convergent service layer functions. Furthermore, mutual authentication and data protection is required on L2 and L3 between MNs and UFA GWs.

We aimed to reuse existing standard technologies. Both signalling schemes use the MIH [10] protocol and proactive context transfer (CXT [11]) in the handover procedures. MIH provides access network independent functions to monitor currently available L2/L1 resources, prepare L2 resources, and commit L2 handovers. Both schemes apply the IPsec protocol to protect the communication between the MN and UFA GW on IP level. EAP-AKA, authentication [23] and EAP re-authentication [24] services are also common in the schemes. The functional elements of these standards appear in UFA. Table 1 summarizes the important terms.

The reference scenarios for the HIP and PMIP-based signalling schemes are illustrated in Fig. 1(a) and (b), respectively. Both reference scenarios put intelligence into UFA GWs deployed at the edge of the network. The UFA GW is the first IP router seen from the MNs. An UFA GW may connect multiple MIH PoAs. Optionally, PoAs can be part of the UFA GW if the UFA GW’s interfaces are on the same link as the MNs, otherwise UFA GWs are non-PoAs. UFA GWs must instantiate MIH PoS functions to monitor local and remote L2/L1 resource availability in serving and candidate PoAs [10] and in connected MIH-enabled MNs. Consequently, MIH enables operator-driven handover decisions and efficient adaptation of services to available resources. Resource monitoring would also be possible over a multi-hop distance on IP level [27], hence it is not mandatory to deploy UFA GWs at the first IP hop. Another important part of UFA GWs is the SIP B2BUA, called UFA-Call Session Control Function (U-CSCF) [1]. It integrates several IMS functions, i.e., the Policy and Charging Rules Function (PCRF) and P-CSCF. As UFA GW gathers SIP session information and resource information from L1 to L4, efficient SIP session establishment, maintenance and QoS provisioning is achievable [2]. The UFA cross-layer (UFA-CL) module is in charge of the coordination of resource allocation, load balancing, and handover decisions in the UFA GWs. UFA-CL is a MIHU in the MIH taxonomy. Optionally, local EAP re-authentication (ER) server is also a part of UFA GWs, in order to provide fast re-authentication on L2 and L3 [12].

There are control functions which are not part of the UFA GWs and remain in the core network. The optimal location of these functions is subject to further research. Such functions are the IMS S-CSCF, the Home Subscriber Server (HSS), the authentication, authorization and accounting (AAA) servers with home ER server functionality, service and configuration provision (DHCP), media independent Information Service (MIIS), Interrogating CSCF (I-CSCF) to support IMS roaming. Existing service
platforms and application servers remain centralized as well. All core functions and the UFA GWs are connected with IP networks. A candidate transport technology to use is the IP over Multiprotocol Label Switching (IP/MPLS) over optical networks.

The two signalling schemes differ in addressing, mobility and security control functions. The HIP-based scheme illustrated in Fig. 1(a), applies HIP for IPsec security association (SA) establishment between the MNs and UFA GWs, and between the UFA GWs. IP-level handovers are prepared and executed using HIP delegation services [12] and CXTP-based context transfer. In this scheme the UFA GWs and MNs must be HIP enabled. Furthermore, the core network must provide HIP control plane functions for maintaining the MN’s reachability. This can be realized either by an RVS [25] function combined with HIP-capable DNS service, or a complete distributed HIP signalling architecture, such as Hi3 [20].

In the PMIP-based scheme, presented in Fig. 1(b), the Proxy Mobile IP protocol provides network-based IP mobility management support to the MNs. The functional entities of PMIP are the LMA and the Mobile Access Gateway (MAG). LMAs are located in the core network. LMA is responsible for maintaining the MN’s reachability state, and it is the topological anchor point for the MN’s home network prefix(es). There can be multiple LMAs in a PMIP domain each serving a different group of MNs. The MAG is the entity that performs the mobility management on behalf of a MN. It is located on the access link where the MN is anchored, therefore, it is deployed in the UFA GW. The PMIP scheme does not require the participation of the MN in any IP mobility related signalling. This signalling scheme applies the Internet Key Exchange Protocol version 2 (IKEv2) [28,29] for IPsec SA negotiation between the MN and the UFA GW. We suppose that during handover execution, IPsec and IKEv2 security contexts are transferred to the target UFA GW in the MIH commitment phase. Another possibility would be using CXTP, because currently IEEE 802.21 does not support the transfer of L3 contexts.

MOBIKE [30] or HIP multihoming and mobility extension [31] are not applicable in UFA to handle mobility because they only support location update. In UFA, IP-mobility leads to the change of the UFA GW’s identity and location besides the alteration of the MN’s IP address.

4. Main procedures of signalling

This section gives a macro-level description of the terminal attachment, session establishment and handover procedures of the HIP and PMIP-based UFA signalling schemes.

4.1. Terminal attachment

The terminal attachment procedure includes the attachment of the MN to the PoA on L2 and the UFA GW on L3. The functions covered are mutual authentication and key agreement on both layers, secure bootstrapping of the

Table 1
Taxonomy of applied standards.

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIHF</td>
<td>The function that realizes the MIH services</td>
</tr>
<tr>
<td>PoS</td>
<td>MIH Point-of-service: network-side MIHF instance</td>
</tr>
<tr>
<td>Non-POS</td>
<td>A MIH network entity that can directly exchange MIH messages with other MIH network entities but not with any MIH enabled MN</td>
</tr>
<tr>
<td>PoA</td>
<td>Network point-of-attachment: L2 access point</td>
</tr>
<tr>
<td>MIHU</td>
<td>MIH user: entities using the MIHF services</td>
</tr>
<tr>
<td>MIIS</td>
<td>Media independent information service: provides details on the characteristics and services supported by the serving and neighbouring access networks</td>
</tr>
<tr>
<td>RVS</td>
<td>Rendezvous server: HIP core network function for dynamic ID/Locator mapping [25]</td>
</tr>
<tr>
<td>LMA</td>
<td>Local mobility anchor: PMIP function for addressing and data forwarding [5]</td>
</tr>
<tr>
<td>MAG</td>
<td>Mobile access gateway: realizes PMIP mobility signalling for the MN [5]</td>
</tr>
<tr>
<td>ER</td>
<td>EAP re-authentication server: provides fast re-authentication service [24]</td>
</tr>
<tr>
<td>CSCF</td>
<td>Call session control function: call session control functions in IMS [26]</td>
</tr>
<tr>
<td>PCRF</td>
<td>Policy and charging rules function: function in IMS to enforce QoS and charging policies</td>
</tr>
<tr>
<td>BEX</td>
<td>HIP base exchange procedure for secure HIP host association establishment [4]</td>
</tr>
</tbody>
</table>

Fig. 1. Ultra flat architecture.
MN with configuration parameters, secure service discovery, and registration to addressing, MIH and IMS services.

4.1.1. L2 attachment

L2 attachment includes authentication and key agreement between the MN and PoA, and preparation for fast L2 and L3 re-authentications. The key concept is to apply ERP for fast re-authentications on L2 as well as L3. This has been presented shortly in [12] for HIP, but now it is also generalized and applied for the PMIP-based scheme.

ERP [24] is a standard solution to support intra-domain, inter-technology L2 handovers with fast L2 re-authentications if the access technologies support EAP [32]. In this work we assume that EAP authentication methods, such as the EAP-AKA or EAP-SIM for 3GPP WLAN Interworking (I-WLAN) [33] are supported on L2. We plan to extend this concept to other L2 authentication methods in the future.

Fig. 2 presents an EAP-AKA-based L2 authentication extended with ERP in UFA. The successful L2 EAP-AKA authentication leads to the derivation of Master Session Key (MSK) and EMSK. MSK is used for deriving encryption key (CK) and integrity key (IK) for data protection on L2. From EMSK a Domain-Specific Master Key (DMSK) [24] is calculated in the MN and the home ER server. DMSKs are transferred in a secure channel to the local ER servers. The EMSK and DSRK are used to create usage specific root keys (RK) [34]. We propose to create a L3 root key (L3RK) at the home ER server and domain-specific L3RK (DS-L3RK) at the local ER server, respectively, as well as in the MN. From the ER server, the L3RK is sent to the UFA GW, while from the L3RK a L3 integrity key (L3IK) is created for L3 re-authentications. Further master session keys (L3MSKs) are derived in the MN and the UFA GW at each L3 re-authentication. In the future this enables the lightweightening of L3 security control protocols, such as the replacement of Diffie–Hellman key exchange protocol in IKEv2 or HIP.

4.1.2. L3 attachment

Fig. 3 presents the L3 terminal attachment steps for both signalling schemes. Successful L2 attachment is followed by a minimal bootstrapping where the MN gets a temporary address and the UFA GW’s address. It is followed by the L3 authentication and key agreement (AKA) procedure described in Section 4.1.3. As a result, an IPsec SA pair is created between the MN and the UFA GW to protect further communication between them. It enables secure bootstrapping including MIH service discovery, configuration of the address of MAG, DNS, RVS or the initialization of other local naming service. The UFA GW may act as a DHCP server or relay when bootstrapping the MN.

After bootstrapping and service discovery, registration to various services is performed. The UFA GW (PoS) discovers the MIH capabilities of the MN, and registers to its MIH event services to monitor the MN’s link-state. In the HIP-based scheme, the MN registers to the UFA GW’s signalling delegation service in the L3 AKA procedure. As a result, the UFA GW gets a temporary authorization to delegate the MN towards the MN’s peers, i.e., it can perform signalling such as location update at RVS, IPsec and HIP association
maintenance, and L3 handover preparation and execution in the MN’s name. To provide reachability from CNs, the UFA GW registers the MN’s current location within the addressing service, i.e., updates the RVS or the LMA in the HIP and PMIP-based scheme, respectively. For SIP-based applications, the MN also performs a SIP registration procedure with the aid of the UFA GW (U-CSCF) as its SIP proxy.

4.1.3. Cross-layer access authorization

Fig. 4 illustrates the AKA part of the L3 attachment procedure. In the PMIP-based scheme, IKEv2 with ERP using the L3IK authentication key mutually authenticates the MN and the UFA GW using the home or local ER server. A new L3MSK must be calculated at each re-authentication in the MN and in the ER server. The ER server must send these keys to the UFA GW in a secure way. The session keys created during the L3 AKA procedure must be bound to L3MSK to assure their authenticity.

In the HIP-based scheme, the MN and the UFA GW mutually authenticate during HIP BEX [4]. The problem is that the MN and the UFA GW must check each other’s role as well. In [12], the following alternative solution to certificate-based authentication has been proposed. The Host Identifier (HI) or the HIT of the MN and the UFA GW are sent to the home/local ER server in an ERP message exchange, and the ER server checks their authorization states.

4.2. Session establishment

The session establishment procedure in UFA can be divided into two main subroutines: architectural support of session establishment (i.e., HIP/PMIP procedures) and application layer support of session establishment (i.e., procedures for SIP/non-SIP applications).

4.2.1. HIP/PMIP procedures

HIP and PMIP-based UFA signalling schemes are different in their session establishment procedures. In the PMIP case no special mechanisms are needed before session initiation on the application level. On the other hand, HIP nodes need to perform a BEX before starting the communication. Then, SIP or non-SIP application sessions can be initiated, as described in Sections 4.2.2 and 4.2.3, respectively.

The HIP session establishment between MN and CN nodes, illustrated in Fig. 5, is achieved by using the HIP delegation services defined in [12]. Based on this approach UFA GWs are able to execute BEX and create host association entries on behalf of MNs. Existing IPsec SAs between MN’s and CN’s UFA GW are re-usable in case of multiple HIP connections. For appropriate traffic mapping to IPsec SAs the MNs and CNs consult their HIP host association databases, and UFA GWs maintain HIT-based traffic forwarding tables. If the CN is also behind an UFA GW, then the HIP and IPsec association is established with the UFA GW of the CN. CN is notified using HIP Notification to create a HIP host association state with the MN. In order to avoid maintenance of unnecessary states in HIP hosts, and to stick to the basic HIP considerations providing DoS resistance, it is important that mandated session establishment requests must be sent only in the I2 message of the BEX sequence. I1 and R1 messages are both simple HIP BEX messages containing only the standard HIP parameters.

4.2.2. SIP-based applications

The UFA CL module is able to check locally (within the UFA GW) whether the negotiated session characteristics and the available access network resources are matching. Therefore, it is the UFA CL who chooses the right codec for the session because it knows both the network and the terminal capabilities. This choice is made thanks to QoS preconditions used in the SDP as proposed in [3]. The UFA GW could also control the user profile to apply a specific policy control in accordance with the user rights. The user profile is stored in the UFA GW during the terminal attachment procedure.
4.2.3. Non-SIP-based applications

The existing common Internet applications are not based on SIP to establish the session. Legacy applications, such as HTTP based browsing, FTP, P2P file sharing, instant messaging etc., are widely used, thus UFA must support them. In our schemes after the establishment of a HIP/PMIP session between the MN and the CN, the MN is able to send the initialization packet of the remote non-SIP service or application running on the CN.

4.3. Mobility procedures

4.3.1. Handover initiation

During the handover initiation phase (illustrated in Fig. 6), the UFA mobility management algorithm decides to initiate the handover process to one of the candidate UFA GWs. Within this phase, the source UFA GW configures the serving access interface of the multimode terminal with the set of QoS parameters required for the serving access link, using MIH procedures. As a result, the serving access interface periodically notifies the registered MIH user (i.e., the UFA CL module in the source UFA GW) about its QoS parameters. Based on this information, the UFA CL has sufficient information about the serving access network and, if necessary, can trigger the handover preparation phase before connectivity is lost.

4.3.2. Handover preparation and decision

During this phase all available information is collected in order to help the handover decision algorithm in the source UFA GW to select the target UFA GW. This phase is constituted by the following sub-phases. (1) Discovery: during this phase, the list of candidate UFA GWs is obtained through the MIIS, which collects information about the candidate access networks, such as their identifiers, L2 addresses, accounting information, etc. UFA GWs may also maintain a local MIIS database. (2) Query: in this phase, the mobility decision algorithm acquires all QoS metrics for all available candidate UFA GWs; (3) Selection: the mobility decision algorithm running either on the network or in the terminal side, decides for the target network. Fig. 7 illustrates the handover preparation phase for a network-initiated handover. The source UFA GW CL module queries the MIIS about the available neighbouring networks, then asks the MN to narrow the list of candidate access networks, and finally checks the available resources at each candidate UFA GW. Thereafter, it decides the selected target UFA GW for the handover procedure.

4.3.3. HIP-based handover preparation and execution

After the selection of the target PoA and the target UFA GW, the necessary HIP and IPsec contexts are proactively established in the network by the source UFA GW using HIP delegation services. HIP delegation services have already been described in [12]. As a reiteration of the main idea, in the case of Type 1 Delegation the delegator asks the delegate to establish HIP and IPsec states for the delegator and the specified peers of the delegator (CNs), and transfer the established states to the delegator. The delegator can hence establish HIP and IPsec associations with the

**Fig. 6.** 802.21 handover initiation.

**Fig. 7.** 802.21 handover preparation.
CN without running the HIP BEX procedure. In the case of Type 2 Delegation service, the delegator asks the delegate to establish HIP and IPsec states in the delegator’s name at specified peer nodes and to further maintain the HIP and IPsec states.

The prerequisites of the handover procedure are that the target UFA GW must register to the Type 1 Delegation service of the source UFA GW, in order to delegate HIP and IPsec association establishment. Furthermore, the source UFA GW (or the MN) must subscribe to the Type 2 Delegation service of the target UFA GW, to authorize the target UFA GW to update the MN’s location at the MN’s active peers, i.e., its CNs or the UFA GWs of its CNs and the RVS.

As depicted in Fig. 8, the source UFA GW initiates a Type 2 Mandated Action Request on behalf of the MN for handing off MN’s sessions. It triggers a Bulk Type 1 Delegation Action Request sent back to the source UFA GW. With this request the target UFA GW authorizes the source UFA GW for the establishment of HIP and IPsec connections with the MN’s peers in the name of the target UFA GW. Then the source UFA GW sends the security contexts to the target UFA GW using CXTP protocol protected with IPsec. Hence, the number of HIP BEX procedures can be reduced and replaced by HIP Updates.

After the successful context transfer, the target UFA GW updates the traffic forwarding policies for the MN at the CNs, RVS, and the MN, as illustrated in Fig. 9. Firstly, Type 2 Mandated Requests are sent by the target UFA GW to the CNs and the RVS in the MN’s name. After updating the MN’s peers, the target UFA GW informs the source UFA GW with a Type 2 Mandated Action Response to prepare for the redirection of the sessions. The target UFA GW updates its HIT-based traffic forwarding table to receive traffic from MN’s peers and send packets towards the source UFA GW. The source UFA GW also updates the MN’s and its own local HIT-based traffic mapping table: the traffic coming from the MN, related to the sessions that will be handed off soon, must be mapped to the IPsec tunnel that has the target UFA GW on the other side. The MN delays the activation of forwarding its traffic to the target UFA GW. Therefore, the traffic of the MN passes through the source and target UFA GW until the physical handover completes. After the MN’s successful L2 attachment to target L2 PoA, the source UFA GW will be simply skipped from the communication path, as described in Section 4.3.5.

Fig. 10 illustrates the execution phase for HIP-based handover. After HIP-level handover preparation, L2 handover execution procedure is initiated by the MIH_N2N_HO_Commit and MIH_Net_HO_Commit request messages, towards the target UFA GW and the MN, respectively. Then, the MN attaches on L2 to the target L2 PoA. This procedure contains fast L2 re-authentication using ERP.

Two options were defined to handle the mobility of SIP applications, as illustrated in Fig. 10. The first option, referred to as Option 1 in the followings, means that SIP and non-SIP applications are treated in the same way for the handover. In this case HIP is responsible for L3 mobility. The available resources may change in the target access network, hence SIP applications may need to update the
SDP information. The SDP update is initiated by the MN after the physical inter-UFA GW handover, in the handover completion phase (see Section 4.3.5). A drawback of this signalling option is that SDP updates may add additional service interruption delay when the MN’s sessions must be downgraded in the new access network, and parallel traffic forwarding in not applicable. Therefore, there is a second option (Option 2) to hand off SIP sessions. The eventual SDP update at the MN and the MN’s CNs is proactively performed, i.e., it is initiated by the source UFA GW when the MN is still attached to the source UFA GW. The benefit of this option is that the SDP update message exchange increases the handover preparation delay instead of the service interruption time.

4.3.4. PMIP-based handover execution

During the PMIP handover execution phase (see Fig. 11) the radio resources are activated in the target access network and the handover is executed. The serving UFA GW informs the target UFA GW (where the target MAG is located) about the handover commitment and requests the latter to prepare resources for the incoming MN. The target UFA GW, i.e., the UFA CL, queries the incoming MN’s profile from an AAA server and sends a Proxy Binding Update in order to register the location of the MN in advance. Thereafter the handover is triggered and the PMIP MAG is informed about the handover. Similarly to the HIP-based signalling scheme, two options were defined to handle the mobility of SIP applications. In Option 1, SIP and non-SIP applications are treated in the same way for the handover, i.e., PMIP is responsible for L3 mobility in a seamless way for applications. In this option, applications must adapt their QoS in a reactive manner, after handover execution, in the handover completion phase. In Option 2, the eventual SIP SDP update is proactively performed, similarly to the HIP-based handover presented in Fig. 10.

5. Evaluation of the signalling schemes

Our aim is to compare the two signalling schemes under three different application scenarios, i.e., (1) establishment and handing off non-SIP sessions, (2) establishment and handing off SIP sessions in the same way as non-SIP sessions (called SIP, Option 1), and (3) establishment and handing off SIP applications with reduced handover execution delay (named SIP, Option 2). The evaluation is based on the following assumptions. We apply the network model presented in Section 5.1. A set of key performance indicators has been defined in Section 5.2. The evaluation also required the specification of criterion weights which influence the overall opinion on the schemes, and the definition of normalization scales which assign grades to performance metrics.

The ranking method used for the evaluation is the Multiplicative Analytic Hierarchy Process (MAHP). It was customized to our needs and described in detail in [6]. Multi-attribute decision making consists typically of the following steps. In the first step the input performance values are normalized with a normalization technique, such as the sum, vectorization, max–min and max methods [35]. In MAHP this step is the assignment of performance values to the overall opinion on the schemes, and the definition of normalization scales which assign grades to performance metrics.

The ranking method used for the evaluation is the Multiplicative Analytic Hierarchy Process (MAHP). It was customized to our needs and described in detail in [6]. Multi-attribute decision making consists typically of the following steps. In the first step the input performance values are normalized with a normalization technique, such as the sum, vectorization, max–min and max methods [35]. In MAHP this step is the assignment of performance grades according to performance values. Then, in the second step, an aggregation method, such as weighted sum, weighted product and MAHP, generates a final score for each alternative, according to the criterion weights and the normalized input parameters. Criterion weight assignment is a subjective part of ranking methods. However, once the weights are fixed, the judgement of alternatives becomes objective under those assumptions.

In most combinations of these two steps the terminal scores become relative, i.e., the terminal scores of the alternatives depend on each other. We must mention that it has never been proven about any ranking method which gives the right answer. In most cases no reference exists to
validate whether a method results in really the best alternative. On the other hand, inconsistencies of a given ranking method, such as the rank reversal problem, can be objectively shown. MAHP is a method which does not have this inconsistency. Furthermore, performance values are categorized during grade assignment, hence MAHP does not differentiate alternatives which perform close to each other.

The MAHP method asks for relatively more input parameters. Unfortunately, every parameter increases the possibilities of errors, but there are advantages as well. A method which has more information about a decision can be expected to bring a better decision. The decisions to be made about the input parameters of MAHP are not so complex questions because the problem is separated into many subproblems. E.g., the performance grade assignment to handover delay values can be defined in advance by asking experts or reading standard requirements. Most of the ranking methods do not ask for so many parameters about the performance grade assignment functions. Consequently, these methods apply the same normalization and scale types, without considering the individual judgement of the performance metrics.

5.1. Network model

The delay and message overhead calculation of the alternatives under the performance criteria is supported by a network model presented in Fig. 12. It is composed of the MN, the UFA GWs in charge of MN or CN attachment and the UFA core nodes. We assume in this model that the UFA equipments are linked to a full mesh IP network. We analyse two different network scenarios, i.e., (S1) the CNs are servers, e.g., HTTP servers, and (S2) the CNs are other MNs attached to UFA GWs. The parameters of the network model are summarized in Table 2. The parameters give the one-way average delay metrics of different network parts. They do not depend on the message size or other parameters. The one-way delay values reflect worst case delays and are based on real measurement experiences of the authors. For some parameters both single and a range of values are given. It depends on the evaluation which values to use. Since our network model contains overestimated delays, the fulfilment of delay criteria means that performance of the given scheme is suitable.

5.2. Validation criteria

We defined three main criteria, i.e., provide (1) low performance costs, (2) high security, and (3) low deployment costs. There are fifteen sub-criteria under them, organized in a criteria tree, illustrated in Fig. 13. Complex solutions have many key performance indicators. We tried to find objective measures. In those aspects our schemes were objectively comparable.

Each message overhead criterion contains five child criteria that are not illustrated in the criteria tree due to space limitations. They take into consideration the message overhead of the procedures on different parts of the network, i.e., parts (I,VII), II, III, VI, V. The following weights have been assigned to these criteria, respectively: {0.61,0.09,0.17,0.10,0.02}. This means that we are mostly concerned with the signalling load of the wireless interface and the inter-UFA GW links.

Some of the main requirements of the UFA, such as high scalability, self-configuring and self-optimizing network were not included in the criteria tree due to the following reasons. High scalability is mainly ensured by the structure of the architecture and influenced by the user traffic. Hence, the signalling schemes do not differ considerably under the scalability criterion. The evaluation of the other aforementioned criteria requires the elaboration of the related mechanisms, which are future research topics of the UFA.

The criteria weights have been defined in an iterative process, based on the consensus of five decision makers. The obtained criteria weights are presented in the first column of Fig. 13. As can be seen from the weights, the dominating sub-criteria under the three main criteria are the low real-time service interruption delay due to inter-UFA GW handovers; the mutual authentication, signalling and user data protection between the MN and the UFA GW; and the low number of additional modules to deploy in the MN in the control plane and user plane, respectively. The three rightmost columns of Fig. 13 describe our input parameters for performance grade assignment. The meaning of these parameters is described in detail in [6]. We apply two types of normalization, i.e., Type-3 and Type-4 scales. Both are step-functions and have seven grades. The former is a smaller-the-better while the latter is a higher-the-better type grade assignment function, respectively. \( g_{\min_i} \) and \( g_{\max_i} \) give the starting and ending of the grade assignment functions, while the progression factor \( \gamma_i \) influences their shape. \( P_{\min_i} \) and \( P_{\max_i} \) are hard constraints for criterion \( i \). If an alternative performs outside

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Delay [ms]</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_{\text{MN}} )</td>
<td>([1, \ldots, 100])</td>
<td>Access network delay bw. MN and Source or Target UFA GW</td>
</tr>
<tr>
<td>( d_{\text{CN}} )</td>
<td>([5, \ldots, 100])</td>
<td>Access network delay bw. CN and UFA GW</td>
</tr>
<tr>
<td>( d_{\text{L}} )</td>
<td>([5, \ldots, 50])</td>
<td>IP network delay bw. MN’s UFA GW 30 and the UFA core nodes</td>
</tr>
<tr>
<td>( d_{\text{core}} )</td>
<td>10</td>
<td>Delay bw. the UFA core nodes</td>
</tr>
<tr>
<td>( d_{\text{U}} )</td>
<td>20</td>
<td>Delay bw. source and target UFA GW</td>
</tr>
<tr>
<td>( d_{\text{core,CN}} )</td>
<td>( S_1: d_{\text{core}} )</td>
<td>Delay bw. UFA core nodes and the CN’s UFA GW</td>
</tr>
<tr>
<td>( d_{\text{U,CN}} )</td>
<td>( S_2: d_{\text{L}} )</td>
<td>Access network delay bw. CN’s UFA GW</td>
</tr>
<tr>
<td>( d_{\text{L2,conf}} )</td>
<td>50</td>
<td>L2 reconfiguration delay of the MN during access to a new UFA GW</td>
</tr>
</tbody>
</table>

Fig. 12. Network model.
the interval given by the hard constraints then it must be rejected by assigning zero grade to it.

5.3. Features of the signalling schemes

In this part we summarize the properties of the alternatives regarding the three main criteria. Easy deployment is one of the main criteria. We introduced the following simplified measure for deployment complexity. Assuming a default TCP/IP stack in each UFA node, which supports MIP and IPsec. We counted the surplus protocol/functional modules in user, control planes in the MNs, UFA GWs, and the UFA core network. Table 3 summarizes the modules needed in the UFA nodes for the two combined signalling schemes. Our hypothesis behind this metric is that the deployment and operational costs increase when adding complexity to a network. This measure disregards complexity of the functional modules and that maintenance costs of each function may be unique.

The security features of the alternatives are summarized in Table 4. In the evaluation, 1 means that a security feature is supported by the signalling scheme. The 0 values represent no support or very weak provision of the security service. The first three criteria are supported in both schemes. HIP supports strong DoS resistance using a built-in puzzle-mechanism, tunable to various levels of DoS threats. On the other hand, IKEv2 provides weak and optional cookie-based protection, hence we assign 0 performance value for its DoS resistance. Both HIP and IKEv2 resist man-in-the-middle attacks. The sigma-compliant Diffie–Hellman key exchange in HIP assures resistance to MITM attacks. In case of IKEv2, the key agreement process is cryptographically bound to the mutual authentication process, hence it also resists to MITM attacks.

**Table 3**

<table>
<thead>
<tr>
<th>Objective: rank UFA signalling schemes</th>
<th>HIP-based UFA</th>
<th>PMIP-based UFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN's control plane</td>
<td>HIP, SIP, 802.21, UFA-CL, EAP, ERP</td>
<td>IKEv2, PMIP, SIP, 802.21, UFA-CL, EAP, ERP</td>
</tr>
<tr>
<td>MN's user plane</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UFA GW's control plane</td>
<td>HIP, 802.21, U-SCSF, UFA-CL, EAP, ERP</td>
<td>IKEv2, MAG, 802.21, U-SCSF, UFA-CL, EAP, ERP</td>
</tr>
<tr>
<td>UFA GW's user plane</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UFA core</td>
<td>HIP-capable DNS, RVS, MIIS, EAP, ERP</td>
<td>LMA, MIIS, EAP, ERP</td>
</tr>
</tbody>
</table>

**Table 4**

<table>
<thead>
<tr>
<th>Security features of the alternatives</th>
<th>HIP-based UFA</th>
<th>PMIP-based UFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutual authentication</td>
<td>1: HIP, ERP</td>
<td>1: IKEv2, ERP</td>
</tr>
<tr>
<td>Signalling protection</td>
<td>1: HIP, IPsec</td>
<td>1: IKEv2, IPsec</td>
</tr>
<tr>
<td>User data protection</td>
<td>1: IPsec</td>
<td>1: IPsec</td>
</tr>
<tr>
<td>DoS resistance</td>
<td>1: strong</td>
<td>0: weak</td>
</tr>
<tr>
<td>MITM resistance</td>
<td>1: supported</td>
<td>1: supported</td>
</tr>
</tbody>
</table>
The overall terminal scores of the signalling schemes are illustrated by the bottommost group of bars named ‘Objective’. These bars show the main results of the evaluation: (1) The HIP is slightly better than the PMIP-based alternative due to the fact that it is better under the deployment and the security criteria. (2) In the second application scenario (SIP, Option 1) both HIP and PMIP get zero. They induce higher than 250 ms real-time service interruption delay. We checked that PMIP (2.2) remains rejected if \( d_{MN} \) and \( d_{CN} \), i.e., the one-way delay of the access network, are set to 0 ms. This happens because the accumulated delay at the other network parts is greater than 250 ms for PMIP (2.2). This result led us to develop a second option (Option 2) for the handover of real-time SIP applications.

HIP is better than PMIP in application scenarios 1 and 3 because of its higher DoS resistance. It can be parametrized based on the existing threats in the network to protect the UFA GWs from malicious MNs flooding the network. Furthermore, from the aspect of deployment, the HIP-based alternative requires one less module in the control plane of the MN and the UFA GW. HIP module is responsible both for the mobility and the security control, while in the PMIP-based scenario this is performed by two modules, i.e., the IKEv2 and the PMIP or MAG.

Handover preparation delay becomes an important measure in case of frequent L3 handovers in dense areas. Even if there are significant differences between the alternatives regarding the number of handover preparation messages, they fall in the same, best category regarding handover preparation delay. The explanation is that we defined a performance grade scale, which assigns the highest grade to alternatives that perform below one second. In our network scenarios all alternatives are below this constraint.

In this evaluation the assumptions regarding the procedures reflect a worst case situation. We assumed that during L2 attachment and re-authentications the full EAP-AKA authentication procedure is executed, i.e., the advantages of ERP are not exploited. During the handover initiation phase the source UFA GW always queries the MIES entity in the core network. SDP update is required at each SIP application handover. Furthermore, SIP sessions remain interrupted until the SDP update is finished. Moreover, the HIP-based scheme, we have always counted with the establishment of new HIP and IPsec associations, while in practice, existing HIP and IPsec associations are reused until their lifetimes expire. On the other hand it was assumed that during MIH handover preparation the source UFA GW requests only one candidate PoA about its available resources, and the number of active sessions (to be handed off) is one for a given MN.

Fig. 15 shows the terminal scores for the alternatives, provided smoother assumptions are given about the probabilities of different events. Only the results for the main
criteria are illustrated. We suppose that 90% of L2 re-authentications execute the ERP. 50% of MIIS requests are made locally in the cache of the UFA GWs. SDP update is needed in 50% of the handoffs of SIP sessions. For the HIP-based signaling scheme, a target UFA GW has an established HIP and IPsec association with the UFA GW of the MN’s CN in 10% of the cases. During 802.21 handover preparation the source UFA GW asks three candidate UFA GWs on average for their available resources. The number of active sessions (to be handed off) is three on average for a given MN.

The results indicate that in the case of smoother assumptions the HIP (2.1) scenario is also accepted. Even if the input parameters were chosen arbitrarily, this analysis aimed to show that the HIP-based scheme has a specific scaling property. By increasing the number of application-level sessions, not every session establishment triggers the generation of new HIP and IPsec associations between UFA GWs. These associations will tend to be established all the time, and can be reused. Compared to end-to-end security solutions like the Secure Mobile Architecture [36], this is a clear advantage of the HIP-based UFA.

We supposed that in 50% of the cases, SIP sessions are handed off without the need for SDP update. In this case, the average service interruption time is below 250 ms, hence the evaluation process accepts the alternative. This result must be treated carefully because when an SDP update is needed the HIP (2.1) will not fulfill the service interruption delay requirement. PMIP (2.2) is still rejected because its average service interruption delay is still higher than 250 ms.

6.2. Sensitivity of the terminal scores to the one-way delay between the MNs and the UFA GWs

In this analysis we aim to analyse the sensitivity of ranking scores of the signalling schemes to the access network delay between the MN and the MN’s UFA GW ($d_{MN} = d_{CN}$) in network scenario S2. We calculate the terminal scores of the alternatives for a range of access network delays ($d_{MN} = d_{CN} = 0, \ldots, 100$ ms.) Fig. 16 presents the results of this analysis. Fig. 16(a)–(c) illustrate the variation of the terminal scores for the low service interruption delay category assignment function for the low service interruption delay criterion gives the grade $v = 7$ to alternatives performing between 0 and 190 ms. Moreover, the performance grade is decreasing to 0 within the interval (190,250) ms, given by a step-function of Type-3. This means that from 0 to

![Fig. 14. Terminal scores of the alternatives (CNs are also MNs).](image)

![Fig. 15. Terminal scores of the alternatives under smoother pre assumptions.](image)
approximately 200 ms of service interruption times all alternatives are considered equally good.

The figures show that the alternatives have a constant terminal score from 0 until a given \( d_{SN} \) value. This is the range where the service interruption delay is below 200 ms. Then, with increasing \( d_{SN} \) the alternatives get worse performance categories. The PMIP application scenarios (2.2 and 3.2) perform worse than the HIP application scenarios (1.1 and 3.1). HIP (1.1) can not fulfil the hard constraint even at an access network delay \( d_{SN} < 5 \) ms. PMIP (2.2) fails even at \( d_{SN} = d_{CN} = 0 \) ms, because the service interruption delay induced by the link delays at the other parts of the network exceed the hard constraint. The PMIP alternatives (2.2 and 3.2) seem to fulfill the 200 ms criterion until \( d_{SN} = d_{CN} = 30 \) ms, and reach 250 ms service interruption delay between \( d_{SN} = d_{CN} = 50, \ldots, 55 \) ms. The HIP alternatives fulfill the delay criterion until \( d_{SN} = d_{CN} = 50, \ldots, 55 \) ms, and are rejected between \( d_{SN} = d_{CN} = 65, \ldots, 75 \) ms. In MAHP the terminal scores are relative, i.e., they depend on the performance of alternatives. This explains that in Fig. 16 some alternatives get increasing scores even if in absolute terms they perform worse. These trends indicate that the relative preferences for certain alternatives grow.

7. Conclusions

In this paper, we have proposed two signalling schemes which extend UFA to support non-SIP applications. We have described terminal attachment, session establishment and handover procedures. Both L2 and L3 have been taken into account during the design. We evaluated the alternatives under security, performance and deployment criteria. The results show that both schemes fulfill these requirements. With our assumptions the HIP-based alternative got slightly better ranking scores than the PMIP-based scheme. This is due to the more sophisticated Denial of Service resistance built into HIP, protecting UFA GWs, and the highly integrated mobility and security control functionalities in the HIP-based signalling scheme. HIP and PMIP scenarios may introduce too large service interruption delay for real-time SIP applications. Therefore, real-time SIP applications may need to update their SDP information proactively, before physical handover. This update should also replace the MIH commitment phase, and immediately trigger the physical handover. We plan to develop a comprehensive simulation environment to evaluate the schemes under realistic traffic demands and topologies, and analyse their scalability.

Acknowledgements

The Ultra Flat Architecture has been studied in the Eurescom funded P1857 project, under the collaboration of France Telecom, Portugal Telecom, and Mobile Innovation Centre Hungary. The authors are grateful to the study supervisor Ádám Kapovits for the administrative management of the project.

References

Philippe Herbelin received his degree in Networking & Telecommunications Engineering from the “Institut National des Télécommunications” (INT) in 1998. He joined France Telecom Network Division to work on IP network architecture evolution with the development of triple play offers. Since 2006, he moved to France Telecom Orange Labs, as a senior architect on IP/MPLS network and Fixed Mobile Convergent Networks. His current interest research is focused on optimized convergent networks based on full IP technology to anticipate mobile Internet challenges. He is involved in the design in of a new flat and fully distributed mobile and convergent architecture.