An Overlay and Distributed Approach to Node Mobility in Multi-Access Wireless Networks

Paolo Dini, Jaume Nin Guerrero, Nicola Baldo
Centre Tecnologic de Telecomunicacions de Catalunya,
av. C.F. Gauss 7, 08860 Castelldefels, Spain.
E-mail:{paolo.dini, jaume.nin, nicola.baldo}@cttc.es.

Abstract— Nowadays many manufacturers are building mobile devices with multiple interfaces. Thus, users have access to different types of wireless access networks, which often, as for WLAN and cellular systems, coexists independently. The challenge is to make such multiple access networks to cooperate to have ubiquitous access and enhanced user quality of service. In this paper we present a scheme to allow inter-technology mobility by the introduction of an overlay network, which works on top of current (and future) networks. The proposed architecture controls all the aspect related to the mobility management: mobile node localization, handover decision and execution. The approach is distributed: it is the mobile node that decides which network to use, based on the offered service quality and the cost of the communication of the available networks, and triggers the handover execution directly to the corresponding host, using optimized SIP-based procedures. The overlay network copes with the mobile node localization. We implemented our solution in the laboratory to prove its validity and to test performance using real equipment. We also simulated the scheme using ns-3 to extend the evaluation to large scale deployments. In both test environments, our solution demonstrates high accurateness in selecting the network with the best quality as well as in supporting seamless vertical handover.

Index Terms—wireless heterogeneous networks, multiple access networks, vertical handover, distributed mobility, overlay networks, IEEE 802.11 WLAN, 3G, UMTS/HSPA, cellular networks, experimental testbed, interface selection, IMS, SIP, VoIP, QoS

1 INTRODUCTION

NOWADAYS, a significant fraction of mobile devices in the market is equipped with multiple wireless interfaces, mainly to connect to cellular systems and WLANs. These two technologies have been developed separately and therefore offer different coverage, data rate and services. Cellular networks have been designed to have high coverage area, low/medium data rates and mainly to provide speech services. On the contrary, IEEE 802.11 WLANs have been designed for local and small areas, medium/high data rates and mainly for data services. Such technologies can coexist in the same environment and also interwork so as to create a heterogeneous platform able to offer more services, higher coverage and higher data rates to mobile end-users. This system can realize the Always Best Connected paradigm introduced in [1], i.e., a system with global coverage (outdoor and indoor) in which mobile users can always be attached to the network which offers the best Quality of Service (QoS).

The recent introduction of femtocells as a solution to the macrocell indoor capacity problem [2] has opened up the interesting possibility of using the same technology to provide both indoor and outdoor high data rate access, thus potentially eliminating the need for interworking between the different networks as well as the need for handsets with multiple interfaces. Though this approach is appealing, its widespread diffusion is still blocked by several open issues, such as (i) spectrum sharing between macro and femto cells, which is not trivial due to the unplanned femto deployment [3], (ii) management of the femtocell backhaul [4] and (iii) network neutrality. On the other hand, enabling cellular-WLAN interworking has several clear advantages, like: (i) the use of the unlicensed band for WLAN transmissions, which is greater and not overlapping with the licensed spectrum used by the cellular system, thus no interference management is needed; (ii) the deployment of WLAN access points is already massive and (iii) most of the mobile devices nowadays have more than one interface.

Seamless inter-technology mobility is a key issue for the creation of an integrated wireless heterogeneous system. Several architectures have been proposed in the literature to solve the integration issue. Basically, we can divide these solutions in two main categories: tightly coupled and loosely coupled architectures. Tightly coupled architectures look at WLANs as another type of access network within the cellular system, for example 3G; in this case WLAN is directly connected to the RNC or to the SGSN, so as to speed up the inter-technology handover procedures. In this case, new interfaces have to be defined among the cellular and WLAN architecture entities, thereby complicating the design of the interworking architecture, especially when the two networks belong to different operators. Besides, high data rate traffic from WLAN should pass through the 3G core network, which has been designed and dimensioned for other types of traffic and services. Bottlenecks and congestions can occur in this situation. Differently, in a loosely coupled architecture WLAN is directly connected to the Internet, thus avoiding the definition of new interfaces as well as the transit of WLAN traffic through the cellular core network. Therefore loosely coupled architecture offers easier implementation than the tightly coupled, even though mobility procedure could suffer a greater delay due to the lack of direct connection between the two networks.

In this paper we propose an inter-technology network architecture divided in two logical planes: transport and vertical mobility plane, as shown in Figure 1.

A transport plane is a set of different data networks. It is devoted to user data and to the needed signaling for single networks management. The vertical mobility (VM) plane is in charge of controlling mobility procedures across the different underlying networks. Connectivity between the entities of the two planes is assured via standard IP-based interfaces. The VM plane can be owned by the same operator managing a transport plane as well as by a third party, who offers vertical mobility as a service to its customers; in this case service agreements between the VM and the transport operators are needed.

The use of Session Initiation Protocol (SIP) has been chosen as communication protocol in the VM plane. The multi-interface mobile node (MIMN) is identified by a unique SIP URI, such as alice@proxy.com. The URI is registered in its SIP server (i.e., proxy.com in this example) together with its multiple IP addresses (one for each of the active wireless interfaces), through which it is reachable by correspondent nodes (CNs). SIP provides a natural way to associate high-level identifiers (i.e., SIP URI) and interface identifiers (i.e., IP addresses) than that defined by Mobile IP, which suffers from the well-known semantic overloading problem of IP addresses. ([6], [14]). Additionally, SIP does not require any modifications to existing network layer protocols, which eases deployment. In fact, it works with both IPv4 and IPv6 and with UDP and TCP ([7]). Finally, though commonly used mostly for Voice over IP applications, SIP can in principle be used to manage any kind of application.

In this paper we focus on the definition of algorithms, methods and procedures to offer always the best quality of service to the mobile users of the heterogeneous platform. Based on the SIP architecture just described, the present paper proposes a complete mobility management framework that addresses in particular the following aspects:

- mobile node localization,
- inter-technology (or vertical) handover decision,
- inter-technology (or vertical) handover execution.

The approach is distributed: mobile nodes decide which network to use, based on the offered service quality and the cost of the communication, and trigger the handover execution directly to the correspondent node; the overlay network copes with the mobile node localization. Such a user-centric approach is advantageous for the end-user since the choice of the transport network to be used can account also for user preferences (e.g., cost, battery energy consumption). Besides, considering that the different transport networks often belong to different operators or even private citizens (e.g., home WLAN hotspots), a distributed solution has more chances to be deployed because it is more technology and operator transparent. One might argue that, in such a user-centric approach, a selfish node could potentially disobey the defined rules to selfishly enhance its performance. However, it is worth highlighting that in such a scenario, there is no incentive for a mobile node to misbehave (e.g., by joining a network which is already at its maximum capacity), since this would lead to poor service quality for the misbehaving node as well.

Though the focus of the paper is on a distributed approach the proposed architecture also envisages the possibility to trigger the vertical handover by the VM plane, for example to offload traffic from a network, thus adding certain flexibility to the proposed solution.

The reminder of the paper is as follows. In section 2 we review related work. In such section we also present the contributions of our proposals with respect to the state-of-the-art solutions. In section 3 we introduce the methods for vertical handover decision and execution. Section 4 describes the implementation of the proposed solutions on the EXTREME testbed® ([12]) as well as the scenarios used for the performance evaluation. Section 5 is devoted to performance evaluation. The paper is concluded with some observations and remarks.

2 Related Work and Paper Contributions

2.1 System Architecture

We focused our reading on solutions where authors introduce or exploit overlay network to perform mobility management. The IP Multimedia Subsystem (IMS) ([5]) is a clear example. It promotes the clear separation between signaling for session control, authentication, authorization and accounting and data transportation. The ambitious goal of IMS is to achieve a common platform for controlling IP multimedia sessions, to support mechanisms to negotiate the QoS, to foster service creation without requiring standardization and to enable interworking with the Internet and the circuit-switched networks.

In [7] the authors propose the IHMAS (IMS compliant Handoff Management Application Server), which exploits the separation between the signaling and data delivery planes to perform advanced data handover management. In particular, clients are provided with a vertical handover predictor that starts session reconfiguration before triggering the handover requests. The authors tested their solution experimentally, thus demonstrating the suitability of an application layer approach to improve performance in session mobility during handovers in the IMS. The main issue we envisage is in the fact that they rely completely on the IMS infrastructure, which is hard to implement in all its entities and procedures.

In fact in [9] the authors highlight the technical difficulties in implementing an IMS or a Multimedia Domain (MMD, which is the 3GPP2 equivalent of the IMS) architecture. They describe their platform focused on MMD-compliant mobility test and analyze different handover mechanisms and the associated functionalities.

On the other hand, an overlay ring to manage mobility is introduced in [10]. The authors propose the use of the peer-to-peer paradigm for mobility management, to avoid the classic problems of client-server like single point of failure, bottlenecks and risk of congestion. They virtualize the functionalities offered by home and foreign agents by the creation of an overlay network of mobile agents, where all mobile agents are fixed nodes playing the role of home agent and foreign agent. The paper experimentally validates the architecture, but does not enter the details of a real implementation in a wireless heterogeneous system. Later the authors extend their work to vertical peer-to-peer mobility in [11], based on the same approach.

2.2 Vertical Handover Decision

Let us specify our approach to vertical handover decision. When the MIMN is in the coverage of more than one network, it has to choose the best one to use for its communication, in terms of monetary cost and offered

QoS. 3G plus WLAN is taken as example of a heterogeneous system in the rest of the paper. Nevertheless, our approach can be extended to any other wireless access technology (e.g., WiMax, LTE).

Due to the higher monetary cost of a connection established through a cellular network, users may prefer to use WLAN, when available. However, WLANs cannot guarantee any QoS as cellular networks, which is a key requirement especially for real-time services (e.g., VoIP). Call admission control (CAC) algorithms running in the MIMNs can help in evaluating the available resources and to estimate the quality offered by the WLAN to the ongoing and the upcoming sessions. Therefore, we concentrate our study on distributed CAC for WLANs, more than on vertical handover algorithms, such as [33], [34], [35].

A limited number of distributed CAC schemes can be found in the literature. In [36] an algorithm is proposed, which relies on each mobile station (STA) doing an active probing of the wireless link to infer the achievable service quality. While able to make an effective CAC decision, this solution has the disadvantage of increasing the control traffic overhead of the network, potentially harming ongoing data communications by other users. In [37] the authors propose a perceptive admission control for 802.11 ad-hoc networks based on the busy-time ratio metric. A similar approach is investigated in [38], where the authors consider an infrastructure WLAN and introduce a new metric, the Time Between Idle Times (TBIT), which is shown to yield a very effective CAC algorithm. In fact, the use of the TBIT metric is more effective than the scheme in [37], and it can be considered the state of the art in distributed CAC for VoIP over WLAN. Nevertheless, we note that the method for measuring the TBIT metric does not consider the time consumed by erroneous transmissions (e.g., collisions), and due to this reason the measured number and duration of the idle time periods are not correct. Furthermore, in [38] the performance evaluation (i) does not consider the presence of heterogeneous voice codecs, (ii) does not consider the case in which STAs use different PHY rates, (iii) does not show any numerical result for the case in which background traffic is present and (iv) does not consider objective quality metrics as E-model [21]. All these conditions are commonly encountered in the vast majority of real WLAN scenarios, and could negatively affect the performance of a distributed CAC scheme.

2.3 Vertical Handover Execution

As stated in Section 1, SIP has been chosen as the communication protocol of the VM plane. It is used for session setup and management as well as for the vertical handover execution (see [13]).

In [17] the authors propose a modification of the standard SIP signaling and architecture described in [11] and [13] to provide support for vertical handovers without disruption of real-time multimedia services. They introduce two new entities in the standard architecture to handle the proposed signaling scheme, which is implemented in a testbed. The solution seems valid, though no experimental result is shown in the paper. Furthermore, it is not totally compliant with the standard SIP architecture by [12] and [13] because of the introduction of new network entities.

In [18] the authors use SIP to perform vertical handovers and also to evaluate its performance when switching between WLAN and UMTS. Mobile node interfaces have no IP address when the handover is triggered. Then, network specific registration and IP assignment procedures are necessary during the handover (e.g., DHCP - Dynamic Host Configuration Protocol and PDP Context Activation procedures). The analysis concludes that the handover from WLAN to UMTS is very slow because of the PDP Context Activation procedure in UMTS. Therefore, seamless interface switching is not feasible due to packet losses. To tackle this problem, they propose a handover procedure based on the duplication of data traffic during switching time, in which the MIMN simultaneously transmits the same data through both interfaces to decrease the number of lost packets. The application at the receiver is in charge of discarding duplicate packets. The proposed method doubles the bandwidth used during the interface switching, due to duplicate packet transmission, thus leading to an inefficient use of the already scarce radio resources.

In [19], to ensure that no packet is lost during the interface switching, the authors introduce the concept of soft handover in a SIP-based vertical handover in a pure IP wireless network. In particular, a SIP proxy agent is embedded within the base stations. Then a new SIP procedure between the old and the new serving base station is added to the standard SIP mid-call mobility procedure. It consists on the insertion of a JOIN header option in the re-INVITE message. Consequently, they define a new architecture with the presence of a SIP user agent also in the base stations. The paper presents a sort of distributed mobility scheme where different network elements share the control of the handover. Nevertheless, it is a modification of the standard SIP architecture defined in [11] and [13]. Three fast handover schemes for SIP mobility have been presented in [46]. They are based on the introduction of logical entities in the end-to-end communication path to decrease the

time in which the handover is executed. In particular they propose to use SIP registrar and RTP translator, SIP outbound proxy and back-to-back SIP user agent. Then, a multi-interface mobility module able to manage both Mobile IP and SIP mobility is proposed in [47]. The focus of the paper is more on the comparison between MIP and SIP procedure though.

In [20] the authors present the issues related to SIP mobility without giving a solution. The concept of registering the multiple interfaces of the mobile node in the SIP server is introduced in the paper but no explanation on how SIP can manage multihomed hosts is given.

A step forward towards SIP-managed multihoming is presented in [21]. The authors propose an association scheme between the SIP URI of a user and its multiple IP addresses (each associated to one network interface) managed by the SIP server. The interfaces are sorted by signal strength and traffic load within the SIP server. According to this scheme, the paper is describing only pre-call mobility. But no explanation about the signaling for mid-call mobility is provided.

2.4 Paper contributions

Our paper has the following contributions compared to the work described above:

- definition of an overlay architecture to manage vertical mobility in a multi-access and multi-technology environment, using legacy SIP procedure for multi-interface mobile node localization and vertical handover execution.
- implementation and experimental evaluation of the proposed architecture; the relevant procedures for the vertical handover execution and the algorithms for the vertical handover decision are tested using real equipment in a fully-configurable testing framework for wireless heterogeneous networks.
- optimization of legacy SIP mobility procedures to obtain:
  - resilient mobile node reachability;
  - preservation of radio resources by limiting signaling and unnecessary data traffic through wireless links;
  - reduced packet losses during the interface switching;
- distributed vertical handover decision based on the estimation of the available resources of the candidate networks. In particular, we focused on the choice between cellular networks and IEEE 802.11 WLANs. The decision is more accurate than other solutions presented in the literature (mainly, the TBIT proposal presented in the above), since it considers not only channel time spent in successful transmissions but also the time consumed by collisions and channel errors.

Our paper aims at having a holistic approach to the vertical mobility issue. In fact, it analyzes the problem in all its complex aspects, from the localization of nodes, to the vertical handover execution, passing through the vertical handover decision. The main aim is to always support the mobile users with the best QoS, which can be offered by the available transport networks.

3 VERTICAL MOBILITY SCHEME

3.1 Multi-interface Mobile Node Localization

When the MIMN, with its SIP URI already registered in the SIP Registrar, moves to one network to another, it changes the IP address of one of its interfaces. The VM plane must have the updated addresses of its mobile nodes in order to always guarantee their reachability, i.e., to forward all the incoming call requests.

The main aim of the optimization proposed in this sub-section is to avoid sending registration update messages to register the new interface IP address during the vertical handover. The goal is to save radio resources as well as to speed up the vertical handover execution procedure and to offer, at the same time, a resilient reachability to mobile nodes.

In compliance with the pre-call mobility procedure described in [13], the MIMN registers its SIP URI and all its available interfaces in its SIP server by sending a REGISTER message. In our scheme, the MIMN has to add weights associated to each of its interfaces in the ‘q field’ of the contact header. The weights are assigned based on user preferences to create a priority list of the registered interfaces within the SIP server. The priority list is used by the SIP server to forward signaling messages for setting up new calls to the MIMN. For instance, a user of a mobile device with WLAN and cellular interfaces may prefer to be always reachable, and thus, he/she may assign higher priority to the cellular interface due to the higher coverage of the network. Therefore, the cellular interface would be used to forward signaling messages first. Signaling would only be sent to the

WLAN interface when no answer is received from the higher priority interface. In this way, a resilient connection between the MIMN and its SIP server is provided.

At the end of this procedure, all the active IP addresses of the MIMN are registered and listed, sorted by user-assigned priority, in the SIP server, which does not need any registration update message during the vertical handover execution procedure.

In summary, the MIMN localization procedure takes place when:

- the MIMN is switched on, after the network specific registration and IP address assignment procedures,
- a new interface is available at the MIMN; for example, when it enters in the coverage area of a new network and it obtains a new IP address,
- the user decides to change his/her interface priority list.

Note that this priority list is used only to forward signaling messages and not data traffic. The decision of the interface to be used for a given data flow is taken when the session starts and it depends on the output of a decision algorithm. The interface may also change during an ongoing session, for example, because the MIMN detects that another of its interfaces can provide better session quality. This situation is managed by the vertical handover execution procedure.

3.2 Vertical Handover Decision

If the MIMN has an ongoing session and it is in the coverage area of both cellular and WLAN networks, our assumption is that user preference is to use the WLAN, because of its lower monetary cost. However, due to the lack of QoS management in 802.11 WLANs, it is necessary to sense the channel and estimate the quality supported by that visible access point, so that users can always be attached to a network providing the required quality for his/her services. The MIMN has to run a CAC algorithm for real-time sessions, before triggering the vertical handover execution (i.e. the initiation of the SIP procedure described in Section 3.3), to understand whether a new call can be admitted to the WLAN with good quality. CAC relies on a load estimation method, which estimate the quality offered to the ongoing sessions and forecasts such quality if the new call enters the network. The figure below draws the vertical handover decision procedure for an already active connection.

3.2.1 Channel Load Estimation and Distributed Call Admission Control

The proposed CAC algorithm, hereinafter also referred to as distributed CAC (DCAC), is designed for IEEE 802.11 Wireless LAN infrastructures. We assume that the eventual presence of hidden nodes has a negligible impact on the communication performance, and that the capture effect can be neglected as well.

We focus on a mobile station (STA) listening for a certain period to the channel where an Access Point (AP) is operating. In such situation, the channel can be either busy, if there is at least one ongoing transmission, or idle if there is none. At a moment when the channel is busy, it can be occupied by either a successful or an unsuccessful frame transmission, depending on whether the frame is correctly received or not. Unsuccessful receptions can be due to packet collisions among active nodes and/or channel errors. Considering typical devices and deployment scenarios, the occurrency of channel errors is significantly reduced by the adoption of rate adaptation mechanisms which dynamically choose an appropriate modulation and coding scheme based on the channel propagation conditions. On the other hand, collisions are relatively frequent in WLAN deployments [40], with collision probabilities typically of the order of a few tenths during normal operation. Additionally, we note that, while the ARQ mechanisms adopted in WLANs effectively prevents network-layer packet losses, link-layer errors and collisions still consume a significant amount of channel resources.
The CAC scheme that we propose works by estimating the channel busyness ratio of a WLAN taking into account as well this link-level overhead due to errors and collisions.

As a representative real-time application, in the remainder of this paper we focus on voice over IP (VoIP) traffic. We consider an heterogeneous scenario where VoIP traffic is coexisting with non-real-time background data traffic. Let ρ_v and ρ_{bg}be the fraction of channel time occupied by successfully received frames belonging respectively to voice and background data traffic; ρ_f be the fraction of channel time occupied by failed frame transmissions; finally, let ρ_{bo} be the fraction of channel time occupied by the back-off procedure. The following equation holds:

$$\rho_v + \rho_f + \rho_{bg} + \rho_{bo} \leq 1 \qquad (1)$$

Our proposed DCAC algorithm runs on a STA and uses the channel load model of (1) to estimate whether a new VoIP session could be started on a certain AP with an acceptable Quality of Experience both for the new

and for the already ongoing VoIP sessions. The algorithm works as follows:

- the STA monitors the wireless channel in which the AP is operating;
- based on the information gathered by channel monitoring ([48]), ρ_v and ρ_f are determined;
- using the method described in Section 3.2.2, the channel load $\widehat{\rho_v} + \widehat{\rho_f}$ expected after the introduction of the new VoIP session is determined as a function of the actual values of ρ_v and ρ_f;
- using the method described in Section 3.2.2, the forecast fraction of channel time $\widehat{\rho_{bo}}$ consumed by the backoff procedure is determined according to the method described in the following subsections.
- the term ρ_{bg} is approximated by the pre-determined fixed values $\overline{\rho_{bg}}$.
- the AP is deemed to be eligible for the new VoIP session if $\widehat{\rho_v} + \widehat{\rho_f} + \widehat{\rho_{bo}} \leq 1 - \overline{\rho_{bg}}$.

We note that the algorithm described is formulated with VoIP in mind; however, it can be applied as well to other types of real-time traffic, such as video conferencing and video streaming. We suggest that a production version of this algorithm shall include some hysteresis mechanism to avoid the ping-pong effect; however, further discussion of this issue goes beyond the scope of this paper.

3.2.2 Calculation of channel load

We assume that the STA running the DCAC algorithm monitors the wireless channel over a period of duration T_W, and gathers the information contained within the MAC and PHY headers of the captured data frames. In the vast majority of commercial IEEE 802.11 devices, only successfully decoded frames can be observed with this procedure. Let the term *frame exchange sequence* denote the sequence of a voice DATA frame plus its following ACK. Let the index i denote the generic observed frame exchange sequence, and let T_i denote its duration, which accounts for the duration of the DATA and ACK frames as per [29], plus one DIFS or AIFS depending on whether QoS support is used, plus one SIFS. Finally, let N denote the total number of frame exchange sequences in the observation period.

The value of ρ_v is calculated as

$$\rho_v = \frac{\sum_{i=1}^{N} T_i}{T_w} \qquad (2)$$

Let us now consider the eventual start of a new VoIP flow by the STA running the DCAC algorithm. Let λ_{new} be the cumulative number of packets per second of the corresponding two new VoIP flows which are to be introduced (from the STA to the correspondent node and vice versa), and let T_{new} be the duration of the frame exchange sequences for these flows. We note that both λ_{new} and T_{new} are known by the STA from SIP signaling and PHY rate adaptation information.

The value $\widehat{\rho_v}$ expected after the new VoIP session is started is then calculated as:

$$\widehat{\rho_v} = \rho_v + \lambda_{new} T_{new} \qquad (3)$$

As for the calculation of ρ_f, we note that a monitoring STA cannot observe directly the occurrence of collisions and channel errors; in fact, most commercial devices are limited to the observation of successfully decoded frames, as we stated previously. A few implementations actually allow gathering frames which are detected as erroneous due to the failure of the MAC CRC; however, even these devices would not be able to collect frames lost due to collisions, since a collision is most likely to corrupt the PHY header of the frame, thus preventing the determination of the duration of the transmission. In principle, it could be possible to estimate the number of collision events by counting the number of PHY errors reported by the device, as proposed in [39]; however, according to our experience, many consumer devices often report more than one PHY error for the same frame, and furthermore PHY errors often happen for causes other than collisions, e.g., due to transmissions by non-802.11 devices and spurious synchronization events. Hence, we suggest that, because of these reasons, it is not practical to use the number of PHY errors to derive an estimation of ρ_f. As a final remark, we report that there are some analytical models, such as [40] and its many derivates, that provide means for estimating the fraction of channel time occupied by unsuccessfully received frame transmissions. These methods are based on the assumption of saturation traffic (i.e., each STA is backlogged). Unfortunately, this assumption does clearly not hold for VoIP STAs, which is the case that we consider.

To summarize, we could not find in the Literature any method for the estimation of ρ_f in our scenario; hence, we derived our own method, which is based only on the observation of successful frame exchange sequences. We leverage on the presence of the so-called *retry* flag in the IEEE 802.11 MAC header, which indicates whether a DATA frame is an initial transmission attempt (value set to false) or is a subsequent retransmission (value set to *true*). Let n_s and n_r be the number of successfully decoded DATA frames (both VoIP

and background traffic) having the retry flag set to *false* and *true*, respectively. Let us assume that whenever a DATA frame is delivered also the subsequent ACK is successfully delivered; this assumption is consistent with our previous assumption of negligible traffic by hidden terminals. Let n_{MSDU} be the total number of MAC Service Data Units (MSDUs) which have been successfully delivered, i.e., the number of DATA packets at the networking layer who experienced a successful transmission attempt within the retransmission limit (r_{max}). Then we have:

$$n_s + n_r = n_{MSDU} \tag{4}$$

The above equation is explained considering that, since only the final successful transmission attempt can be observed for every successfully delivered MSDU, then a generic successfully delivered MSDU will be counted either in n_s or in n_r, depending on whether the first transmission attempt was successful or not. We now make the further assumption that transmissions fail mostly due to collisions, and that all frame transmission attempts in the observation period have the same failure probability P_f; we note that this hypothesis is commonly adopted in the literature (see [40] and [49]). Based on this assumption, the failure probability of the first transmission attempt for a successfully delivered MSDU is also equal to P_f. Hence, P_f can be estimated as:

$$P_f = \frac{n_r}{n_{MSDU}} = \frac{n_r}{n_s + n_r} \tag{5}$$

Considering that the ARQ procedure adopted by the IEEE 802.11 standard is of type stop-and-wait, we can derive an expression the expected value E[k] of the number k of failed transmission attempts per MSDU, conditioned on the event that the MSDU is delivered within the retransmission limit. This is a function of P_f, which can be expressed as:

$$\begin{aligned} E[k] &= \sum_{k=0}^{r_{max}-1} k \Pr(k \text{ failures} \mid \text{MSDU delivered}) \\ &= \sum_{k=0}^{r_{max}-1} k \frac{1-P_f}{1-P_f^{r_{max}}} P_f^k \end{aligned} \tag{6}$$

Based on this, we can estimate ρ_f by means of the following equation:

$$\rho_f = \frac{\frac{n_{MSDU}}{T_w} E[k] \, E[T_f]}{E[c]} \tag{7}$$

where $E[T_f]$ is the expected value of the duration of a single collision event, and E[c] is the expected value of the number of stations involved in a collision. In [41], an expression for $E[T_f]$ is derived assuming saturation traffic; unfortunately, as previously argumented, this assumption does not hold for the VoIP scenarios that we are considering. To overcome this limitation, we propose an alternative approximate formulation of $E[T_f]$, similar to the one in [40], which we deem suitable for the VoIP case. Let $E[T_{DATA}]$ be the sample mean duration of all DATA frames which are observed by the monitoring STA. $E[T_f]$, is approximated with:

$$E[T_f] \cong E[T_{DATA}] + T_{DIFS} \tag{8}$$

Analogously, an accurate expression for E[c] cannot be derived without the saturation assumption. On this matter, we note that the collision probability is much lower in presence of VoIP traffic than with saturation traffic, and therefore the probability of having more than two STAs involved in a collision is negligible. Hence, we choose to approximate E[c] with a value of 2.

We define our estimation of $\widehat{\rho_f}$ as follows:

$$\widehat{\rho_f} = \frac{\left(\frac{n_{MSDU}}{T_w} + \lambda_{new}\right) \widehat{E[k]} \, E[T_f]}{E[c]} \tag{9}$$

where $\widehat{E[k]}$ is calculated using the formulation of (6) but substituting P_f with the transmission attempt error probability $\widehat{P_f}$ that is expected after the new VoIP flows are introduced.

About the calculation of the fraction of time ρ_{bo} consumed by the backoff procedure, we focus on the backoff at the AP, since the downlink is known to be the bottleneck in a VoIP over WLAN scenario [30][30].

We assume that Immediate Access ([29], section 9.2.5.1) is unlikely to occur, which is reasonable when the wireless channel is close to the VoIP capacity, which is the region of operation where the DCAC algorithm needs most of the accuracy. With these assumptions, ρ_{bo} corresponds to the idle time necessary for the AP to

decrement the backoff counter prior to starting a transmission, and can be estimated as

$$\rho_{bo} = \frac{n_{MSDU_{DL}}}{T_w} T_{bo} \qquad (10)$$

In (10), $n_{MSDU_{DL}}$ represents the number of MSDUs transmitted in downlink during the observation period of duration T_w, and T_{bo} is defined as follows:

$$T_{bo} = \sigma \frac{CW_{min}}{2} \sum_{i=0}^{r_{max}-1} 2^i P_f^i \qquad (11)$$

In other words, T_{bo} represents the channel time spent in backoff for each MSDU transmitted by the AP. In (11), σ is the duration of a time slot according to the PHY specification being used.

Finally, the expected value of the fraction of time $\widehat{\rho_{bo}}$ that will be spent in backoff after the introduction of the new VoIP flow is calculated as

$$\widehat{\rho_{bo}} = \frac{(n_{MSDU_{DL}} + \lambda_{new})}{T_w} T_{bo} \qquad (12)$$

The calculation of $\widehat{P_f}$ and $\overline{\rho_{bg}}$ in (12) has been done experimentally using the test platform described in Section 4. For the details the reader is referred to [42].

3.3 Vertical Handover Execution

Three optimizations of the mid-call mobility procedure defined in [13] are proposed in the following. They are called hard, hybrid, and soft procedures. The main aim of those is to minimize the number of lost packets during the vertical handover. Looking at Figure 3 facilitates the comprehension of the procedures.

As stated in the previous section, no REGISTER message is sent to the SIP server during the mid-call procedure due to the previous registration of all the available interfaces and their corresponding IP addresses.

3.3.1 Hard Procedure

When the procedure is triggered, the MIMN sends a re-INVITE message, and subsequently, its data packets through the new interface. The old interface is closed performing a 'break-before-make' procedure. The CN continues sending its data traffic to the old interface of the MIMN until the reception of the re-INVITE message and the dispatch of the OK message.

3.3.2 Hybrid Procedure

Similarly to the previous case, when the handover execution is triggered, the MIMN sends a re-INVITE message, and subsequently, its data packets through the new interface; but in this case the old interface is kept open. CN continues sending its data packets to the old interface of the MIMN until the reception of the re-INVITE message and the dispatch of the OK message; then, it changes the destination IP address. MIMN closes the old interface when the OK message is received.

3.3.3 Soft Procedure

When the execution procedure is triggered, the MIMN sends a re-INVITE message through the new interface, but unlike the previous procedures, data packets still travel through the old interface, which is kept active. CN continues sending its data packets to the old interface until the reception of the re-INVITE message. Then, it changes the destination IP address. MIMN closes its old interface after the reception of the OK message from the CN. A 'make-before-break' procedure is carried out in this case.

4 Implementation of the Test Platform

4.1 Architecture of the test platform

The MIMN architecture is depicted in Figure 4, the diagram displays the hardware and the software within the node and the relationships between the components.

The SIP/VoIP client used is a modified version of Ekiga [25], an open source client for GNU/Linux. The necessary modifications to implement the vertical handover signaling and execution have been performed on the client engine to allow vertical mobility.

A WLAN packet sniffer and vertical handover decision algorithm calculator have also been developed. The WLAN packet sniffer computes the raw parameters such as the number of exchanged, retried and failed frames or the time in which the medium has been busy. The algorithm calculator gets the input reports from the sniffer

to determine the value of the decision algorithm. The calculator also needs information from the SIP/VoIP client, namely the employed codec and the active interface. With all the input information, the module selects the best interface to support the communication and triggers a handover request, if necessary.

It is important to highlight that the mobile node described here, and used on the realization of the experiments, has two WLAN interfaces, one used for active communication and one devoted solely to medium analysis.

The VM scheme performance is assessed within the scenario presented in Figure 5, built using the EXTREME Testbed® [15].

One PC is acting as multi-interface mobile node, other two as WLAN access point and SIP server, running SER [27], and a fourth as correspondent node. All the machines run a Fedora Linux OS with kernel 2.6.17.11. MIMN is equipped with a WLAN card (Atheros chipset) and a UMTS-HSDPA card (OPTION GT-MAX). Multihoming is handled by means of iproute2 Linux utility suite [28] and MUSA [16] has been used as UMTS-HSDPA network that permits a complete control on cellular network configuration parameters. Finally, a fifth node is acting as packet gateway of the UMTS network and a set of 16 interfering nodes are used to generate background traffic on the WLAN.

The traffic sources used are Ekiga [25] and MGEN [26]. In particular, three voice codecs are considered in our test scenarios: G711, G729 and G723. Table 1 summarizes the characteristics of such codecs.

Table 1: Voice Codecs packet level characteristics

Codec	Bitrate (kbps)	Packet inter-arrival time (msec)	RTP Packet size (bytes)
G711	64	20	60
G729	8	20	20
G723.1	6.3	30	24

4.2 Description of test scenarios

4.2.1 Vertical handover decision test scenarios

These scenarios only consider the WLAN part of the test platform, since the main aim of the tests is to evaluate the performance of the DCAC algorithm, when used to estimate WLAN channel load. The test environment is composed of one AP and several mobile nodes sending/receiving traffic to/from an external fixed node. The ns-3 simulator [43] has been also used to scale our performance evaluation to scenarios involving higher number of nodes than the ones supported by the testbed. The parameters of the simulator have been opportunely tuned to match the characteristics of the testbed; in particular, we verified that the results obtained for the experiments that we ran on EXTREME matched those obtained with the simulator for the same scenarios. For a detailed description of this tuning process, the reader is referred to [44].

Definition of the considered scenarios

The following four scenarios are used to evaluate the DCAC algorithm and also to experimentally obtain the values of $\hat{P}_f$ and $\overline{\rho_{bg}}$.

Homogeneous traffic scenario

This is the type of scenario which is considered in the most of prior works dealing with VoIP over WLAN [30]-[38].

For this scenario, noise-related channel errors are minimized by using RF cables in the testbed and locating wireless nodes close to the AP in the ns-3 simulations. For each STA, two synthetic voice flows are started, one from the STA to the AP and the other vice versa. In every experiment, all the nodes use the same fixed PHY rate and the same voice codec. We repeated several experiments varying the number of STAs (from 1 to N_{max}), the voice codec (G.711, G.723, G.729) and using different PHY rates belonging to different PHY specifications[1]. Note that N_{max} always exceeds the maximum number of users that an AP can support with an adequate voice quality. We used the extended E-Model ([22], [23], [23]) as method to evaluate the perceived voice quality by the user. It is assumed that a user is satisfied when $R > 70$, where R is the voice quality rating ([24]).

The following table summarizes the settings used in this scenario. For every particular setting, 20 inde-

[1] For the PHY specifications, we used the DSSS specification for the 2.4GHz ISM band (commonly referred to as 802.11b) and the OFDM specifications for the 5 GHz band (commonly referred to as 802.11a). We executed DSSS experiments both with the testbed and with the simulator, whereas 802.11a experiments were run on the simulator only. This was needed to cope with the high value of N_{max} required to go beyond the VoIP capacity when using 802.11a.

pendent repetitions of the same experiment have been run to get a satisfactory statistical confidence in the results.

Table 2. Homogeneous scenario settings

voice codec	PHY specification	PHY rate	N_{max}
G.711	DSSS 2.4 GHz	1 Mbps	6
G.711	DSSS 2.4 GHz	2 Mbps	8
G.711	OFDM 5 GHz	6 Mbps	21
G.711	OFDM 5 GHz	12 Mbps	36
G.723	DSSS 2.4 GHz	1 Mbps	12
G.723	DSSS 2.4 GHz	2 Mbps	19
G.723	OFDM 5 GHz	6 Mbps	45
G.723	OFDM 5 GHz	12 Mbps	95
G.729	DSSS 2.4 GHz	1 Mbps	9
G.729	DSSS 2.4 GHz	2 Mbps	12
G.729	OFDM 5 GHz	6 Mbps	38
G.729	OFDM 5 GHz	12 Mbps	55

Heterogeneous codec scenario

This scenario is identical to the previous scenario, except that different codecs are used within the same experiment. In particular, STAs number 1 to N_1 use a first type of codec, while STAs number $N_1 + 1$ to $N_1 + N_2$ use a second type of codec. Experiments are repeated for values of N_2 ranging from 0 to $N_{max} - N_1$. All experiments for this scenario were run on the EXTREME testbed using the DSSS PHY only. The following table summarizes the experimented settings with their most relevant parameters.

Table 3. Heterogeneous scenario settings

scenario	1st codec	2nd codec	PHY rate	N_1	N_{max}
A	G.711	G.729	1 Mbps	3	7
B	G.729	G.723	2 Mbps	8	14

Multi-rate scenario

The purpose of this scenario is to test the performance of the DCAC scheme in a situation in which nodes use different PHY rates to adapt to the differences in signal propagation that they perceive to and from the AP due to their different position. We note that this situation is potentially critical for the CAC scheme, (i) due the presence of hidden nodes and (ii) since the presence of data frames with different durations might potentially be detrimental for the estimation of the residual channel capacity.

For these scenarios, we ran all experiments using the ns-3 simulator with the OFDM PHY specification. In the two sets of experiments, N STAs are randomly placed in a circle of radius 60m (scenario 1) and 80m (scenario 2). We used a log distance propagation model with a propagation exponent of 3 and a reference loss of 46 dB at a distance of 1 m. All STAs use an ideal rate adaptation algorithm which selects the highest PHY rate that provides communications with a Bit Error Rate less than 10^{-5}. These settings result in topologies that have an average number of hidden nodes (i.e., STAs that are not seen by the STA running the DCAC algorithm) equal to 0.29 and 0.34, in the scenario 1 and 2 respectively. Several experiments were repeated varying N between 1 and 80. For every value of N, we ran 30 independent replication of the experiment.

TCP background traffic scenario

This scenario is similar to scenario 1, with the difference that, in addition to the STAs performing voice communication, N_{tcp} further STAs have been added which performed a TCP file transfer in the downlink direction. For this scenario we used the EDCA access method (commonly referred to as 802.11e) for traffic differentiation. In particular, we used the AC_VO and AC_BE access categories defined by the standard [29] for VoIP and TCP traffic, respectively. The use of EDCA is motivated by the fact that only traffic differentiation can make VoIP perform satisfactorily when TCP flows are present. Experiments for this scenario have been run repeating the same settings of the homogeneous scenario and adding $N_{tcp} = 1, 2, 3$, with 15 independent replications for every resulting combination. All experiments for this scenario were run using the ns-3 simulator.

4.2.2 Vertical handover execution test scenario

The objective of these tests is to assess the quality degradation introduced by the vertical handover execution process to real-time services, such as VoIP. Since only the impact of the handover procedure is to be measured, interfering nodes are disabled in the test platform.

The experiments consist in performing a VoIP call and changing the interface during the duration of the

call. The passages from WLAN to UMTS-HSDPA and from UMTS-HSDPA to WLAN are analyzed. Each call lasts for one minute and the handover is executed in the middle of the call.

Packet losses (in terms of call interruption time), one way delay and perceived voice quality are calculated to understand the causes of the degradation of the voice session during the interface switching in the three introduced procedures. The extended E-Model is used to estimate the perceived voice quality. This method is assumed as more accurate than E-Model [24] for the bursty packet losses occurring in our test scenario. The metrics are all calculated at IP level by parsing captured packet traces. Then, the effect of playout buffer is not considered to calculate the R-factor. Anyway, IP level delay obtained in our tests is always less than 10 msec for WLAN and between 40 and 80 msec for UMTS-HSDPA: the playout buffer presence can be considered as negligible in the calculation of the R-factor and then on the proposed analysis. It is worth noting that results are averaged over 50 repetitions of the same experiment before plotting the considered graph.

5 PERFORMANCE EVALUATION

5.1 Analysis of the DCAC algorithm

We ran both our DCAC algorithm and the TBIT scheme [38] in all the scenarios described in Section 4.2.1. The performed experiments consisted in evaluating the voice quality (measured using the R-factor metric defined within the E-Model [24]) obtained after the addition of a new voice call. As performance metric for the comparison, we introduce a parameter which describes the difference between the voice capacity, expressed as the maximum number n_{real} of user which can be accepted with a good quality (i.e. R>70) in a given scenario and the maximum number n_{alg} of users accepted by the algorithm being considered. It is expressed as follows:

$$\Delta = \frac{n_{real} - n_{alg}}{n_{real}} \qquad (13)$$

Δ assumes positive values when $n_{real} > n_{alg}$, i.e., the algorithm overestimates the network congestion and admits fewer users than the voice capacity, and it assumes negative values when $n_{alg} > n_{real}$, i.e., the algorithm underestimates the network congestion and admits more users than the voice capacity. The latter is much worse than the former, since all the users will experience a bad voice quality when the voice capacity is exceeded.

Table 4 compares the performance of DCAC and of TBIT for all the considered test scenarios. We can appreciate that our solution performs better in every studied case, most notably in the homogeneous scenario with high PHY rates, in Multirate scenarios and in presence of TCP background traffic.

Two examples of such cases are analyzed in the following. They are namely, the homogeneous scenario with the G711 codec at 12 Mbps (Figure 6) and TCP background scenario with the G711 codec at 2 Mbps (Figure 7). Figure 6 and 7 represent the fraction of blocked voice calls averaged over all the repetitions of the same experiment versus the number of voice users. n_{real} and n_{alg} of our solution are highlighted in the pictures.

In Figure 6, we can notice that, in our test scenarios TBIT always admits new flows regardless the number of voice users already present. This is due to the fact that TBIT does not consider the channel time spent in collisions. In the same scenario, our algorithm performs very close to the real case, thus estimating properly the congestion and taking the decision accordingly. TBIT, on the contrary, underestimates the congestion, thus causing a quality degradation of all the active voice communication.

Also in the case described in Figure 7, our solution is close to the real case, whereas TBIT completely misinterprets the network status. It is well known that, when EDCA is used, TCP occupies the capacity left by voice traffic (i.e. UDP). TBIT recognizes it as a congestion and therefore does not admit new voice calls. When most of the traffic in the network is voice, then TBIT underestimates the network congestion (as in Figure 7) and admits a high number of voice sessions, again, degrading the perceived quality of all the active users.

5.2 Analysis of the SIP-based handover execution procedures

Figure 8 presents the results of the measurement campaign when the MIMN switches its communication from UMTS-HSDPA to WLAN. Each graph represents the behavior of the R-factor vs. time for a given codec, just before and after the vertical handover execution instant. Uplink (UL) and downlink (DL) are separated in the graphs so as to analyze both directions in detail. Notice that, though we use the UL and DL terminology, in

fact, measurements are taken end-to-end, i.e., from MIMN to CN or vice versa, which includes the wireless link but also the wired portion of the network.

TABLE 4. Results of the comparison between DCAC and TBIT for every tested scenario

Scenario	Codec	Bitrate	DCAC	TBIT
Homogeneous	G711	1M	0	1
		2M	0	0.14
		6M	0.05	Δ_{err}
		12M	0.03	Δ_{err}
	G723	1M	0	0.18
		2M	0	0.13
		6M	0.03	Δ_{err}
		12M	0.07	Δ_{err}
	G729	1M	0.13	0.13
		2M	0	0.09
		6M	0.03	Δ_{err}
		12M	0.02	Δ_{err}
heterogeneous	Scenario A		0	0.25
	Scenario B		0	0.33
Multirate	Scenario1		-0.07	Δ_{err}
	Scenario 2		-0.19	Δ_{err}
TCP back-ground	G711	1M	0	Δ_{err}
		2M	0.17	Δ_{err}
		6M	0.18	Δ_{err}
		12M	0.24	Δ_{err}
	G723	1M	-0.29	Δ_{err}
		2M	0.08	Δ_{err}
		6M	0.15	Δ_{err}
		12M	0.21	Δ_{err}
	G729	1M	0	Δ_{err}
		2M	0.13	Δ_{err}
		6M	0.18	Δ_{err}
		12M	0.24	Δ_{err}

Δ_{err} is representing situations in which TBIT algorithm totally misbehaves, as detailed in the figures 6 and 7.

Since interfering nodes are disabled in this campaign, a satisfactory value of R-factor (R > 70) can be noticed under stationary conditions, i.e., before and after the change of the interface. In this testing scenario and according to the E-Model definition in [22] and [24], WLAN can provide better voice quality than UMTS-HSDPA due to its higher bitrate and its consequent lower transmission delay. The same reasoning applies for UMTS-HSDPA due to the different bitrates and delays in UL and DL. Moreover, the three codecs present different values of the R-factor due to their intrinsic design characteristics.

It is interesting to note that data packets in the UL start passing through the new interface in distinct instants depending on the interface switching procedure followed. In particular, the soft procedure is the last one in changing the interface, as may be foreseen by looking at its message chart in Figure 3 (i.e., looking at the instant in which RTP packets start being transmitted through the new interface of the MN). It is also worth highlighting that the one way delay observed in all performed test is very low (around 10 ms for WLAN and between 40 and 80 ms for UMTS-HSDPA). Consequently, packet loss, rather than delay, is the metric with more impact on the perceived voice quality, according to [23] and [22].

Let us consider the transition from one interface to the other in Figure 8. The UL flow does not present any problem for all the introduced procedures. A transient phase appears before reaching the new stable situation with a new value of the R-factor depending on the new network and codec.

On the contrary, the R-factor for the DL flow presents a breakdown (of around 0,5) before reaching the new stationary value of the R-factor, when considering the hard procedure. This breakdown is due to the number of lost packets during the transition in the hard procedure as shown in Table 5, which reports the average and the

standard deviation (between brackets) of time in which the call is interrupted due to the lack of reception of the transmitted packets (called interruption time) for all the repetitions of the experiment, for each codec and each execution procedure. On the other hand, call is not interrupted (i.e., zero lost packets) when the hybrid and soft procedures are considered (see Table 5), thus explaining the absence of a breakdown when these two procedures are applied. Graphs showing R-factor behavior during the switch from WLAN to UMTS-HSDPA interface are omitted due to space limitations. Nevertheless, also in this case several packets are lost when the hard procedure is applied, as confirmed by Table 5.

It is interesting to notice that the interruption time for the transition (i.e., re-INVITE-OK handshake) from WLAN to UMTS-HSDPA is higher than from UMTS-HSDPA to WLAN (see Table 5) when using the hard procedure. All packets from the CN are lost during the transition from one interface to the other, since the old interface of the MIMN has been closed just before the dispatch of the re-INVITE message (see the big cross in Figure 3a). UMTS-HSDPA presents higher delays than WLAN, and then the re-INVITE-OK handshake takes more time in such network, thus leading to a higher number of lost packets and consequently a higher call interruption time.

TABLE 5.Interruption time (ms) due to vertical handover

Codec	Handover Direction	Hard	Hybrid	Soft
G.711	W→U	76 (15.5)	0 (0)	0 (0)
	U→W	46.66 (16.32)	0 (0)	0 (0)
G.723	W→U	57.99 (7.74)	0 (0)	0 (0)
	U→W	45.99 (19.2)	0 (0)	0 (0)
G.729	W→U	65.34 (20.66)	0 (0)	0 (0)
	U→W	30.66 (16.68)	0 (0)	0 (0)

W=WLAN; U=UMTS. In brackets the standard deviation is reported

The above mentioned results are confirmed by the audio tests in our lab: a listener at the MIMN can hear a very short voice glitch when the hard procedure is applied. On the contrary, no interruption is perceived by the listener in the hybrid and soft procedures.

The problem of the hard procedure is intrinsic to the 'break-before-make' nature of the process: the old interface is closed before that the last packet with the IP address of the old interface reaches the MIMN. On the other hand, shutting down the old interface after the completion of the re-INVITE-OK handshake procedure minimizes the probability of losing packets sent by the CN to the IP address of the old interface, as occurs in the hybrid and soft procedures.

Though the three considered codecs implement packet loss concealment, they reveal a certain impact in the performance in the presence of bursty packet losses, e.g., the breakdown commented above. Then, the soft or hybrid procedures are appropriate solutions to handle seamless interface switching for voice services.

5.2.3 On the robustness of the procedures

The main aim of this sub-section is to study the robustness of the three procedures in order to understand the limits of the proposed solutions. This evaluation is done by analyzing two worst-case scenarios.

The first case analyzes the impact of losing a re-INVITE (or OK) message, for example, due to channel errors. In this case, the SIP client has to re-send the message. The hard procedure is the most affected by the time wasted during the re-INVITE-OK handshake. Its performance will be still worse than the results shown in the previous subsection. The other two procedures (hybrid and soft) are more robust, since the old interface data socket remains open until the OK message arrives and no packet is lost even in this case.

Under real conditions, the decision to switch to another interface has to be taken when the new network can provide a better quality to the active sessions, e.g., the new network has a higher available bitrate than the old one. If the two networks present very unbalanced bitrates, even the soft and hybrid procedure may present

packet losses if the re-INVITE-OK handshake time is faster than the arrival of the last packet sent by the CN to the IP address of the old interface. The in-lab wireless heterogeneous network has been appropriately set up to emulate the scenario presented above. Therefore, interface switching is performed from a very low bitrate to a high bitrate network. In particular, the transition from UMTS at 64 kbps to WLAN at 54 Mbps is analyzed.

In Table 6 the average and the standard deviation (in brackets) of the interruption time when using G729 and G723.1 codecs are reported. G711 has not been tested in this case, due to the low bandwidth provided (64Kbps) by cellular network.

TABLE 6.Interruption time (ms) due to vertical handover

Codec	Hard	Hybrid	Soft
G.723	99.99 (21.72)	3.99 (15.48)	2.01 (7.74)
G.729	109.34 (16.68)	1.34 (5.16)	1.34 (5.16)

In brackets the standard deviation is reported

As expected, this situation worsens the already bad performance of the hard procedure: higher re-INVITE-OK handshake delay makes lose a higher number of packets during the transition, as the procedure breaks the link before the instauration of the new connection. On the other hand, communication is interrupted for a very short period, when the soft and hybrid procedures are applied. In particular, the latter presents a lower interruption time because the CN changes the destination IP address just after the reception of the re-INVITE, whereas the former waits until the completion of the re-INVITE-OK handshake. Such results confirm the robustness of the hybrid and the soft procedures.

5.3 Overall System Performance Analysis

In this section the vertical handover process is analyzed in its entirety, i.e., considering the joint effects of handover decision and execution in the overall procedure.

Scenario in which users are communicating with G711 voice codec, WLAN data rate is fixed at 2 Mbps and UMTS bearer is at 512 kbps in uplink and 2 Mbps in downlink has been chosen as reference scenario for this analysis. Also, out of the three handover execution strategies presented, only the soft procedure is considered.

As for UMTS to WLAN handover, the experiment consists in decreasing the number of voice users in the 802.11 network, so that the MIMN, transmitting through UMTS and sensing the WLAN, can switch the interface of the communication. As for WLAN to UMTS handover, the experiment consists in increasing the number of voice users in the 802.11 network, so that the MIMN, transmitting and monitoring the WLAN, can switch to UMTS.

Let us define the vertical handover delay as the period between the entrance/exit of the call that damages/improves the quality experienced by the users in the considered 802.11 channel and the reception of the first packet through the new interface. The vertical handover process is analyzed from the perspective of the mobile node, which triggers the handover and it is divided in five components, namely: time to take the decision, time to send the INVITE message, time to receive the OK message, time to send the first packet through the new interface and time to receive the first packet through the new interface.

Results in the table below helps us in introducing two interesting considerations.

TABLE 7.Different components of the Vertical Handover delay

Delay component	UMTS to WLAN		WLAN to UMTS	
	Avg (ms)	Std (ms)	Avg (ms)	Std (ms)
Decision	937.22	671.68	1205.47	169.19
INVITE	949.27	671.84	1220.37	169.81
OK	978.54	671.63	1321.26	168.33
ACK	986.44	672.15	1324.29	168.59
First packet received	1002.33	674.32	1334.21	170.41
First packet sent	1009.13	675.45	1343.37	175.09

First, handover decision is the component with the biggest influence on the overall vertical handover procedure. In fact, the decision process takes around 90% of the total vertical handover time for every studied scenario. The main issue is the duration of the monitoring window to have a consistent decision to avoid false handover initiation, which can lead to a ping-pong effect ([45]). After empirical studies, the monitoring win-

dow has been set at 1 second, which is the minimum time in which the number of false decisions is negligible in our setup. The ergodicity of the defined decision process can be utilized to optimize the duration of the monitoring window. Though, such optimization may be able to decrease the time dedicated to take the handover decision, it will remains as the critical part of the entire vertical handover management.

Second, CSMA/CA used by 802.11 WLAN is not well designed for voice traffic. WLAN offers higher bandwidth than 3G, even in a situation in which users do not have a satisfying voice quality. In fact, the SIP messages needed to switch from UMTS to WLAN are exchanged more quickly than in the opposite direction (i.e., handover from UMTS to WLAN is quicker than from WLAN to UMTS). In such situation WLAN is far from a congested situation, but it cannot serve more voice users with adequate quality due to the overhead introduced by its random medium access protocol, which is not optimized for such type of traffic (very small packet size). In this respect, it is also interesting to notice that the standard deviation of the obtained delay components is lower in UMTS, due to its circuit-oriented bearer management, which provides a more stable bandwidth allocated for a single user. This result confirms the benefit of using circuit-oriented access for voice services to have better bandwidth usage and user experienced quality.

CONCLUSIONS

In this paper we have presented an approach based on the introduction of an overlay network to provide seamless terminal mobility in multi-access wireless networks. The proposed architecture does not require any modification to the already existing wireless access networks. Moreover, it offers the necessary flexibility to allow interoperability in a multiple operator environment, as well as to easily include future access technologies and networks in the heterogeneous system.

SIP protocol has been selected as the signaling protocol of the overlay network, since it provides a natural way to associate high-level identifiers and interface identifiers and it requires no modification to standard Internet protocols. Additionally, SIP enables distributed mobility management, where mobile nodes can decide which network to use and initiate the handover procedure. Moreover, SIP also allows the wireless access networks to trigger handover execution, providing a really flexible and scalable solution to vertical mobility. In this paper, we have studied mainly the distributed approach.

The work has analyzed the three components of the vertical mobility management: node localization, handover decision and handover execution. As confirmed by the experimental tests and simulations, our scheme allows resilient mobile node reachability, preservation of the radio resources by limiting the unnecessary signaling through wireless links, reduction of packet losses during the vertical handover, and accurate distributed vertical handover decision, based on the estimation of the available resources of the candidate access networks.

ACKNOWLEDGEMENTS

This work has been made possible through joint collaboration with Cisco Advanced Architecture & Research Group.

REFERENCES

[1] E. Gustafsson, A. Jonsson, “Always Best Connected”, IEEE Wireless Communications, Issue 1, Feb. 2003
[2] D. Calin, H. Claussen, H. Uzunalioglu, On femto deployment architectures and macrocell offloading benefits in joint macro-femto deployments, IEEE Communications Magazine Volume: 48 , Issue: 1
[3] S. M.Cheng; S.Y. Lien; F.S. Chu; K.C. Chen, On exploiting cognitive radio to mitigate interference in macro/femto heterogeneous networks, IEEE Wireless Communications, Volume: 18, Issue: 3
[4] O. Tipmongkolsilp, S. Zaghloul, A. Jukan, The Evolution of Cellular Backhaul Technologies: Current Issues and Future Trends, IEEE Communications Surveys & Tutorials, Volume: 13 , Issue: 1
[5] G. Camarillo, M. A. Garcia Martin, “The 3G IP Multimedia Subsystem (IMS)” WILEY ed., 2008
[6] IETF MEXT webpage: http://www.ietf.org/html.charters/mext-charter.html
[7] S. Guha, Y. Takeda, P.Francis NUTSS: A SIP-based Approach to UDP and TCP Network Connectivity, in: Proceeding FDNA '04 workshop on Future directions in network architecture of the ACM SIGCOMM
[8] P. Bellavista, A. Corradi, L. Foschini “IMS-compliant Management of Vertical Handoffs for Mobile Multimedia Session Continuity”, IEEE Communications Magazine, April 2010.
[9] A: Dutta, K. Manousaki, S. Das, F. J. Lin, T. Chiba, H. Yokota, A. Idoue, H. Schulzrinne, “Mobility Testbed for 3GPP2-based Multimedia Domain Networks”, IEEE Communications Magazine, July 2007
[10] R. Farha. K. Khavari, N. Abji, A. Leon-Garcia “Peer-to-peer Mobility Management for all-IP Networks”, in proc. of IEEE ICC 2006
[11] R. Farha. K. Khavari, A. Leon-Garcia, “Peer-to-peer Vertical Mobility Management” in proc. of IEEE ICC 2007
[12] J. Rosenberg, H. Schulzrinne,G. Camarillo, A. Johnston et al., SIP: Session Initiation Protocol, RFC 3261, June 2002

[13] A. Johnston, S. Donovan, R. Sparks, C. Cunningham,K. Summers, Session Initiation Protocol (SIP) Basic Call Flow Examples, RFC 3665, December 2003
[14] R. Wakikawa, V. Devarapalli, T. Ernst, K. Nagami, Multiple Care-of Addresses Registration, Internet Draft, March 2009
[15] M. Portoles-Comeras, M. Requena-Esteso, J. Mangues-Bafalluy, M. Cardenete-Suriol, "EXTREME: Combining the ease of management of multi-user experimental facilities and the flexibility of proof of concept testbeds", in proc. IEEE TridentCom 2006
[16] P. Dini, M. Portoles-Comeras, J. Mangues-Bafalluy, L. Dai, S. Addepalli, A Real-Time Cellular System Architecture to Experiment with UMTS/HSDPA in a laboratory, in Proc. IEEE TridentCom 2009, April 2009
[17] S. Salsano, A. Polidoro, C. Mingardi, S. Niccolini, L. Veltri, SIP-based Mobility Management in Next Generation Networks, IEEE Wireless Communications, April 2008
[18] W. Wu, N. Banerjee, K. Basu, K. Das, SIP-based Vertical Handoff between WWANs and WLANs, IEEE Wireless Communications, June 2008
[19] N. Banerjee, S. K. Das, A. Acharya, SIP-based Architecture for next generation Wireless Networks, in Proc. IEEE PerComm 2005
[20] A. Dutta, Y. Ling, W. Chen, J. Chennikara, Multimedia SIP Sessions in a Mobile Heterogeneous Access Environment, in Proc. 3G Wireless 2002
[21] D.D. Nguyen, Y. Xia, M.N. Son, C.K. Yeo, B.S. Lee, A Mobility Management Scheme with QoS Support for Heterogeneous Multi-homed Mobile Nodes, in Proc. IEEE GlobeCom 2008, New Orleans, Novemeber/December 2008
[22] A. Clark, Extended E-model T1A1.1/2001-037 Extensions to the E Model to incorporate the effects of time varying packet loss and recency April 30, 2001
[23] Ie values ITU-T Rec. G.113 (11/2007) Transmission impairments due to speech processing
[24] E-model ITU-T Rec. G.107 (03/2005) The E-model, a computational model for use in transmission planning.
[25] Ekiga application webpage: http://www.gnomemeeting.org
[26] Mgen: Multi-Generator webpage: http://cs.itd.nrl.navy.mil/work/mgen/index.php
[27] SIP Express Router webpage: http://www.iptel.org/ser
[28] IPROUTE2 Utility Suite webpage: http://www.policyrouting.org/iproute2-toc.htm
[29] IEEE Std. 802.11-2007, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications", 2007.
[30] S. Garg, M. Kappes "Can I add a VoIP Call" in proc. IEEE ICC 2003, Seattle, USA, May 2003
[31] S. Garg, M. Kappes "Admission Control for VoIP Traffic in IEEE 802.11 Networks" in Proc. of IEEE GLOBECOM, 2003
[32] H. Zhai, X. Chen, Y. Fang, "How Well Can the IEEE 802.11 Wireless LAN Support Quality of Service?", IEEE Trans. On Wireless Comm., vol.4, n.6, November 2005.
[33] X. Yan, Y. A. Şekercioğlu, S. Narayanan "A Survey of Vertical Handover Decision Algorithms in Fourth Generation Heterogeneous Wireless Networks", *Computer Networks*, Vol. 54, No. 11. (02 August 2010)
[34] E. Van der Berg, et al. "Dynamic Network Selection using Kernels" in proc. IEEE ICC 2007
[35] J. Sachs et al. "Generic Abstraction of Access Performance and Resources for Multi-Radio Access Management", in proc. IST In Mobile and Wireless Communications Summit, 2007
[36] P. McGovern, S. Chung, S. Murphy, L. Murphy "Endpoint Admission Control for VoIPoWLAN" in Proc. ICT 2006, May 2006
[37] I.D. Chakeres, E. M. Belding-Royer "PAC: Perceptive Admission Control for Mobile Wireless Networks" in Proc. QSHINE 2004, Washingthon D.C., USA, 2004
[38] K. Yasukawa, A. G. Forte, H. Schulzrinne "Distributed Delay Estimation and Call Admission Control in IEEE 802.11 WLANs", in Proc. of IEEE ICC, June 2009
[39] B. R. Tamma, N. Baldo, B.S. Manoj, R. Rao, "Multi-Channel Wireless Traffic Sensing and Characterization for Cognitive Networking", in Proc. of IEEE ICC, June 2009
[40] G. Bianchi "Performance Analysis of the IEEE 802.11 Distributed Coordination Function" IEEE Journal Selected Areas in Communications, 18(3), March 2000.
[41] N. Baldo, A. Zanella, "A Game Theoretic evaluation of Rate Adaptation strategies for IEEE 802.11 based Wireless LANs", in Proc. of ICST GameComm, October 2009.
[42] P. Dini, N. Baldo, J. Nin Guerrero, J. Mangues, S. Addepalli, L. Dai, Distributed Call Admission Control for VoIP over 802.11 WLANs based on Channel Load Estimation, in Proceedings of the IEEE International Conference on Communications (ICC-2010), 23-27 May 2010, Cape Town (South Africa).
[43] The ns-3 network simulator [online] available at: http://www.nsnam.org
[44] N. Baldo, M. Requena, J. Nuñez, M. Portolès, J. Nin Guerrero, P. Dini, J. Mangues, "Validation of the IEEE 802.11 MAC model in the ns3 simulator using the EXTREME testbed", in Proc. of Simutools, March 2010.
[45] W.I. Kim, et al., "Ping-Pong Avoidance Algorithm for Vertical Handover in Wireless Overlay Networks," in *Proc. of IEEE 66-th Vehicular Technology Conference,* VTC-2007, pp.1509-1512, Sept. 30 2007-Oct. 3 2007.
[46] A. Dutta, B. Kim, T. Zhang, S. Baba, K. Taniuchi, Y. Ohba, Experimental analysis of multi interface mobility management with SIP and MIP, in proc. of Wireless Networks, Communications and Mobile Computing, 2005 International Conference, 13-16 June 2005, page(s): 1301 - 1306 vol.2 Print ISBN: 0-7803-9305-8
[47] A. Dutta, S. Madhani, W. Chen, O. Altintas, H. Schulzrinne, Fast-handoff schemes for application layer mobility management, in proc. of IEEE PIMRC 2004
[48] IEEE Standard for Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications
[49] Ken Duffy, David Malone, and Douglas J. Leith "Modeling the 802.11 Distributed Coordination Function in Non-Saturated Conditions" IEEE COMMUNICATIONS LETTERS, VOL. 9, NO. 8, AUGUST 2005

Authors' Biography

Paolo Dini received the "laurea" degree in Electronic Engineering and the PhD degree in Telecommunications in 2001 and 2005 respectively, both from the Università di Roma "La Sapienza". From 2001 to 2005 he worked as a research assistant in the Computer Science and Systems (DIS) and Information and Communication Theory (INFOCOM) departments of Università di Roma "La Sapienza". In 2005 he worked with the Research Center on Software Technologies (RCOST) as a research assistant. In February 2006 he joined IP Technologies Area of CTTC as a post doc researcher. Since November 2006 he is a Research Associate in the IP Technologies Area. He was granted twice by the Cisco University Research Program for his research activities in 2009 and 2011. His current research interests are in the field of mobility and radio resource management for mobile and wireless networks and green networking.

Jaume Nin Guerrero obtained his Telecommunications Engineering degree from the Technical University of Catalonia (UPC) on 2007. He wrote this Master Thesis as guest student at Aalborg Universitet (AAU) in collaboration with Nokia Siemens Networks Denmark.
Before joining the IP Technologies group of the CTTC in October 2008 as Network Software Engineer, he worked as Software Engineer at Abiquo. His research interests are on mobility management for wireless networks and testbed deployments.

Nicola Baldo received his Laurea (BE) and Laurea Specialistica (ME) degree in Telecommunications Engineering in 2003 and 2005, respectively, from the University of Ferrara, Italy, and the PhD in Information Engineering from the University of Padova, Italy, in 2009. In summer 2003 he was an internship student at the Ericsson Eurolab Deutschland, Aachen, Germany. In 2005 he was on leave at the STMicroelectronics Advanced System Technology group, Agrate Brianza (MI), Italy. In 2008 he was on leave at the Calit2 department, University of California, San Diego, USA. Since February 2009 he is with the CTTC. His research interests include Cognitive Radio and Networks, Cross-layer Optimization, Multimedia Communications and Network Simulation Tools. He is a member of the IEEE and the IEEE Computer Society.

Figures

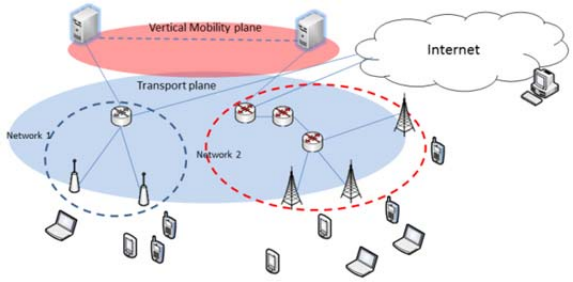


Fig.1 System level view of the proposed inter-technology architecture. Two networks composing a possible example of multi-access environment are shown.

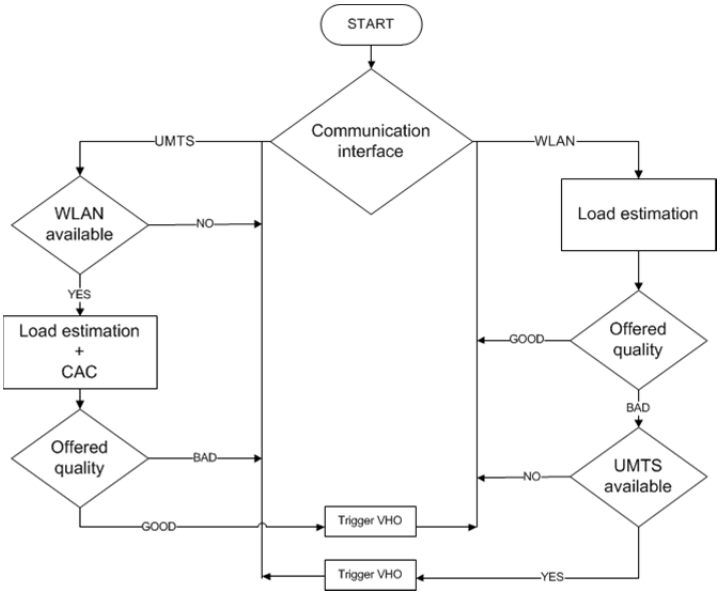


Fig.2 Flow chart describing the vertical handover decision process.

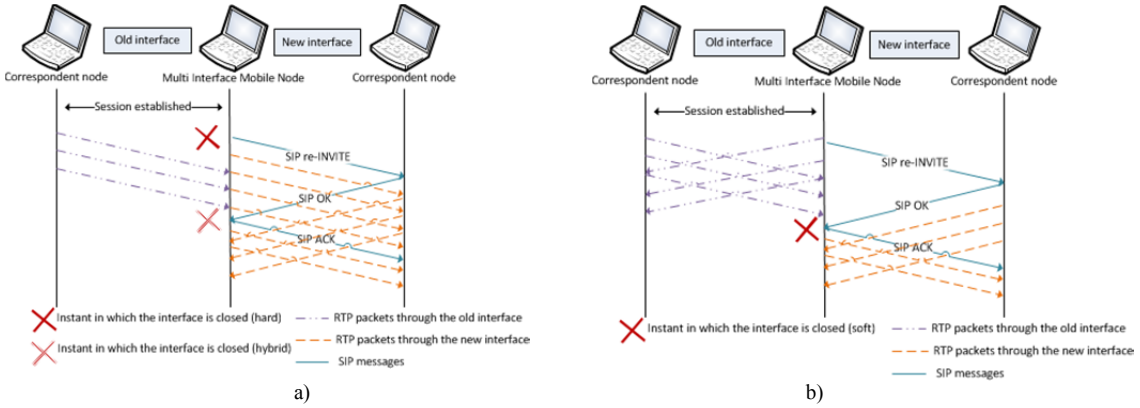


Fig.3. Message sequence charts of the three proposed vertical handover procedures. 3a represents the hard and the hybrid procedure; 3b represents the soft procedure.

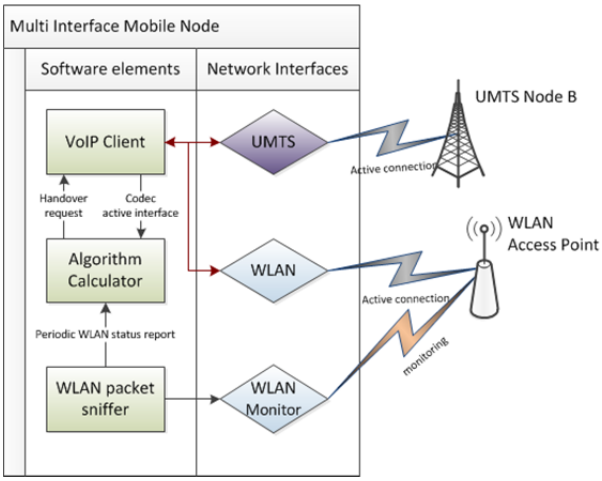


Fig.4. Multi Interface Mobile Node internal architecture

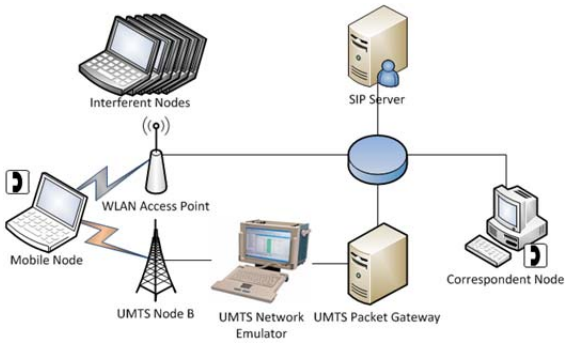


Fig.5. Test platform used in the experiments

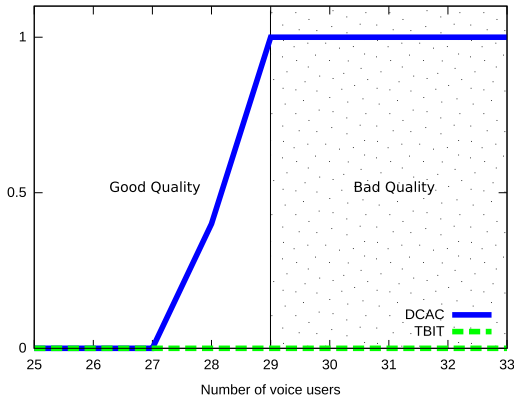


Fig.6. Fraction of blocked users for G711 voice codec at 12 Mbps

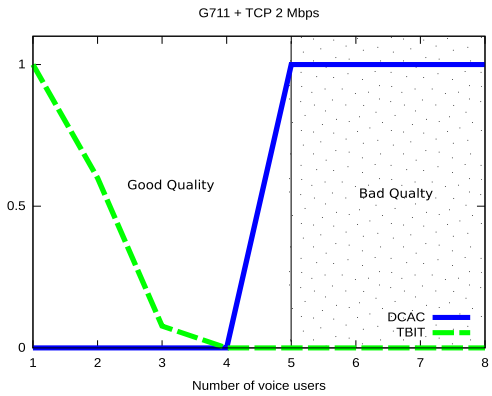
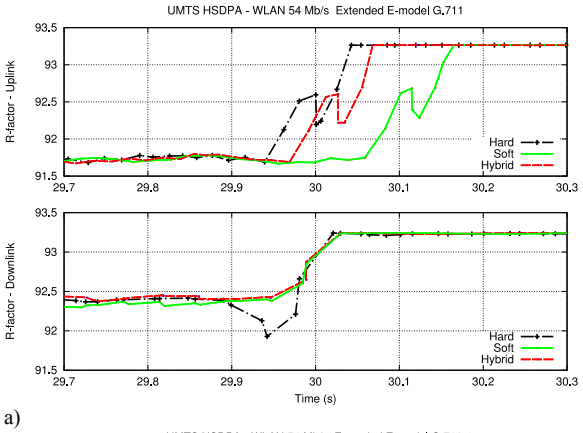
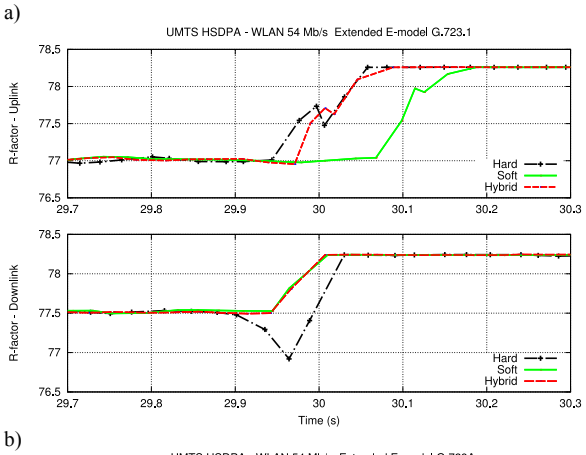


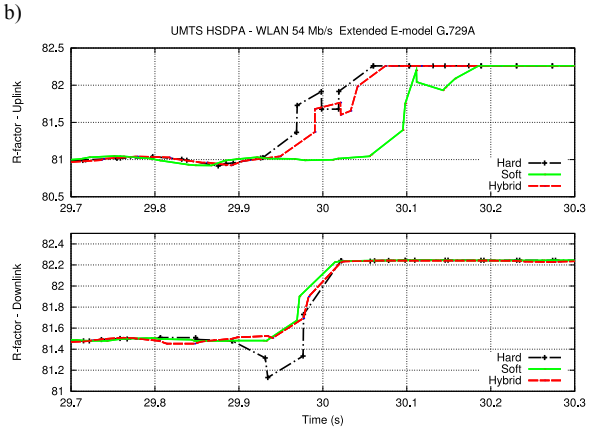
Fig.7. Fraction of blocked users for G711 plus 1 TCP connection at 2 Mbps



a)



b)



c)

Fig.8. R-factor calculated for the transition from UMTS-HSDPA to WLAN interface of the G711, G729, G723.1 codecs and for each of the three considered procedures (hard, hybrid and soft)