Enhanced fast handoff scheme for heterogeneous wireless networks

Christian Makaya *, Samuel Pierre

Abstract

Mobility management, integration and interworking of existing wireless systems are important factors to obtain seamless roaming and services continuity in next generation or 4G wireless networks (NGWN/4G). Although, several IPv6-based mobility protocols as well as interworking architectures have been proposed in the literature, they cannot guarantee seamless roaming, especially for real-time applications. Moreover, mobility management protocols are designed for specific needs, for example, the purpose of IPv6-based mobility schemes consists of managing users roaming while ignoring access network discovery. This paper proposes an efficient handoff protocol, called enhanced Handoff Protocol for Integrated Networks (eHPIN), which carries out localized mobility management, fast handoff, and access network discovery. It alleviates services disruption during roaming in heterogeneous IP-based wireless environments. Performance evaluation results show that eHPIN provides significant gain with respect to signaling traffic overhead cost, handoff latency, packet delivery cost, handoff failure and packet loss compared with existing IPv6-based mobility schemes.

C211 2008 Elsevier B.V. All rights reserved.

Keywords: Mobility management; IP mobility; Quality of service; Heterogeneous wireless networks; Handoff; Seamless roaming; Services continuity

1. Introduction

Fourth or next generation wireless networks (4G/NGWN) are expected to exhibit heterogeneity in terms of wireless access technologies, user-oriented services and greater capacities. Users will have increasing demands for seamless roaming across different wireless networks, support of various services (e.g., voice, video, data) and quality of service (QoS) guarantees. Hence, with this heterogeneity, users will be able to choose radio access technology (RAT) that offers higher quality, data speed and mobility which is best suited to the required multimedia applications. Moreover, technological advances in the evolution of portable devices make it possible to support different RATs. Heterogeneity in terms of RATs and network protocols in 4G/NGWN requires common interconnection element. Since the Internet Protocol (IP) technology enables the support of applications in a cost-effective and scalable way, it is expected to become the core backbone of 4G/NGWN [1]. Thus, current trends in communication networks evolution are directed towards an all-IP principles in order to hide heterogeneities of lower-layers technologies from higher-layers and to achieve convergence of different networks.

Mobility management, with provision of seamless handoff and QoS guarantees, is one of the key topics in order to support global roaming of mobile nodes (MNs) in NGWN. Providing seamless mobility and service continuity (i.e., minimal service disruption during roaming) support based on intelligent and efficient techniques is crucial. This means that seamless handoff schemes should have following features: minimum handoff latency, low packet loss, low signaling overhead and limited handoff failure or blocking rate. Handoff latency represents the time interval during which an MN cannot send or receive any data traffic during handoffs. It is composed of L2 (link switching) and L3 (IP layer) handoff latencies. The overall handoff latency may be sufficiently long to cause
packet loss, which is intolerable for real-time applications such as voice over IP (VoIP). Furthermore, subscribers are more sensitive to session/call blocking during handoff than to session blocking during call initiation. The handoff blocking probability refers to the likelihood that a session connection is prematurely terminated due to an unsuccessful handoff over a session lifetime. Hence, minimization of handoff blocking probability is crucial for mobility management schemes. The signaling traffic overhead is defined as the total number of control packets (for registration, binding update and binding refresh procedures) exchanged between an MN and a mobility agent (e.g., home agent).

Several IPv6-based mobility schemes such as Mobile IPv6 (MIPv6) [4], Hierarchical Mobile IPv6 (HMIPv6) [7] and Fast Handovers for Mobile IPv6 (FMIPv6) [8], have been proposed by the Internet Engineering Task Force (IETF) to enable an MN to remain reachable when moving out of its home network. However, these protocols are hindered by several drawbacks such as signaling overhead, handoff latency and packet loss. To achieve seamless mobility across various access technologies and networks, an MN needs to have information regarding the wireless network to which it can attach. To enable this, Candidate Access Router Discovery (CARD) protocol [10] was proposed by the IETF. When coupled with CARD protocol, traditional fast handoff schemes may work inefficiently since some operations may be redundant, which results in higher handoff delay and wastage of network resources.

Enhancing those protocols for efficient mobility management in heterogeneous NGWN is highly necessary. This paper proposes a mobility management scheme, called enhanced Handoff Protocol for Integrated Networks (eHPIN), that enables seamless roaming, services continuity and QoS guarantees for real-time applications in heterogeneous IPv6-based wireless environments. eHPIN performs access network discovery, localized mobility and fast handoff management. In other words, eHPIN aims to provide efficient access network discovery and roaming support in order to alleviate services disruption during handoff. It is designed for both heterogeneous and homogeneous wireless networks. The main contributions of this paper can be summarize as follows:

(i) The proposal of an appropriate usage of fast handoff, media independent handover (MIH) and access networks discovery concepts in one suite protocol that provides efficient mobility management in heterogeneous wireless environments.

(ii) To further reduce handoff delay and packet loss, we propose the anticipated binding update procedure and tunnels establishment as well as efficient context transfer.

(iii) Performance evaluation is done based on the proposed analytical model in order to compare eHPIN with some existing IP-based mobility management protocols.

The remainder of this paper is organized as follows. In Section 2, an overview of basic concepts and related work are depicted. An interworking architecture for 4G/NGWN is presented in Section 3. The proposed mobility management protocol, eHPIN, is described in Section 4. Performance analysis and numerical results are shown in Sections 5 and 6, respectively. Finally, Section 7 concludes the paper.

2. Background and related work

Mobility management enables a communication system to locate roaming terminals in order to deliver data packets (i.e., location management) and to maintain connections with them as they move into a new subnet (i.e., handoff management). Handoff management is a major component of mobility management since an MN can trigger several handoffs over a session lifetime as it will be the case in NGWN. It is crucial to provide seamless mobility and service continuity support based on intelligent and efficient techniques. Various schemes have been proposed in the literature and by the IETF for mobility management in IP-based wireless networks.

Mobile IPv6 (MIPv6) [4] was proposed for mobility management at the IP layer and allows MNs to remain reachable in spite of their movements within IP wireless environments. Each MN is always identified by its home address, regardless of its current point of attachment to the network. While away from its home network, an MN is also associated with a care-of address (CoA), which provides information about the MN's current location. After acquisition of CoA, an MN sends a binding update (BU) message to the home agent (HA), informing it of the new address and also to all active correspondent nodes (CNs) to enable route optimization. However, MIPv6 has some well-known drawbacks such as signaling traffic overhead, high packet loss rate and handoff latency, thereby causing user-perceptible deterioration of real-time traffic [5,6]. These weaknesses led to the investigation of other solutions designed to enhance MIPv6 and support micro-mobility of MNs.

Two main MIPv6 extensions proposed by the IETF are Hierarchical MIPv6 (HMIPv6) [7] and Fast Handovers for MIPv6 (FMIPv6) [8]. HMIPv6 handles handoffs through a special node called Mobility Anchor Point (MAP). The MAP, acting as a local HA in the network visited by the MN, limits the amount of MIPv6 signaling outside its domain and reduces delays associated to location update procedure. However, HMIPv6 cannot meet the requirements for delay sensitive traffic, such as voice over IP (VoIP), due to packets loss and handoff latency. FMIPv6 was proposed to reduce handoff latency and to minimize services disruption due to MIPv6 operations during handoffs such as movement detection, binding update and addresses configuration. The link layer information (L2 trigger) is used either to predict or respond rapidly to handoff events.
Although FMIPv6 paves the way on improving MIPv6 performance in terms of handoff latency, it remains hindered by several problems such as QoS support and scalability. In fact, FMIPv6 does not effectively reduce global signaling and packet loss, which cause unacceptable services disruption. In FMIPv6, a new access router (NAR) consumes storage space to buffer the forwarded packets by previous access router (PAR) before delivering packets to the MN. Moreover, these transferred packets lack QoS guarantee before the new QoS path is setup. Combining HMIPv6 and FMIPv6 motivates the design of Fast Handover for HMIPv6 (F-HMIPv6) [9] to allow more network bandwidth usage efficiency. However, F-HMIPv6 may inherit drawbacks from both FMIPv6 and HMIPv6, such as synchronization issues and signaling traffic overhead that result in combining both schemes [5,6].

To achieve seamless mobility across various access technologies and networks, an MN needs information about the wireless network to which it could attach. Also, it is necessary to transfer information (context transfer) related to the MN from the current access router to the next one. The Candidate Access Router Discovery (CARD) protocol [10] and the Context Transfer Protocol (CXTP) [11] have been proposed to enable this procedure. They prevent the usage of limited wireless resources, provide fast mobility and secure transfers. Their key objectives consist of reducing latency and packet loss, and avoiding the re-initiation of signaling to and from an MN from the beginning. However, context transfer is not always possible, for example, when an MN moves across different administrative domains. The new network may require the MN to re-authenticate and perform signaling from the beginning rather than to accept the transferred context. Moreover, the entities which exchange context or router identities must authenticate each other. This could become a tedious process in NGWN. All of the aforementioned remarks show that seamless mobility and services continuity are not guaranteed in the current IPv6-based mobility management protocols.

The IEEE802.21 or Media Independent Handover (MIH) [12] standard has been proposed in order to provide interoperability, generic link layer intelligence and to enable efficient handoff management between heterogeneous access networks, that includes both 802 and non-802 based networks (e.g., 3GPP and 3GPP2). A new logical network entity called the MIH Function (MIHF) is defined between link layer (L2) and network layer (L3) which provides three main services: Media Independent Command Service (MICS), Media Independent Information Service (MIIS) and Media Independent Event Service (MIES). The MICS provides a set of commands that enables MIH users to issue commands for handoff control and mobility. In fact, the MICS commands are used to determine the status of the connected links and also to execute connectivity and mobility decisions of the higher-layers to the lower ones. While the MIIS provides event classification, event reporting and event filtering corresponding to dynamic changes in link characteristics, status and quality. It reports both local and remote events to the upper layers.

Some of the events that have been specified by MIH are Link Down, Link Up, Link Going Down, Link Parameters Report, and Link Detect. Finally, the MIIS provides the capability to discover available neighboring network and for obtaining the necessary information to make effective handoff decisions. Note that, the MIIS defines a set of information elements (IEs) to provide access to dynamic and/or static information and higher-layer services supported by the networks. Such information include details on the services and characteristics provided by the serving and neighboring networks, roaming agreements, available networks, access costs, QoS parameters (e.g., delay, jitter, data rate), and security capabilities. The MN can access the relevant information via its current active network interface and the other interfaces do not need to be turned on simultaneously.

3. Interworking architecture for NGWN

Heterogeneity, in terms of radio access networks in NGWN, requires the integration and interworking of various existing wireless systems. Two majors architectures (loose and tight coupling) for 3G/WLAN interworking have been proposed by both 3G wireless network initiatives, 3GPP and 3GPP2, for their respective systems [2,13]. However, this integration brings new challenges such as architecture and protocols design, mobility management, QoS guarantees, interworking and security. All scenarios listed in [2,3] are not yet fulfilled. Moreover, both interworking models have as well as pros and cons.

An interworking architecture, called Integrated Inter-System Architecture (IIA) based on 3GPP/3GPP2-WLAN interworking models, was proposed in [14] and shown in Fig. 1. For the sake of simplicity, only UMTS, CDMA2000 and WLAN networks are illustrated. However, IIA may integrate any number of radio access technologies (RATs) and mobile devices may be equipped with any number of interfaces. Instead of developing new infrastructures, IIA extends existing infrastructures to tackle integration and interworking issues and provide mobile users with ubiquity or always best connected. The serving GPRS (general packet radio service) support node (SGSN) and packet control function (PCF) are enhanced with the AR functionalities and called Access Edge Node (AEN). Similarly, the gateway GPRS support node (GGSN) and packet data serving node (PDSN) are extended with the MAP or HA functionalities (to enable message format conversion, QoS requirements mapping, etc.) and called Border Edge Node (BEN). The WLAN Interworking Gateway (WIG) acts as a route policy element and ensures message format conversion.

A novel entity, Interworking Decision Engine (IDE), is introduced to enable the interworking and handoffs between various networks by reducing signaling traffic, services disruption during handoff and handles authentica-
tion, authorization and accounting (AAA) and mobility management. The usage of the IDE could be considered as a value-added service that network operators offer to their subscribers to allow roaming in other networks. To avoid additional signaling overhead due to the execution of the AAA procedure every time an MN performs handoff and requests registration, we propose a token-based approach. The token includes security association parameters to setup secure tunnel between an MN and AR/AENs.

The logical components of the IDE are illustrated in Fig. 2. The Authentication Module (AuM) is used to authenticate users moving across different wireless networks and avoids the required direct security agreements or association between foreign networks and home network. The AuM stores information such as the subscribers’ identities, users’ preferences/profiles and terminal mobility patterns. The Accounting Module (AcM) enables billing between different wireless networks and stores billing information associated with the resource usage. It acts as common billing/charging system between various network operators. The CIBER (Cellular Intercarrier Billing Exchange Roamer Record) protocol can be used for the exchange of roaming billing information, for voice and data, among wireless telecommunication companies through the IDE.

Usually, different administrative domains have different QoS policies for resources allocation. Thus, when an MN moves from one administrative domain to another, QoS re-negotiation may be required. Such re-negotiation will be based on service level agreements (SLAs) between both domains. The Resource Management Module (RmM) enables QoS mapping and re-negotiation. Furthermore, the RmM allows fast transfer of user profiles and QoS parameters between two administrative domains during handoff. The SLRA Module stores information about service providers or network operators which have SLAs and roaming agreements (RAs) with the IDE manager. The Handover Decision Module (HdM) is used when inter-system or inter-domain handoff should be granted or not. In other words, it provides support for roaming and handoffs. For further details about the IISA architecture, please refer to [14].

4. Proposed eHPIN protocol

Assuming that mobile devices are becoming increasingly powerful, intelligent and sensitive to link layer changes, a network-assisted and mobile-controlled handoff strategy is adopted. eHPIN combines both mobile-monitored and network-probed information to provide reliable handoff control. Prior to handoff, an MN can obtain information regarding candidate wireless networks to which it is likely to handoff and uses such information to optimize handoff performance. On the other hand, if mobile device capabilities are limited, handoff decision is taken by mobility agents in the network side. L2 trigger generation may be imprecise because it is a link layer event and depends on L2 technology and channel conditions. Thus, two modes (predictive and reactive) have been proposed for FMIPv6. When coupled with CARD protocol, FMIPv6 can be inefficient. In fact, some operations of those both protocols may be redundant, which results in higher handoff delay, signaling overhead and wastage of network resources. Hence, it is necessary to find an effective way to perform access router discovery procedure and handoff anticipation in a one-suite protocol. eHPIN is proposed to reach this goal.

In eHPIN, the handshake procedure for access router discovery and fast handoff as well as all time consuming operations such as bi-directional tunnels setup between the MAP/BEN and candidate access routers (CARs) or AENs, duplicate address detection (DAD) procedure and CAR/AEN pre-selection are performed before the L2 trig-
ger generation. According to MIH standard, an MN which implements our proposed handoff decision function [16] is able to obtain periodically the information about available networks in its moving coverage area by using its current network interface. In other words, the information provided by the MIIS is used to estimate the parameters of the QoS capabilities and requirements.

4.1. Handoff initiation with eHPIN

With the information exchanged between the MAP/BEN and AR/AENs by using Router Information eXchange (RIX Request/Reply) messages, the BEN maintains a global view (i.e., load status of AENs, connection state of any MN in its domain as well as their movement patterns) of its domain and can learn both L2 and L3 information of an access network. The L2 (link layer) information may include the specific link layer wireless access technology, system parameters (e.g., channel frequency and number) and QoS status such as bandwidth availability and signal strength. The L3 (IP layer) information can include the global address of AR/AEN, the address of the prefix advertised in the wireless network, the current QoS status and parameters. The QoS parameters may include information such as the supported data rate, the video coding rate, and maximal delay.

The L2 and L3 information are then forwarded to the IDE and allows it to maintain a global view of all MAP/BEN domains having SLAs with the IDE manager. The exchange of RIX messages is quite similar as that of the routing information protocol (RIP) [15] works to allow neighboring routers to exchange their routing table with one another. The update interval time (e.g., 30 s) for each information depends on its property: static or dynamic. Thus, the backbone signaling increment does not require higher additional costs for system deployment.

The MN decides whether to send the CARD Request message to MAP/BEN according to the generation of the anticipated triggers (ATs). For example, high bit error rate, link going down and weak signal strength, security risks, monetary cost and geographical location can be used as anticipated triggers. Anticipated triggers are defined based on IEEE 802.21/MIH and are slightly different of basic or general L2 trigger used in FMIPv6. In other words, AT just announces or anticipates the need of possible handoff while general L2 trigger generation results into link layer switching. For example, when an MN moves, we can consider Link Going Down as an anticipated trigger while classical L2 trigger is Link Up.

The CARD Request message contains user preferences as well as information regarding the applications required QoS capabilities. To allow seamless service continuity, the requirements specified in the CARD Request message need to be set consistently with the QoS negotiated in the previous subnet. Crucial for real-time applications, QoS consistency is handled by the IDE, which allows QoS mapping between different networks. Upon receipt of the CARD Request message, the MAP/BEN checks its local CAR/AEN table to retrieve information about CAR/AENs’ capabilities. Moreover, the MAP/BEN performs pre-filtering of available AR/AENs in order to have potential CAR/AENs, address auto-configuration (AA) process on behalf of an MN in order to form one or more new on-link CoAs (NLCoAs). For address auto-configuration, we assume that the new CoAs pool is located at the MAP/BEN and which is updated by an out-of-band signaling based on RIX message exchanged between the MAP/BEN and AR/AENs. The MAP/BEN relieves the MN of the burden of LCoAs and RCoAs computation or configuration.

Note that if the MAP/BEN lacks information regarding this user profile, it requests his information to the IDE rather than to the MN’s HA, which is likely to be far away from the current location. After receiving the CARD Request message, the MAP/BEN sends the handoff initiate (HI) message containing the corresponding NLCoAs to the potential CAR/AENs. When all potential CAR/AENs receive the HI message containing NLCoA, they perform a duplicate address detection (DAD) procedure and acts as a proxy for the MN to defend this temporary address in its network. The HI is also used to trigger the request of context transfer. In other words, the MAP/BEN transmits a Context Transfer Data (CTD) message, piggybacked in HI, to CAR/AENs. Example of features contained in CTD message are QoS context information, header compression, security details, and AAA information. This paper focuses mainly on QoS context information. Performing a DAD procedure for all possible NLCoAs with eHPIN requires some extra overhead compared to basic F-HMIPv6 and FMIPv6. However, the DAD procedure is performed prior to the L2 trigger generation, then it reduces L3 handoff latency and the impacts of imprecise L2 trigger timing.

When the CAR/AEN receives a CTD message, it may generate a CTD Reply (CTDR) message optionally to report the status of processing the received contexts and this message is piggybacked in the handoff acknowledgment (HACK) message. The CAR/AEN installs the contexts once it is received from the MAP/BEN. This context will be activate upon receiving a fast binding update acknowledgment (FBAck) message. The CAR/AEN will send a HACK message to the MAP/BEN only after relocation of traffic bearers and resources are reserved for the new path in order to indicate that handoff may be done and packets forwarding may be initiated. The MAP/BEN binds previous on-link CoAs (PLCoAs) and the NLCoA, but marks its state idle and sends a CARD Reply message to the MN which contains the NLCoAs set, CAR/AENs list and capabilities. The idle state means that, the MAP/BEN does not start buffering and forwarding packets at this stage, nor does it uses reserved resources for this handoff preparation request. Contrary to FMIPv6 and F-HMIPv6, where forwarded packets lack QoS guarantees before the new QoS path is set up, eHPIN solves this issue.
With the CARD Request/Reply messages exchange, an MN knows the CARs to which it is likely to handoff. Then, the MN will activate only the interface associated to the CAR/AEN list, rather than setting all air-interfaces always on as is the case with traditional IPv6-based mobility management schemes. This selective interface activation enables better trade-off between system discovery time and power consumption efficiency. After receiving the CARD Reply, an MN can start a handoff any time. The CARD Request/Reply messages exchange no longer delays the handoff procedure, as it is carried out while the MN uses the previous on-link CoA (PLCoA). Whenever an MN receives the L2 trigger, it initiates a target AR/AEN selection among the CAR/AENs set. This selection is based on the handoff decision function proposed in [16].

4.2. Handoff execution with eHPIN

Once the handoff decision step is completed the MN sends a fast binding update (FBU) message containing the selected target AR/AEN information to the MAP/BEN. Unlike the basic fast handoff schemes, the FBU message is not used to trigger bi-directional tunnel establishment or handoff initiate/acknowledgment (HI/HAck) messages exchanges, but rather triggers the packet forwarding procedure. Upon receipt of the FBU message, the MAP/BEN activates the idle binding, sends the fast binding update acknowledgment (FBAck) message to the MN on both links (previous and new) and establishes a binding between PLCoA and NLCoA. The MAP/BEN can start packets forwarding to the target NAR/AEN.

4.2.1. Intra-BEN roaming scenario

When the selected NAR/AEN among CAR/AENs receives the FBAck message, it activates the transferred context. The MN performs a L2 handoff and sends fast neighbor advertisement (FNA) message to announce its presence on the new link. Upon receiving the FNA, the NAR/AEN starts to deliver the buffered packets, if any, to the MN. The disordered packets problem can be reduced significantly with buffering at the NAR/AEN or MN. In fact, a routing header extension is added to the forwarding packets addressed to PLCoA towards NLCoA. This procedure refers to the reactive mode of eHPIN while the predictive mode is explained above (i.e., the MN sends FBU message through PAR/AEN’s link and FBAck is received before the L2 handoff). The reactive mode is carried out either intentionally or serve as a fall-back mechanism when the predictive mode cannot be completed successfully, for example, if the L2 handoff is completed before the FBAck message is received at the MN. Signaling messages exchange of eHPIN is shown in Fig. 3 during intrasystem or intersystem handoff for intra-MAP/BEN roaming. Contrary to basic fast handoff schemes (i.e., FMIPv6, F-HMIPv6), only one round trip message exchange for FBU/FBAck and FNA messages are required for handoff after L2 trigger with eHPIN scheme.

4.2.2. Inter-BEN roaming scenario

If the CAR/AENs are located within another MAP/BEN domain, the serving MAP/BEN (MAP1/BEN in Fig. 4) sends a handoff request (HOReq) message to the candidate MAP/BEN (MAP2/BEN in Fig. 4) and encapsu...
lates a HI message within HOReq. The candidate MAP/BEN forwards HI message virtually in parallel to all CAR/AENs belonging to its domain by including the CTD message. Note that, if the context information of this MN are not available at the candidate MAP/BEN, the later sends Context Transfer Request (CTReq) message to the IDE in order to obtain the session management parameters of the MN for establishment of traffic bearers on the new path. In response to the CTReq message, the IDE transmits a CTD message that includes the MN’s feature contexts. When the new MAP/BEN receives a CTD message, it installs the contexts as received them from the IDE.

Once the application requirements are validated, the CAR/AENs send a HAck message to the candidate MAP/BEN, which then encapsulates a HAck message with the handoff reply (HORep) message and sends it to the current MAP/BEN. HORep[HAcc] contains NLCoAs, CAR/AENs capabilities and other adequate information. Upon receipt of HORep[HAcc] message, the serving MAP/BEN sends a CARD Reply message including the CAR/AENs list and capabilities, associated NLCoAs and other information. Similar operations as those in the case of intra-BEN roaming ensue when a L2 trigger is generated. Fig. 4 shows the message sequence exchange for eHPIN for inter-BEN roaming.

5. Performance evaluation

In IP-based wireless networks, QoS may be defined in terms of packets loss, handoff latency and signaling traffic overhead cost. Analysis of these metrics are very useful to evaluate the performance of mobility management protocols. The notation used in this paper is outlined in Table 1.

Table 1

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_g$</td>
<td>Global binding update cost to HA/CNs</td>
</tr>
<tr>
<td>$C_l$</td>
<td>Local binding update cost to MAP/BEN</td>
</tr>
<tr>
<td>$M$</td>
<td>Number of subnets in the MAP/BEN domain</td>
</tr>
<tr>
<td>$N_{CN}$</td>
<td>Number of CNs with a binding cache entry for an MN</td>
</tr>
<tr>
<td>$d_{XY}$</td>
<td>Average number of hops between nodes $X$ and $Y$</td>
</tr>
<tr>
<td>$C_{XY}$</td>
<td>Transmission cost of control packets between nodes $X$ and $Y$</td>
</tr>
<tr>
<td>$PC_X$</td>
<td>Processing cost of control packet at node $X$</td>
</tr>
<tr>
<td>$t_f$</td>
<td>Time period between the L2 trigger and the start of the link switch</td>
</tr>
<tr>
<td>$t_p$</td>
<td>Time period between the transmission of FBU and the start of the link switch</td>
</tr>
<tr>
<td>$P_s$</td>
<td>Success probability of the anticipated handoff</td>
</tr>
</tbody>
</table>
occurs after the L2 trigger, will be larger. We assume that the success probability $P_s$ to be dependent on the timing of the L2 trigger. In ideal case, when the time taken from the occurrence of the L2 trigger event to the start of the real L2 switching handoff is 0, the success probability $P_s$ will be 1. On the other hand, if this time approaches infinity, $P_s$ will be 0. This means that handoff anticipation is done too early without any exact information as to the link condition.

User mobility and traffic models are crucial for efficient system design and performance evaluation. Usually, MN mobility is modeled by the cell residence time and numerous random variable types are used for this purpose [17]. We consider a traffic model composed of two levels, session and packets. The session duration follows exponential distribution with the inter-session rate $\lambda_s$ while the packet generation and arrival rate follow a Poisson process. The evaluation of time that an MN stays within the subnet is usually based on two distributions: Gamma and exponential. The Gamma distribution is very realistic for mobility model by considering changes in the speed and direction of the MN while the exponential distribution is a particular case of Gamma distribution.

The exponential distribution provides an acceptable trade-off between complexity and accuracy. Thus, most cost analyses adopt exponential assumption [17]. We consider that subnet and MAP/BEN domain residence time follow exponential distribution with a mean of $1/\mu_s$ and $1/\mu_d$, respectively. Note that, $\mu_s$ is the border crossing rate for an MN moving out of an AR/AEN coverage area and $\mu_d$ for an MN moving out of the MAP/BEN domain. When an MN crosses the MAP/BEN domain border, it also crosses an AR/AEN border. Hence, the rate for AR/AEN crossing for which the MN remains in the MAP/BEN domain is $\mu_t = \mu_s - \mu_d$. If we assume that all subnets are made up of circular shapes forming together a contiguous area and that each MAP/BEN domain is composed of $M$ equally subnets, then $\mu_d = \frac{\mu_s}{\sqrt{M}}$ [18].

5.1. Total signaling cost

The performance analysis of wireless networks must consider the total signaling cost induced by mobility management schemes. As for wireless cellular networks, signaling traffic overhead cost must be performed for NGWN or IP-based mobile environments. In NGWN, there are two kinds of location or binding update signaling. One takes place from an MN’s subnet crossing and another occurs when the binding is about to expire. Moreover, delivery of data packets induces usage of network resources, thus generating additional costs. Hence, the total signaling cost, $C_T$, could be divided into the binding update signaling cost, $C_{BU}$, and the packet delivery cost, $C_{PD}$: $C_T = C_{BU} + C_{PD}$.

5.1.1. Binding update signaling cost

The binding update cost heavily depends on the average number of location updates during the inter-session arrival time. Depending on the type of movement, two kinds of location or binding updates could be performed: local and global binding update. The global binding update procedure refers to the registration of RCoA to HA/CNs. On the other hand, if an MN changes its current address (LCoA) within a MAP/BEN domain, it only needs to register this new LCoA to the MAP/BEN. Hence, the average location binding update cost for IPv6-based mobility management schemes during inter-session time can be expressed by:

$$C_{BU} = E(N_i)C^l + E(N_d)C^g$$  \hspace{1cm} (1)$$

where $E(N_i)$ is the average number of subnets (AR/AENs) that an MN crosses while remaining within a given MAP/BEN domain during an ongoing session and $E(N_d)$ denotes the average number of MAP/BEN domains crossing.

To perform a signaling overhead analysis, a performance factor called session-to-mobility ratio (SMR) is introduced. It is similar to the call-to-mobility ratio (CMR) defined in cellular networks [19]. The SMR represents the relative ratio of session arrival rate over the user mobility rate: $SMR = \lambda_s/\mu_s$. The binding update signaling cost, $C_{BU}$, is then given by:

$$C_{BU} = \frac{1}{\mu_s} (\mu_s C^g + \mu_t C^l) = \frac{1}{SMR \sqrt{M}} [(C^g + (\sqrt{M} - 1)C^l)]$$

(2)

In IP-based mobile environments, not all L2 handoffs result in L3 handoffs. Hence, handoff procedure anticipated by using L2 trigger may lead to unnecessary signaling traffic. The critical phase of the fast handoff approach starts when a L2 trigger is generated to indicate the impending handoff. We assume that if an MN receives a FBAck message from the MAP/BEN, that it will inevitably start L3 handoff to the NAR/AEN without exceptions. Hence, if there is no real handoff after a L2 trigger generation, all messages exchanged from FBU to FBAck may be unnecessary. The global and local binding update signaling cost for eHPIN are expressed as follows:

$$C^g = P_s S^g_f + (1 - P_s)(S^f_f + S^o_f) + C_{iu}$$
$$C^l = P_s S^g_f + (1 - P_s)(S^f_f + S^o_f) + C_{mu}$$  \hspace{1cm} (3)$$

where $C_{iu}$ represents the binding update cost at the HA/CNs, $C_{mu}$ denotes the binding update cost at the MAP/BEN, $S^g_f$ (resp. $S^o_f$) is the global (resp. local) signaling cost for a successfully anticipated handoff, and $S^f_f$ (resp. $S^o_f$) is the global (resp. local) signaling cost for control messages if no real L3 handoff occurs and $S^o_f$ (resp. $S^o_f$) denotes the global (resp. local) signaling cost for the reactive mode. Table 2 shows their expressions.

The packet transmission cost in IP-based networks is proportional to the distance in hops between the source and the destination nodes. Furthermore, the transmission cost in a wireless link is generally superior than that in a wired link [19]. Thus, the transmission cost of a control packet between nodes $X$ and $Y$ belonging to the wired part
of a network can be expressed as $C_{XY} = \tau d_{XY}$ while $C_{MN,AEN} = \tau k$, where $\tau$ is the unit transmission cost over wired link and $k$ the weighting factor for the wireless link.

### 5.1.2. Packet delivery cost

Similar to Koodli and Perkins [20], we divide handoff latency into three components: link switching or L2 handoff latency ($t_{L2}$), IP connectivity latency ($t_{IP}$) due to movement detection, and address configuration and location update delay ($t_{L}$). The IP connectivity latency reflects how quickly an MN can send IP packets after the L2 handoff, while the location update latency represents the delay required for forwarding IP packets to MN’s new IP address.

On the other hand, the time period between the starting point of L2 handoff and the moment an MN receives IP packets for the first time through new link refers to packet reception latency ($t_{IP}$) or data latency. Moreover, the following delay components are introduced: movement detection delay ($t_{MD}$), address configuration and DAD procedure delay ($t_{AC}$), binding update delay ($t_{BU}$) and delay from completion of binding update and reception of the first packet by an MN through the new IP address ($t_{NR}$).

The timing diagram of eHPIN for intra-MAP/BEN roaming is illustrated in Fig. 5. When two endpoints have an ongoing session, a packet delivery cost incurs. The packet delivery cost consists of the packet transmission cost and the packet processing cost. By using the handoff timing diagram illustrated in Fig. 5, the packet delivery cost could be defined as the linear combination of the packet tunneling/forwarding cost ($C_{tun}$) and the packet loss cost ($C_{loss}$). Let $\alpha$ and $\beta$ be the weighting factors (where $\alpha + \beta = 1$), which emphasize the tunneling and dropping effects. The packet delivery cost, $C_{PD}$, is computed as follows:

$$C_{PD} = \alpha C_{tun} + \beta C_{loss}. \quad (4)$$

Let $s_{C}$ and $s_{D}$ be the average size of control and data packets, respectively and $\eta = s_{D}/s_{C}$. The cost of transferring data packets is $\eta$ greater than the cost of transmitting control packets. Let $\lambda_{C}$ be the packet arrival rate in unit of packet per time. The packet loss in fast handoff schemes may be due either to L2 handoff or in case of wrong spatial prediction of NAR/AEN. The packet loss due to L2 handoff delay is inevitable without efficient buffering mechanisms [20]. Since a bi-directional tunnel is established before L2 trigger, there is no packet loss cost for the predictive mode of eHPIN (i.e., $C_{loss}^{p} = 0$). Moreover, as the packets forwarding process is not supported in the reactive mode, packet tunneling cost equal zero ($C_{tun}^{r} = 0$). Due to wrong spatial prediction of NAR/AEN or if FBAck message was not received through the previous link, the packets forwarded by the MAP/BEN to an erroneously predicted NAR/AEN can be lost. Packets forwarding to the wrong NAR/AEN stops when the FBU message sent through the NAR/AEN’s link is received at the MAP/BEN. In this case, the reactive mode of eHPIN is used.

The packet tunneling cost for predictive mode $C_{tun}^{p}$ and the packet loss cost $C_{loss}^{l}$ of eHPIN are expressed as follows:

$$C_{tun}^{p} = \lambda_{p} C_{tun}^{p} (t_{L2} + t_{IP}^{p} + t_{U}^{p})$$
$$C_{loss}^{l} = \lambda_{p} C_{loss}^{l} (t_{L2} + t_{IP}^{l} + t_{U}^{l})$$

where $t_{L2}^{p}$ is the location update latency for intra-MAP/BEN movement and $t_{U}^{l} = t_{BU}^{l} + t_{NR}^{l}$, $t_{IP}^{l}$ is the IP connectivity latency of reactive mode, $t_{IP}^{p}$ is the IP connectivity latency excluding IP addresses configuration, DAD procedure and movement detection. In fact, these operations are performed in anticipation prior an MN leaves the PAR/AEN’s link. The cost of transferring data packets from CN to MN through to the MAP/BEN is

$$C_{CM}^{l} = \eta (C_{CN,BEN} + C_{BEN,AEN} + C_{AEN,MN})$$

and $C_{CM}^{l} = \eta (C_{CN,BEN} + C_{BEN,AEN} + C_{AEN,MN})$. The average packet delivery cost of eHPIN scheme is given by:

$$C_{PD}^{l} = P_{s} C_{PD}^{p} + (1 - P_{s}) C_{PD}^{r}$$

where $C_{PD}^{p}$ and $C_{PD}^{r}$ indicate packet delivery costs for the predictive and reactive modes of eHPIN, respectively, and are computed using (4).

![Fig. 5. Handoff delay timeline of eHPIN for intra-BEN roaming.](image-url)
On the other hand, for inter-MAP/BEN roaming case with eHPIN, the timing diagram is illustrated in Fig. 6, where $t_{pl}$ is the delay to perform anticipated BU or to register a new RCoA to the HA, $t_{gr}$ is the delay for the return routability procedure and $t_{CN}$ represents the delay for performing anticipated BU or registering a new RCoA to all active CNs. The packet loss cost $C_{loss}$ and the packet tunneling cost $C_{tun}$ are expressed as follows:

$$
C_{loss}^{eHPIN} = \lambda_{p} C_{cm}^{eHPIN} (t_{L2} + \rho_{IP}^{eHPIN} + t_{U})
$$

$$
C_{tun}^{eHPIN} = \lambda_{p} C_{cm}^{eHPIN} \max(t_{L2} + \rho_{IP}^{eHPIN}, t_{RA} - t_{F}) + t_{RR} + t_{CN} + t_{NR}
$$

(7)

where $t_{U} = t_{BU} + t_{RR} + t_{NR}$, $\rho_{IP}^{eHPIN}$ is the IP connectivity latency of reactive mode for inter-MAP/BEN roaming, the cost of transferring data packets from CN to MN by transiting to the previous and new MAP/BEN is

$$
C_{cm}^{eHPIN} = \eta(C_{CN,pBEN} + C_{pBEN,aBEN} + C_{aBEN,aEN} + C_{AEN,MN})
$$

and $C_{cm}^{eHPIN} = \eta(C_{CN,pBEN} + C_{pBEN,DAR} + C_{AEN,MN})$. The average packet delivery cost for eHPIN associated to inter-MAP/BEN roaming is computed similarly as in (6) and by using (4), eHPIN eliminates all sources of packet loss except for the unavoidable loss due to the link layer switching handoff. However, with efficient buffering mechanism at AR/AENs packet loss during L2 handoff may be avoided.

5.2. Handoff latency and packet loss

Handoff latency and packet loss are computed according to the following parameters: $t_{L2}$ represents the L2 handoff latency and $t_{X,Y}$ one-way transmission delay of a message of size $s$ between nodes $X$ and $Y$. If one of the endpoints is an MN, $t_{X,Y}$ is computed as follows:

$$
I_{X,Y}(s) = \frac{1 - q}{1 + q} \left( \frac{s}{B_{w}} + L_{w} \right) + (d_{X,Y} - 1) \left( \frac{s}{B_{u}} + L_{u} + \sigma_{q} \right)
$$

(8)

where $q$ is the probability of wireless link failure, $\sigma_{q}$ the average queuing delay at each router on the Internet [21], $B_{w}$ (resp. $B_{u}$) is the bandwidth of the wireless (resp. wired) link and $L_{w}$ (resp. $L_{u}$) denotes the wireless (resp. wired) link delay.

Let $L_{u}$ be the time elapsed between the reception of FBAck on the previous link and the beginning of the L2 handoff when there is no good synchronization between L2 and L3 handoff operations. Moreover, let $\Delta_{u}$ be the time between the last packet received through the previous link and the L2 handoff beginning when the FBAck arrives on the new link. Note that, $\Delta_{u}$ and $\Delta_{l}$ can equal zero. For eHPIN, handoff latency depends on the information available, and on which link fast handoff messages are exchanged. If information regarding the NAR/AEN and impending handoff are available and, if the FBAck message is received through the previous link, the handoff latency is expressed as follows:

$$
O_{eHPIN}^{l} = \Delta_{l} + t_{L2} + 2t_{MN,AEN}.
$$

(9)

However, if a FBAck message is not received on the previous link, it will be received through the new link. Hence, in this case, the handoff latency for eHPIN is expressed as follows:

$$
N_{eHPIN}^{l} = \Delta_{l} + t_{L2} + 2t_{MN,AEN} + 3t_{AEN,BEN}.
$$

(10)

The average handoff latency of eHPIN for intra-BEN roaming is given by:

$$
D_{eHPIN}^{l} = P_{s} O_{eHPIN}^{l} + (1 - P_{s}) N_{eHPIN}^{l}.
$$

(11)

For inter-MAP/BEN movement, when FBAck message is received through the previous link, the handoff latency of eHPIN is identical as for intra-MAP/BEN roaming: $O_{eHPIN}^{l} = O_{eHPIN}^{l}$. In fact, the handoff procedure depends only on intra-MAP/BEN communication delay, since the inter-MAP/BEN signaling is completed before the L2 handoff. On the other hand, when the FBAck message is received through the new link for inter-MAP/BEN movement, it is assumed that appropriate information about NAR/AEN is already available and NLCoA is already configured. Hence, the handoff latency of eHPIN for inter-MAP/BEN roaming is given by:

$$
N_{eHPIN}^{l} = \Delta_{l} + t_{L2} + 2t_{MN,AEN} + 3t_{AEN,aBEN} + 2t_{AEN,pBEN}.
$$

(12)

The average handoff latency of eHPIN for inter-MAP/BEN roaming is computed similarly as in (11).

In theory with eHPIN, there are no packets loss, unless buffers overflow at NAR/AEN or MAP/BEN. However, without efficient buffer management (EBM), packets forwarded can be lost during handoff latency. In fact, the number of packets lost is proportional to handoff latency:
where $B$ is the buffer size of an AR/AEN and $BS_{eHPIN}^i$ is the required buffer space at NAR/AEN for intra-MAP/BEN roaming with eHPIN during packets forwarding and is computed as follows:

$$BS_{eHPIN}^i = \max(BS_{eHPIN}^i - B, 0)$$

(13)

Similarly we can compute the number of packets lost $P_{loss}^{eHPIN,g}$ and the required buffer space at NAR/AEN, $BS_{eHPIN}^g$, for inter-MAP/BEN roaming.

6. Numerical results

The parameter and default values used in performance evaluation are given in Table 3, except when the wireless link delay ($L_{w}$), packet arrival rate ($\lambda_p$) and prediction probability ($P_s$) are considered variable parameters. An analytical framework for the performance evaluation of IPv6-based handoff schemes proposed by the IETF, i.e., MIPv6, HMIPv6, FMIPv6 and F-HMIPv6 is presented in [22]. These evaluation methods are used to compare the performance of the IETF’s protocols with eHPIN. The network topology considered for this analysis is illustrated in Fig. 7. We assume that distance between different domains are equal, i.e., $c = d = e = f = 10$ and $a = 1$, $b = 2$, and $g = 4$. All links are considered to be full-duplex in terms of capacity and delay. Other parameters used to compute signaling costs are defined as follows: $\tau = 1$, $\kappa = 10$, $\alpha = 0.2$, $\beta = 0.8$, $PC_{AEN} = 8$, $PC_{HA} = 24$, $PC_{CN} = 4$, $PC_{IDE} = 15$ and $PC_{BEN} = 12$. Most parameters used in this analysis are set to typical values found in [19,23,24]. Performance analysis is conducted using MATLAB and OPNET software.

Fig. 8 illustrates the binding update signaling cost during handoff as a function of the SMR. When the SMR is small, the mobility rate is larger than the session arrival rate, the MN changes frequently its point of attachment resulting in several handoffs. These handoffs will cause the exchange of several messages between different entities and will increase signaling overhead. However, when the session arrival rate is larger than the mobility rate (i.e., SMR is larger than 1), the binding update is less often performed. In other words, the signaling overhead decreases as the frequency of the subnet change decreases. eHPIN allows significant signaling overhead cost saving compared to other protocols. The RIX messages exchange introduces additional signaling similarly as with routing information protocol (RIP). However, this signaling increment only occurs in the wired part of network. Compared to the wireless part, the wired one has much superior bandwidth and resources.

Fig. 9 illustrates the binding update signaling cost during handoff as a function of the prediction probability ($P_s$) when $SMR = 0.1$. HMIPv6 and MIPv6 are not affected by the prediction probability contrary to fast handoff-

### Table 3

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Symbols</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 handoff time</td>
<td>$t_{l2}$</td>
<td>50 ms</td>
</tr>
<tr>
<td>Time period between L2 trigger and L2 handoff</td>
<td>$t_f$</td>
<td>10 ms</td>
</tr>
<tr>
<td>Prediction probability</td>
<td>$P_s$</td>
<td>0.98</td>
</tr>
<tr>
<td>Wireless link failure probability</td>
<td>$q$</td>
<td>0.50</td>
</tr>
<tr>
<td>Wired link bandwidth</td>
<td>$B_w$</td>
<td>100 Mbps</td>
</tr>
<tr>
<td>WLAN link bandwidth</td>
<td>$B_{wl}$</td>
<td>11 Mbps</td>
</tr>
<tr>
<td>3G wireless link bandwidth</td>
<td>$B_{wl}$</td>
<td>384 Kbps</td>
</tr>
<tr>
<td>Wired link delay</td>
<td>$L_w$</td>
<td>2 ms</td>
</tr>
<tr>
<td>Wireless link delay</td>
<td>$L_{dl}$</td>
<td>10 ms</td>
</tr>
<tr>
<td>Number of ARs by domain</td>
<td>$M$</td>
<td>2</td>
</tr>
<tr>
<td>Control packet size</td>
<td>$s_c$</td>
<td>96 bytes</td>
</tr>
<tr>
<td>Data packet size</td>
<td>$s_d$</td>
<td>200 bytes</td>
</tr>
<tr>
<td>Packets arrival rate</td>
<td>$i_p$</td>
<td>10 packets/s</td>
</tr>
</tbody>
</table>
based schemes since they do not use L2 trigger to anticipate the handoff. The signaling overhead decreases when the prediction probability accuracy increases for fast handoff-based schemes (i.e., FMIPv6, F-HMIPv6 and eHPIN). For small values of $P_s$, HMIPv6 performs better than eHPIN. However, when $P_s$ increases, eHPIN outperforms all other schemes.

The packet delivery cost is shown in Fig. 10 as a function of the packet arrival rate ($\lambda_p$). Combined hierarchical and fast handoff-based schemes (i.e., F-HMIPv6 and eHPIN) perform better than FMIPv6, MIPv6 and HMIPv6. Moreover, they are more efficient when $\lambda_p$ increases. This means that eHPIN and F-HMIPv6 are more adequate for real-time applications where periodic packets are sent at high rates. We observe that eHPIN enables lower packet delivery cost compared to F-HMIPv6. For varying prediction probability ($P_s$), Fig. 11 shows the packet delivery cost which decreases when the accuracy of $P_s$ increases for fast handoff schemes. The high value of $P_s$ means that the FBAck message is received through the previous link (i.e., via PAR/AEN). Then, buffered packets are delivered to an MN just after it attaches to the new link. Hence, service disruption delay is reduced. We observe that, regardless of the prediction probability value, eHPIN outperforms all other schemes by providing a lower packet delivery cost. The prediction probability has a greater effect on F-HMIPv6. In fact, when $P_s = 0$, F-HMIPv6 turns to HMIPv6, which is its reactive mode.

Fig. 12 shows that the handoff latency increases proportionally to the wireless link delay. The handoff latency is very high for MIPv6 followed by HMIPv6 while FMIPv6 and F-HMIPv6 enable its reduction. eHPIN allows significant handoff latency reduction compared to other mobility management protocols. It is well known that the maximum tolerable delay for interactive conversation is approximately 200 ms. Hence, eHPIN can meet this requirement when the wireless link delay is set up below 50 ms.

Fig. 13 shows the total packet loss in terms of packet arrival rate. Note that packet loss is much less prominent for eHPIN than for other IPv6-based handoff protocols. The effect of handoff in IPv6-based wireless environments is dominated by packet loss, which is due to the L2 handoff and the IP layer operations. In fact, due to the lack of any buffering and anticipated handoff mechanisms in MIPv6 and HMIPv6, all in-flight packets are lost during handoff. However, in fast handoff schemes (i.e., FMIPv6, F-HMIPv6 and eHPIN) packet loss begins when L2 hand-
off is detected and until the buffering mechanism is initiated or if buffers overflow.

Fig. 14 provides comparison of forwarded packets delay induced by fast handoff schemes (i.e., FMIPv6, F-HMIPv6 and eHPIN). eHPIN allows the lowest delay for these packets, thus guaranteeing QoS for sessions with many forwarded packets. Fig. 15 shows that eHPIN has much lower handoff blocking probability than other IPv6-based handoff schemes. This result is due to the ability of eHPIN to reduce signal message exchanges and handoff latency. Thus, eHPIN can safely provide seamless handoff with services continuity.

7. Conclusion

The interworking of networks and mobility management are key issues in 4G/NGWN. Several proposals are available in the literature for these two topics. However, they fail to satisfy basic requirements such as seamless roaming and services continuity for real-time applications across various heterogeneous networks.

This paper proposes an efficient handoff management protocol, called enhanced Handoff Protocol for Integrated Networks (eHPIN), to enable a better network performance in IPv6-based heterogeneous wireless networks.
eHPIN is a one-suite protocol to cope with access network discovery, fast handoff, context transfer and hierarchical mobility mechanisms. Moreover, eHPIN provides efficient handoff management for homogeneous networks as well as for heterogeneous networks.

Performance analyses demonstrate a significant improvement in quality of service (QoS), which is defined in terms of signaling traffic overhead, packet delivery cost, handoff latency, packet loss and handoff blocking probability, compared with existing IPv6-based mobility management protocols. eHPIN alleviates services disruption during handoff by allowing the selection of the best available network and anticipation of handoff. In other words, eHPIN can guarantee seamless handoff, services continuity and QoS in heterogeneous systems. Future work plans consist of validating numerical results through intensive simulation experiments.

References

[13] 3GPP2 TS, 3GPP2-WLAN Interworking; Stage 1 Requirements, 3GPP2 S.R0087-0 v1.0, July 2004.