# UFA: Ultra Flat Architecture for high bitrate services in mobile networks

Khadija Daoud, Philippe Herbelin
Orange Labs
Issy Les Moulineaux, France
{khadija.daoud, philippe.herbelin}@orange.ftgroup.com

Noël Crespi
Institut TELECOM SudParis
Evry, France
Noel.crespi@it-sudparis.eu

**Abstract— The challenge in the coming years for mobile networks will be to offer high bitrate data services to customers in mobility. Future mobile architectures are being standardized to offer mobility between heterogeneous access technologies. The design of these architectures does not take into account scalability requirement since they are centralized with many network levels and dependency. This paper proposes a new architecture, called Ultra Flat Architecture (UFA), that integrates scalability requirement. The key idea of UFA is the reduction of the number of network nodes to one node which is the base station, by the distribution of traditional user and control plane functions in this node. We detail UFA architecture and show how it optimizes service establishment and mobility procedures. We perform a first performance evaluation of the solution performance and expose the first requirements that guarantee seamless handover for real-time services.**

***Index Terms*— scalability, network architecture, mobility, SIP, QoS, future internet.**

## I. INTRODUCTION

SINCE few years, traffic on the Internet is growing exponentially due to an increased subscriber number and to new applications such as P2P and multimedia services, Internet Protocol Television (IPTV) and Video on Demand (VoD). In fixed networks, it is possible to access simultaneously from home to two or three services through a modem connected to a high bitrate last mile link such as Digital Subscriber Line (DSL), cable or fiber. This concept is called multi play service, the most common example being triple play services, widely deployed today. In mobile networks, the last mile link is getting more and more capacities with the new standardized systems, solely based on IP, such as Third Generation Partnership Project (3GPP) Long Term Evolution (LTE) [1] or Worldwide Interoperability for Microwave Access (WiMAX). Moreover, current operator trend is to have Fixed Mobile Convergence (FMC) offers for their customers aiming at providing them in a mobile network the same services as they would have in their fixed network. In this context, we foresee a new kind of service called Multiplay in Motion service (MIM) where the user is able to access his multi play service while in mobility. MIM service will introduce a high load in the network due to the nature of services it transports and to the exponential number of subscribers it would attract.

MIM is a novel and different concept that will have a high impact on the network evolution. We believe that it will introduce scalability issues if deployed on the current specified architectures such as GPRS [2], GAN [3], I-WLAN [4], SAE [5] [6], etc. A new architecture should be defined to solve these issues and to take into account MIM service requirements, i.e. providing high bitrates for users in mobility.

This paper introduces and describes a framework of a new mobile architecture that brings modification to the physical location of a mobile architecture functions and distributes them from centralized nodes to the Base Station (BS). The obtained architecture is called Ultra Flat Architecture (UFA) since it is constituted of only one node. In addition to solving scalability issues, the proposed architecture enables to optimize service establishment and mobility procedures. Furthermore, it applies to terminals using one active interface (e.g. those moving within a homogenous access network) as well as terminals using simultaneous active interfaces (e.g. those moving between heterogeneous systems). UFA is techno-independent and can be used for any kind of access network, wireless or wired. This paper focuses on terminals using one active interface.

The subsequent parts of this paper are organized as follows: section II defines scalability problems on current mobile architecture, section III describes the proposed architecture, section IV evaluates UFA performance in terms of handover delay and finally section V concludes the work and gives the future steps.

## II. PROBLEM DEFINITION

Scalability issues are mainly foreseen for anchor-based mobile architectures. When a Mobile Node (MN) is moving, it may have to change its L2 connectivity and possibly its IP address. Mobility protocols are in charge of updating the path towards the new MN location. A number of mobility protocols have been proposed over the years, they can be classified based on the layer in which they operate [7]. For example, GTP [8] works in the subnetwork layer, Mobile IP (MIP) [9] and Proxy Mobile IP (P-MIP) [10] in the network layer, m-SCTP [11] in the transport layer and SIP [12][13] in the application layer. GTP, MIP and P-MIP are anchor-based and use network tunnels for managing mobility whereas SIP and m-SCTP use end-to-end signaling with no need of anchors and no impact on the infrastructure.

### A. Anchor-based mobility protocols

Subnetwork layer mobility is used when the mobile moves inside one IP subnetwork. It is used in GPRS and UMTS. When the MN asks for a service within UMTS, it is allocated by the GGSN (Gateway GPRS Support node) with an IP address which remains the same for the connection lifetime. Two GTP (GPRS Tunneling Protocol) tunnels are then set between the GGSN and the SGSN (Serving GPRS Support Node) and between the SGSN and the RNC (Radio Network Controller). When the MN moves within UMTS, if its serving RNC changes, the initial GTP tunnels are updated from the SGSN towards the new serving RNC via GTP protocol. Note that the GGSN is seen as an IP anchor point for the MN and does not change whatever the MN movement within UMTS.

Mobile IP (MIP) [9] is used to handle the MN IP address change. It requires a Home Agent (HA) in the network to which the MN IP address (HoA) belongs. When the MN is away from its home network, a care-of-address (CoA) is temporarily assigned to the visiting MN, either by foreign agent (FA) or by other means such as DHCP. Location updates are sent by the MN to the HA to bind the HoA to the CoA. Packets addressed to the MN with the HoA as a destination address are intercepted by the HA and encapsulated within an IP tunnel towards the MN using the CoA as the outer IP address of the tunnel. Currently, MIP is one of the protocols proposed within 3GPP SAE architecture [5] [6] to handle the mobility between 3GPP and non 3GPP accesses (Wifi, WiMax). The MIP HA will be likely to be implemented in the PDN GW.

Proxy mobile IP [10] uses mobility anchors and is similar to Mobile IP. The difference is that in proxy mobile IP, mobility is not executed by the terminal but by a proxy mobility agent located in the network.

### B. End-to-end based mobility protocols

Protocols that manage mobility in an end-to-end way are numerous, such as m-SCTP and SIP. These protocols are specific to a subset of applications depending on their session protocols and/or transport layers. For example, native SIP is not used to manage the mobility of TCP-based applications like FTP. SIP performs mobility in the application layer and does not require additional network infrastructure. If during an active session, the MN gets a new IP address, it sends to the Correspondent Node (CN) a *Re-INVITE* or an *UPDATE* message to update session description. With this update, it informs the CN that it should use the new MN IP address to send the following data packets.

### C. Scalability issues of anchor-based mobility protocols

Mobile architectures with anchors for mobility management are not scalable in MIM context because of their centralized nature. Indeed, IP addresses for users are allocated in high level network elements, called anchor points (GGSN in the UMTS networks and PDN GW in SAE architecture) and there are other intermediate network anchor points: SGSN, RNC in UMTS and SAE GW for SAE. Each of these anchor points maintain a context per MN that binds the MN identity, tunnel identifier, required QoS, etc. Network elements are limited in terms of simultaneous active context, therefore in case of traffic growth new equipments should be added or existing ones should be replaced with more powerful ones. If the traffic grows rapidly and continuously, adopting this solution will be challenging for operators and cannot ensure Return On Investment (ROI) of these equipments.

Fixed networks were subject to the same scalability problems. IP routing function that provides IP connectivity to the users was first implemented within centralized IP routers. When triple play services took off, network architecture has been modified pushing IP routing function close to the subscribers, for example in the DSLAM [14] for DSL services. This has solved scalability problems since the centralized routers are no more needed. Besides, the number of required DSLAMs was not impacted by this modification since the most dimensioning criterion is the number of users physically linked to the network. With this flat and distributed architecture, fixed network investments have been reduced. The same solution can be proposed for mobile network by implementing IP function in compulsory rolled-out equipments, i.e. the BSs. The architecture proposed in this paper is based on this principle; it is called Ultra Flat Architecture (UFA).

## III. PROPOSED ARCHITECTURE: ULTRA FLAT ARCHITECTURE (UFA)

### A. UFA description

Putting the IP function within the BS requires an optimized mobility mechanism as the MN will change its IP address each time it moves from one BS to another. SIP protocol enables to manage mobility without additional equipments; moreover it is the dominant protocol for multimedia services and is adopted by 3GPP as the signaling protocol for the IP multimedia subsystem (IMS). Therefore, we choose to use it to execute mobility, but in a specific way.

Mobility execution by SIP has been widely studied in several papers. Although SIP-based approach does not necessitate any network infrastructure addition, some papers [15][16] mention its drawback regarding the handoff delay it introduces compared to Mobile IP. This delay is due to the time necessary to execute the following steps: 1) attachment to the new BS (L2 handover) 2) acquisition of the new IP network configuration over the new BS, including the IP address 3) local notification of the SIP layer in the MN by the IP layer of the IP address change 4) transmission by the MN and propagation of the *Re-INVITE* message until the CN (application handover) and finally 5) receiving the first data packet from the CN over the new IP address. Some interesting solutions were proposed to reduce this delay due to the five subsequent steps. For example, [17] presents a solution called PAR-SIP in which a major part of IP configuration and application handover steps is anticipated and performed before L2 handover occurs. More detailed, when the MN detects the need of handover, it determines a target BS (T_BS) and informs its serving BS (S_BS). This latter contacts the T_BS

to request an IP address for the MN. The reserved IP address is then transmitted via the S_BS to the MN which sends an advance *Re-INVITE* over the S_BS to the CN before performing L2 handover. The handover delay is thus reduced. One major difference between PAR-SIP and UFA is that PAR-SIP is limited to the case where the handover is decided and executed by the MN. In UFA, the handover is decided and executed by the network which required a specific use of SIP. Another difference is that PAR-SIP neglects QoS aspects which are very important for the operators and users.

Our architecture implements not only IP function in the BS but also higher layers until SIP layer. The BS acts as a SIP B2BUA in order to execute handover on behalf of the MN and reduce handover delay as described in subsection C. B2BUA is introduced in [12] and its internal functioning is not standardized. It consists of two SIP user agents where one user (the server) receives a SIP request, possibly transforms it, and then relays it to the other user (the client) which re-issues it. It maintains the dialog state and can thus participate in all requests sent on the dialog and even send requests on behalf of the MN. Figure 1 gives the protocol stacks in UFA base stations and describes a handover scenario from UFA_S_BS to UFA_T_BS.

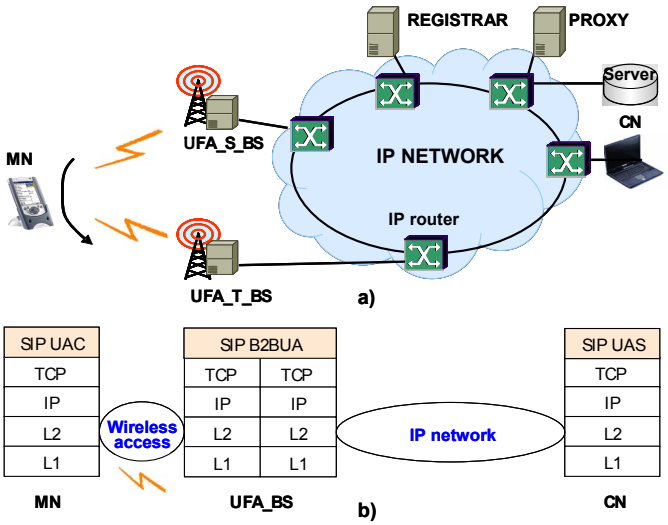


Figure 1: a) UFA architecture; b) UFA protocol stack

Gathering SIP layer and other layers (Transport, IP, layer 2) in the same entity (BS) has other advantages that will be given through the description of service establishment and mobility management procedures.

### *B. Service establishment within UFA*

As illustrated in Figure 2, the MN initiates a video call by sending an *INVITE* message (E1) containing the session description protocol (SDP).

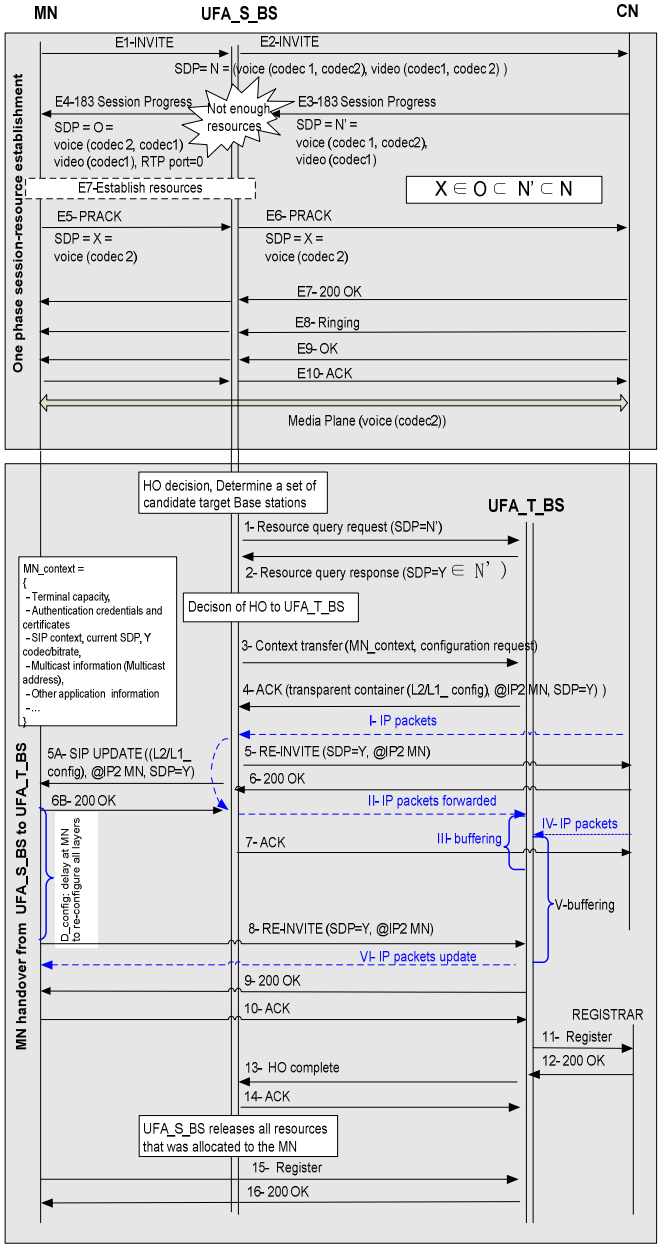


Figure 2: UFA flow chart for service establishment and handover

The SDP contains mainly the MN IP address and a set of proposed media with their related codecs (SDP=N=voice (codec 1, codec2), video (codec1, codec 2)). The CN answers via a 183 session progress message (E3) and indicates the media and the codecs it supports (SDP=N'=voice (codec 1, codec2), video (codec1)) among the proposed ones. UFA_S_BS receives the E3 message and modifies it before generating a new 183 message (E4) towards the MN. SDP modification is needed to take into account UFA_S_BS QoS conditions such as its capabilities in terms of supported bitrates and its available bandwidth. In our case UFA_S_BS has not the necessary resources to support voice and video media flows described in SDP N', thus it proposes another SDP (O=voice (codec2, codec1)) that requires less resources.

In *PRACK* message (E5) the MN will choose an SDP (X= voice (codec2)) based on the SDP O received from the BS. Therefore there is no risk that the session establishment fails due to resource problems. Compared to an IMS session establishment [18] over UMTS access network, UFA session establishment procedure is considered as much more optimized. Indeed, with classical IMS, session establishment is performed in two separate steps. Firstly, the session is negotiated at the application level between the MN and the CN regarding their common capabilities. Secondly, a PDP context is established to reserve resources adapted to the negotiated SDP on all network segments (GGSN, SGSN, RNC, NodeB). If resources are not available, the session establishment will fail.

### C. Handover Management within UFA

The flow chart for handover procedure is shown in Figure 2. The MN scans the neighboring cells and reports the measurements to UFA_S_BS. When UFA_S_BS detects a need to handoff the MN, it sends a *Resource Query Request* message (1) to a list of selected candidate BSs. The requested resources in (1) are expressed using the whole application context (SDP=N') independently of the current allocated resources corresponding to SDP=O on UFA_S_BS. Consequently, the target BS is able to propose in its response (*Resource Query Response* message (2)) a higher SDP (SDP=Y) if its resources enable this. The combination in one step of mobility and service adaptation procedures does not exist in current systems. In 3GPP intra technology handover preparation procedures [2], the requested resources are equal to the current ones allocated on the source BS, even if the initial application request was larger. Another advantage of using SDP for requesting resources is to make UFA handover solution working for inter technology HO situations.

When UFA_S_BS gets the message *Resource Query Response* (2) from the set of BS candidates, it selects one target BS (UFA_T_BS) and begins handover preparation procedure (*Context Transfer* message (3)). UFA_S_BS transfers to UFA_T_BS the whole MN context and requests it provide accordingly the new MN configuration after handover. The transferred context contains for example the MN SIP dialog items and identifier (Call_ID, To TAG, From TAG) in order to configure the B2BUA in UFA_T_BS and prepare it to support the MN after HO. If the context transfer procedure is successful, the UFA_T_BS sends an *ACK* message (4) containing the MN configuration related to all layers from the application layer to the physical layer such as the new MN IP address (@IP2 MN), the new MN L2 resource configuration, etc. After receiving the *ACK* message (4), the UFA_S_BS begins the handover execution procedure. Being a B2BUA, it launches on behalf of the MN a *Re-INVITE* message (5) to the CN containing the new application configuration (SDP=Y) and the new MN IP address. In parallel to this, it sends to the MN a *SIP UPDATE* message (5A) ordering it to attach to the UFA_T_BS. *SIP UPDATE* message (5A) updates the session as does any similar message, and includes the new MN configuration on the other layers (e.g. IP and layer 2). The MN configures itself and attaches to the UFA_T_BS by sending a *Re-INVITE* message (8). At the SIP level, since the SIP context is already transferred from UFA_S_BS to UFA_T_BS in (3), message (8) is not mandatory. It is used to only notify the UFA_T_BS of the MN attachment. When receiving this message (8), the UFA_T_BS sends the *ACK* message (9), transfers data packets to the MN (VI), registers (11) the MN in the REGISTRAR and orders the UFA_S_BS (13) to release the MN associated context and resources.

The concept of B2BUA in the BS has been already proposed in [15] to provide a solution for lossless handover at the IP layer to terminals using two active interfaces simultaneously. UFA uses the B2BUA differently to control the application mapping to resource availability and to make the execution of the handover by the UFA_S_BS feasible. It provides lossless handover for terminals using one active interface or simultaneous active interfaces thanks to buffering mechanism explained in the next section.

## IV. UFA PERFORMANCE EVALUATION

### A. Analytical model

To evaluate UFA performance, we draw on Figure 3 UFA handover timing diagram based on the flow chart given in Figure 2. Performance Key Indicators (PKI) are calculated based on the processing and transmission delays (Dbs, Dw, Dn, Dp, D_config) of the messages involved in handover procedure, as detailed in Table I.

Table I: List of delay components

| Notations | Meaning |
|---|---|
| Dbs | Transmission delay of a packet over the interface between two BSs |
| Dw | Transmission delay of an IP packet over the wireless link between the MN and the BS |
| Dn | Transmission delay of an IP packet over the wireless link between the MN and the CN |
| Dp | Processing delay within equipment corresponding to the treatment of the received or transmitted packet. It is considered the same for all messages and equipments. |
| D_config | Time necessary for the MN to perform L2 synchronization and to configure its layers according to the content of message (5A) |

Due to the varying nature of internet delay and the computing power of intermediate nodes, it is difficult to estimate Dn and Dbs. For this reason, we consider them as constant and independent of the message length. On the contrary, Dw can be determined based on the wireless bitrate (w_bitrate) and on the length (X) of transmitted messages.

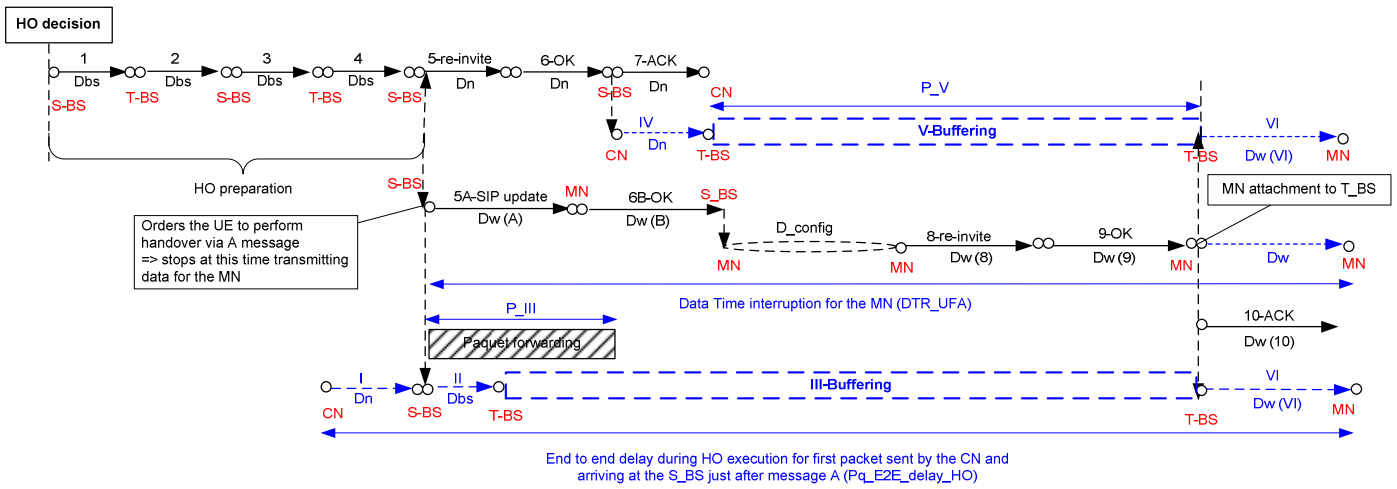


Figure 3: Handover timing diagram

Dw(X) is given by:

$$Dw(X) = k * \tau \qquad (1)$$

$$k = \frac{X}{w\_bitrate * \tau} \qquad (2)$$

Where:

- k is the number L2 frames per an IP packet
- $\tau$ is the frame duration

During handover procedure, data packets are buffered in UFA_T_BS until the attachment of the MN (message 9). Buffered packets are of two types. The first type is shown in step *III- buffering* and corresponds to the packets forwarded by UFA_S_BS to UFA_T_BS after the departure of the *Re-INVITE* message (5). Packet forwarding lasts for (P_III=2*Dn). The second type is shown in step *V-buffering* and is related to packets sent by the CN to UFA_T_BS after the transmission of the OK message (6). This buffering lasts for (P_V= Dw(5A) + Dw(6B) + D_config + Dw(8) + Dw(9) + 2Dp – 3* Dn) and occurs only when this expression is positive which is roughly equivalent to the condition (Dw> 3* Dn).

In order, to provide seamless handover, the E2E delay of the buffered packets should not exceed 200 ms for voice services [20]. Also to avoid packet loss, the buffer size in UFA_T_BS should be adapted to the quantity of data received during the period ($P = P\_III$ or $P = P\_III + P\_V$ if P_V>0). We define thus the following PKIs which expression can be deduced rapidly from Figure 3:

• Pq_E2E_delay_HO: The maximum end to end (E2E) delay between the CN and the MN for all buffered data packets (calculated for the first packet waiting in the buffer for transmission).

• Pq_W_delay_HO: The delay introduced by the wireless link in the Pq_E2E_delay_HO.

• UFA_T_BS_buffer_size: The minimum UFA_T_BS buffer size.

$$Pq\_E2E\_delay\_HO = Dn + 2 * Dp + Pq\_W\_delay\_HO \qquad (3)$$

$$Pq\_W\_delay\_HO = Dw(A) + Dw(B) + Dw(8) + Dw(9) + D\_config + Dw(VI) + 9 * Dp \qquad (4)$$

$$BS\_buffer\_size = \lambda * P \qquad (5)$$

Where λ is the packet arrival rate at the UFA_T_BS

## B. *Numerical results*

Our aim is to evaluate the performance of UFA for three access systems (UMTS, HSDPA, LTE) and to determine the adequate bitrate over which handover related signaling shall be mapped. For the messages *5A-UPDATE, 6B-OK, 8-Re-INVITE, 9-OK*, we considered respectively the following lengths: 1200 bytes, 900 bytes, 900 bytes, 500 bytes. These lengths are similar to those considered in IMS [18].

In a first step we use typical values for Dbs, Dn and Dp (Dbs=20 ms, Dn=60 ms, Dp=2 ms, D_config=40 ms) and measured the Pq_E2E_delay_HO. Figure 4 shows the evolution of this parameter for different possible bitrates and for each access system. For a given value of bitrate, we observe different values of Pq_E2E_delay_HO due to the difference between the considered systems in their way of providing access to the wireless media ($\tau$ = 20 ms for UMTS, 2 ms for HSDPA, 1 ms for LTE). If we consider voice service, which is the most constringent service in terms of E2E delay (200ms), we conclude that UFA should not be applied for bitrates inferior to 384 kbps as in UMTS R99. Note that this is a severe conclusion since we can relax the condition (Pq_E2E_delay_HO<200ms) and tolerate that the delay for the first packets in the buffer exceeds 200 ms.

Results given above apply for a specific value of Dn. This parameter is difficult to estimate and varies from one correspondent node (CN) to another (the CN could be located anywhere on an IP backbone (locally, regionally, nationally or worldwide)). One way to control and tune Dn is to anchor SIP sessions in additional intermediate equipments (B2BUA) under

the control of the operator. The Dn value is thus reduced to the transport delay between the BS and the added B2BUA. Another advantage of the B2BUA is to handle the cases where the CN is not able to process the *Re-INVITE* message (5). B2BUA should be located on the same Autonomous System (AS) and/or under the same network authority and respect the maximum Dn value. We plot in Figure 5, for HSDPA, the parameter Pq_W_delay_HO which represents the delay introduced by the wireless link in the Pq_E2E_delay_HO. The difference between this parameter and the E2E delay requirement (e.g. 200 ms) enables to determine the maximum Dn value. We observe that for 384 kbps Dn value should not exceed 50 ms which is feasible on high speed network interfaces.

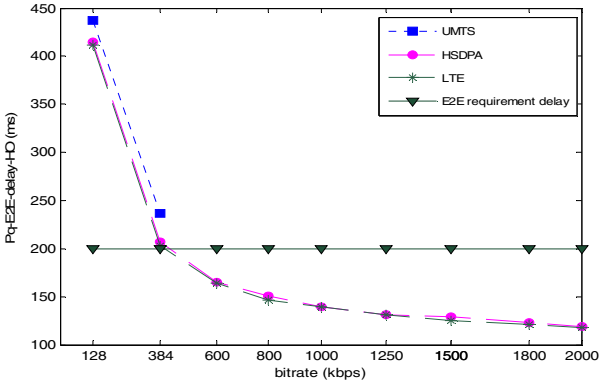


**Figure 4:** E2E delay during HO (Pq_E2E_delay_HO)

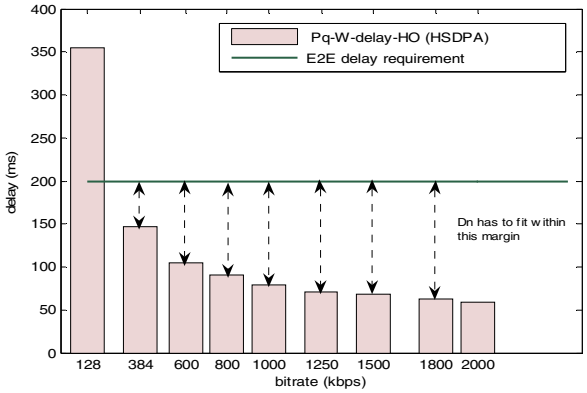


**Figure 5:** Impact of the wireless link on the E2E delay (Pq_W_delay_HO)

## V. CONCLUSION AND FUTURE WORKS

To solve scalability issues and optimize QoS aspects during call establishment and mobility, we proposed in this article a new mobility architecture, called UFA, that modifies current mobility architectures from many nodes to only one node by incorporating all functions in the BS. One of the characteristics of this architecture is to enable the execution of the handover by the network via the SIP protocol. A first evaluation of this architecture regarding handover performance shows that handover related messages should be mapped on bitrates higher than 384 kbps.

Work will be continued regarding the following aspects: UFA testbed to better measure handover performance and to provide a proof of concept, quantification of UFA advantages regarding QoS aspects in comparison with the policy control and charging (PCC) framework defined by the 3GPP, the support by UFA of the mobility of terminals using non-SIP applications, adaptation of UFA to IMS related functions such as user authentication, etc.

## REFERENCES

[1] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description".

[2] 3GPP TS 23.060, "General Packet Radio Service (GPRS); Service description"

[3] 3GPP TS 43.318 , "Generic access to the A/Gb interface".

[4] 3GPP TS 23.234, "3GPP system to wireless Local Area Network (WLAN) Interworking; system Description".

[5] 3GPP TS 23.401, "3GPP System Architecture Evolution: GPRS enhancements for LTE access".

[6] 3GPP TS 23.402, "3GPP System Architecture Evolution: Architecture Enhancements for non-3GPP accesses".

[7] N. Banerjee et al., "Mobility Support in Wireless Internet," IEEE Wireless Commun., vol. 10, no. 5, 2003, pp. 54–61.

[8] 3GPP TS 29.060, "General Packet Radio Service (GPRS), GPRS Tunnelling Protocol (GTP) across the Gn and Gp Interface", stage 3.

[9] C. Perkins, "IP Mobility Support for IPv4," IETF RFC 3344, Aug 2002.

[10] Gundavelli, S., et.al, "Proxy Mobile IPv4", draft-ietf-netlmm-proxymip6-16.txt, November 2007.

[11] M. Riegel, and M. Tuexen, "Mobile SCTP", draft-riegel-mobile-sctp-09.txt, November 2007.

[12] J. Rosenberg et al, " SIP: Session Initiation Protocol", IETF RFC 3261, june 2002.

[13] H. Schulzrinne and E. Wedlund, "Application-Layer Mobility Using SIP," Mobile Comp. and Commun. Rev., vol. 4, no. 3, 2000, pp. 47–57.

[14] DSL Forum TR-101, Migration to Ethernet-Based DSL Aggregation, April 2006.

[15] N. Banerjee et al., "Seamless SIP-Based Mobility for Multimedia Applications", IEEE Network , March/April 2006.

[16] W.wu, N. Banerjee, K. Basu, S.K. Das, "SIP-based vertical handoff between WWANs and WLANs Wireless Communications", IEEE wireless communications, Special Issue: Towards Seamless Interworking of Wireless LAN and Cellular Networks, Vol 12, Issue 3, June 2005, pp. 66 – 72.

[17] Wooseong Kim, Myungehul Kim, Kyounghee Lee, Chansu Yu, Ben Lee, "Link Layer Assisted Mobility Support Using SIP for Realtime Multimedia Communications", *MobiWac'04*, October 2004.

[18] 3GPP TS 23.228, "IP Multimedia Subsystem (IMS)", stage 2, december 2006.

[19] A. Nasir, M. Rukh, "Internet Mobility using SIP and MIP", Proceedings of Third International Conference on Information Technology: New Generations, (ITNG'06), 10-12 April 2006; pp. 334 – 339.

[20] ITU-T Rec. G.114 (05/2003) One-way transmission time.