A solution for transparent mobility with route optimization in the IP multimedia subsystem

Ivan Vidal *, Jaime Garcia-Reinoso, Ignacio Soto, Antonio de la Oliva

Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 Leganés, Madrid, Spain

Abstract: This paper presents TRIM+, an architecture for transparent mobility management with route optimiza tion in IMS based networks. The design of our architecture is based on a previous work referred to as TRIM. TRIM was originally devised to provide transparent mobility support in the IMS, although transpar ency came at the cost of using a suboptimal data path between communicating end points. TRIM+ main tains transparency as a design criterium, and thus end user applications, running at the mobile node and its correspondent communication peers, are unaware of mobility management procedures. Additionally, the proposed design defines a set of route optimization procedures, allowing compliant devices to use the optimal data path for media communications. Furthermore, TRIM+ addresses packet loss management in scenarios where the media path cannot be maintained during the handover of the MN. To this end, our architecture enables the MN to request buffering capacity in its home network to temporarily store incoming media traffic during the handover, which would otherwise be dropped. This mechanism, as well as route optimization procedures, are executed transparently to the end user applications running at the communicating end points. As a proof of concept, we have implemented a software prototype of the TRIM+ architecture, deploying it over a real IMS testbed. By means of a set of experiments, we have val idated the mechanisms proposed in this paper, considering both UDP and TCP user traffic.

Keywords: IMS, Transparent mobility, Handover management, Route optimization

1. Introduction

Convergence and mobility are the two main trends in the cur rent evolution of communication networks. Convergence means that services traditionally offered through circuit switched net works, such as telephony, are being moved to packet based net works, specifically to IP based networks. Telecommunication operators see the IP Multimedia Subsystem (IMS) [1] as the key enabler to support convergence, providing access control and ses sion signaling, functionalities that are needed to offer traditional services of circuit switched networks through IP based networks.

Nevertheless, providing mobility on these converged IP based networks is not a trivial task. Although there are several solutions developed to enable mobility support at the IP layer, their integra tion in an IMS network poses several challenges that have to be considered. Other solutions, which do not have an impact in the IMS, are also available, but they have some limitations such as offering mobility support only within an administrative domain or within a restricted geographical scope, or affecting the applica tions that have to be adapted to deal with changes produced by mobility. In Section 2, we give an overview of the different ap proaches for mobility support in IP networks with IMS, and their limitations.

In order to address the challenges of providing IP mobility in an IMS based network, in [2] we proposed TRIM, a solution to support mobility in IP networks with IMS, making mobility transparent to applications and to communication peers of mobile nodes. TRIM overcomes the limitations of previous alternatives for mobility support in IP networks with IMS. Nevertheless, this solution is lim ited on performance, because transparency comes with the cost of utilizing a suboptimal path to exchange data between end user applications.

This paper presents TRIM+, an architecture for transparent mobility management with route optimization in IMS based net works. As in TRIM, our solution can be easily integrated into the IMS infrastructure, and preserves transparency as a design crite rium. However, TRIM+ defines a set of route optimization proce dures, which allow configuring the optimal data path for the exchange of media between compliant user equipment. In addi tion, TRIM+ addresses packet loss management in scenarios where the media path cannot be maintained during the handover of the MN (hard handovers). Finally, TRIM+ is compatible with legacy ter minals, so a TRIM+ terminal can communicate with a standard 3GPP terminal or other terminal without TRIM+ functionality,

* Corresponding author.

E-mail addresses: ividal@it.uc3m.es (I. Vidal), jgr@it.uc3m.es (J. Garcia-Reinoso), isoto@it.uc3m.es (I. Soto), aoliva@it.uc3m.es (A. de la Oliva).

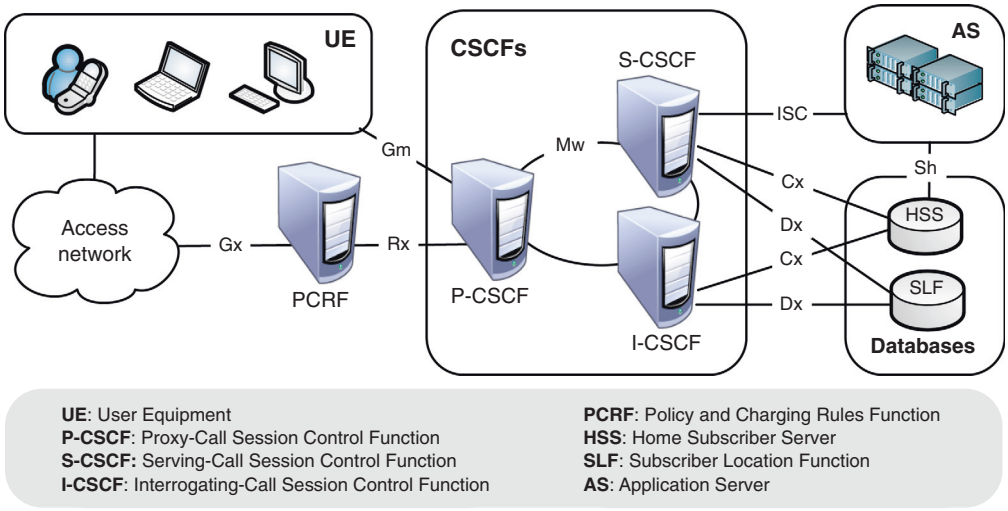

Fig. 1. Simplified overview of the IMS reference architecture.

and still have mobility support. The architecture and operation of TRIM+ is detailed in Section 3. A thorough evaluation of the behav ior of TRIM+ considering both TCP and UDP traffic, is presented in Section 4, where we also compared its performance with that of the 3GPP IMS Service Continuity. Finally, in Section 5 we summa rize the conclusions of our work.

2. Mobility management in the IMS

Communication networks are converging towards packet switched IP based technologies. To be able to accommodate tradi tional voice and video services over a packet network, operators have pushed the introduction of the IP Multimedia Subsystem (IMS) that provides access control and session control, according to a specific profile for the Session Initiation Protocol (SIP) [3] and the Session Description Protocol (SDP) [4] defined by the 3GPP in [5]. IMS was initially proposed by the 3GPP,[1] but currently it is being developed in conjunction with ETSI TISPAN to ensure that IMS works with any type of access network. A simplified view of the IMS architecture is presented in Fig. 1.

Mobility support is also an essential feature in communication networks nowadays. Users expect network access everywhere and while moving. The IETF[2] has developed several solutions for mobility support based on the IP layer of the protocol stack, being significant examples: Mobile IPv4 [6] and Mobile IPv6 [7]. These two protocols account for the traditional approaches for solving the general problem of IP mobility in the Internet. Mobile IP uses two IP addresses, one that is permanent and does not change with mobility, and another one that is temporal and changes for each vis ited access network. The permanent address, called Home Address, is the one seen by applications, while the temporal address, called Care of Address, is used to direct traffic to the current location of the MN. Mobile IP has been in place for a long time and it seems a good approach to solve the problem of mobility management in the IMS. Unfortunately the integration of Mobile IP and related solu tions with IMS is not straightforward due to the two different ad dresses that Mobile IP manages but that IMS does not expect. The challenges in the integration are explained in detail in [8 11], and summarized next. IMS performs resource reservation and authoriza tion within the network (part of the policy and charging control architecture), based on firewall like rules which allow flows in the network if they match a certain template. This template is based on the IP addresses used by the flow that, in the case of using Mobile IP, in the terminal side correspond to the Care of Address. The filter ing rules are setup during the signaling exchange between the IMS core and the user terminal to request the session. The problem ap pears since the IP address used by the control application at the ter minal (SIP based) can only see the Home Address assigned to it, being unaware of the temporal Care of address the terminal is cur rently using. Hence there is a mismatch between what the control signaling indicates and the actual IP addresses used. This mismatch cannot be solved without modifications in IMS and/or the policy and charging control architecture of 3GPP to make it aware of Mobile IP.

Therefore several works can be found in the literature trying to integrate different architectures through the use of modified Mo bile IP and IMS approaches. Examples of these works can be found in [10,12 15]. Basically all of these approaches which try to tightly integrate the different WiMAX/WLAN/UMTS architectures, relay on Mobile IP mechanisms for managing the mobility between het erogeneous accesses. Some of these works just ignore the problems identified above for the interworking between IMS and Mobile IP, while others provide a variety of solutions, from the modification of the actual signaling used for the handover, the inclusion of new features in Mobile IP or even using some other specifications such as IEEE 802.21 to perform hybrid signaling mechanisms. Sum marizing, all these approaches require modifications to standard protocols and procedures, in particular in the IMS infrastructure, in order to provide Mobile IP based mobility with IMS to the user.

There are some other works that look at the IP mobility man agement issue through a different perspective. These works under stand mobility as a particular case of IP multi homing, in which the terminal uses several IP addresses in one or several interfaces and is able to control which IP address is used for each flow. Examples of these approaches can be found using the Shim6 [16] and Mobile SCTP [17] protocols. In [18], authors try to provide a framework for service continuity for IMS based networks using the Shim6 layer to provide IP mobility support. Although these approach has some benefits as the lack of modifications to the network side and mobil ity transparency at the IMS layer, it has a major drawback, since it requires all UEs to implement the Shim6 protocol and communica tion with legacy nodes is not possible.

Another work, which shares the same core ideas with the previ ous one, can be found in [19]. This work uses mSCTP instead of

[1] The 3rd generation partnership project: http://www.3gpp.org.
[2] Internet engineering task force: http://www.ietf.org.

Shim6 to provide the IP agility function, but in order to support legacy nodes, it relies on proxy mSCTP nodes scattered through the network to translate between mSCTP and standard TCP. These proxies create suboptimal paths in the communications and, although minimal, the solution requires modifications to the IMS.

Another option is using mobility support based in functionality in the network (localized network based mobility), in such a way that the Mobile Node (MN) can move in the network without changing its IP address. This is the common approach in UMTS net works that use GTP [20] to hide mobility from the IP layer of MNs while they move within the same UMTS network. PMIPv6 [21] is conceptually a similar approach defined by the IETF, that also hides mobility from the IP layer of MNs while they move within a certain part of the network that uses PMIPv6 to offer localized network based mobility support. These solutions are restricted to localized parts of the network and do not allow moving among networks with different administrative management, a case of increasing importance with the current evolution in the use of communica tion networks.

Finally, as IMS is based on SIP, we can use SIP based mobility approaches, such as the ones presented in [22 25]. The idea is to manage at the application layer the change of IP address caused by the movement of the MN. SIP provides the identification of the end point of a communication, using a SIP Uniform Resource Identifier (URI), and the means to locate it (to find out the associ ated IP address). If there are changes of IP address due to move ments of the MN, the association of the URI and the new IP address is updated through SIP signaling. This is the basis of the 3GPP Service Continuity [26,27], which defines how to support movement of MNs between packet switched access networks using IMS. Note that the 3GPP Service Continuity also covers other sce narios that we do not cover in this paper, namely the movement between circuit switched access networks and packet access net works with IMS, and the transfer of sessions between different ter minals. With the SIP based mobility approach, mobility is not transparent to applications. SIP allows the change of IP address due to mobility by keeping the end points of the communication informed of the IP addresses currently in use, but the changes of address are visible to applications, in both sides of the communica tion, that have to deal with it. This is unfortunate because it means that a communication based application software that works per fectly fine when both end points of the communication are fixed nodes, can nevertheless fail when one, or both, of the end points of the communication are mobile terminals. TCP applications will definitely be affected and will not work without having special purpose code to close the old and open a new TCP connection after the movement. Depending on how they are programmed UDP applications are also likely to be affected needing to reconfigure or open new network sockets.

In [2], we proposed TRIM, a solution for providing transparent mobility in IMS based networks using SIP. TRIM uses a SIP applica tion server (AS) to make mobility transparent to the signaling plane of the communication peer of a MN (a node we call corre spondent node or CN). Two SIP sessions are created when a MN communicates with other node (the CN), one between the MN and the TRIM AS and another between the TRIM AS and the CN. If the MN moves, only the first signaling leg has to be updated with the new IP address of the MN, the second signaling leg between the TRIM AS and the CN is kept without modifications. But with this arrangement mobility still affects the data plane of both the MN and the CN. In the TRIM proposal, a Multimedia Resource Function (MRF) is included in the data path of the communication. So the data path leg between the CN and the TRIM MRF never changes, packets are forwarded between the CN and the TRIM MRF during the lifetime of the session. Only the data plane leg between the TRIM MRF and the MN changes after a MN movement. SIP signaling is used to reconfigure the TRIM MRF to forward the traffic received from the CN to the new IP address of the MN, and vice ver sa, the traffic coming from the new IP address of the MN is received in the TRIM MRF and is forwarded to the CN. Address translations are performed in the MN, so upper layers are kept unaware of modifications in the IP address used in the link, upper layers al ways see an internal unchanged address; and also in the MRF so packets that are received in an address of the MRF, are sent to the right address of the CN or MN. This solution provides transpar ent mobility to the CN (both in the signaling and in the data plane) and to the transport layer and applications in the MN that are kept unaware of the mobility and changes of IP address in the MN. The limitation of TRIM is that the data path is anchored in the home network of the MN, i.e., in the network of the operator of the MN that is where the MRF resides. This leads to suboptimal data paths that can result in a significant increase of data path delay and the corresponding loss of performance. In [28], we explore the poten tial of TRIM to support flow mobility, where handover decisions are governed by policy rules defined by operators, but this proposal still uses suboptimal paths for media communications.

3. The TRIM+ architecture

This section describes TRIM+, an architecture that provides transparent mobility support with route optimization in IMS based networks. The design of our architecture evolves from a previous work, i.e., TRIM [2]. As we have described in the previous section, TRIM was originally devised to provide transparent mobility in the IMS, although this transparency requires the use of suboptimal paths for the media exchange. TRIM+ maintains transparency as a primary design criterion, although it additionally provides the fol lowing core functionalities:

- During a session setup, our architecture allows each involved party to determine if TRIM+ is supported by the other end of the SIP communication. This information can then be used to trigger route optimization procedures. The different use cases corresponding to session setup, where a CN could be either a TRIM+ UE (mobile or fixed) and a UE not supporting TRIM+ (i.e., a legacy terminal), are covered Section 3.2.
- If TRIM+ is supported by both ends of the communication, our proposal allows configuring the optimal data path for the exchange of media between the communicating endpoints, avoiding data forwarding through intermediate network enti ties. Route optimization procedures are described in Section 3.3.
- Handover management is always transparent to end user appli cations, running at the MN and the CN, even if the media exchange takes place using the optimal data path. The different use cases of handover management are detailed in Section 3.4.
- TRIM+ can address packet loss in case that the media path can not be maintained during the handover of the MN. To this end, our solution allows the MN to request buffering capacity from its home network, to momentarily store incoming media during the handover. This solution is described in Section 3.5.
- The case where a MN communicates with a CN connected to a fixed access network is expected to become a common scenario under an IMS deployment as it is nowadays in the Internet (e.g. a mobile user utilizing a handheld to access a video streaming service). Our prior proposal (i.e., TRIM) supported mobility management transparently to the CN in this case, by anchoring the media session at a network element located at the MN home network. TRIM+ also preserves transparency in this scenario. However, if the fixed CN implements the enhancements pro posed in this paper, route optimization is also possible, avoiding the suboptimal route through the home network of the MN.

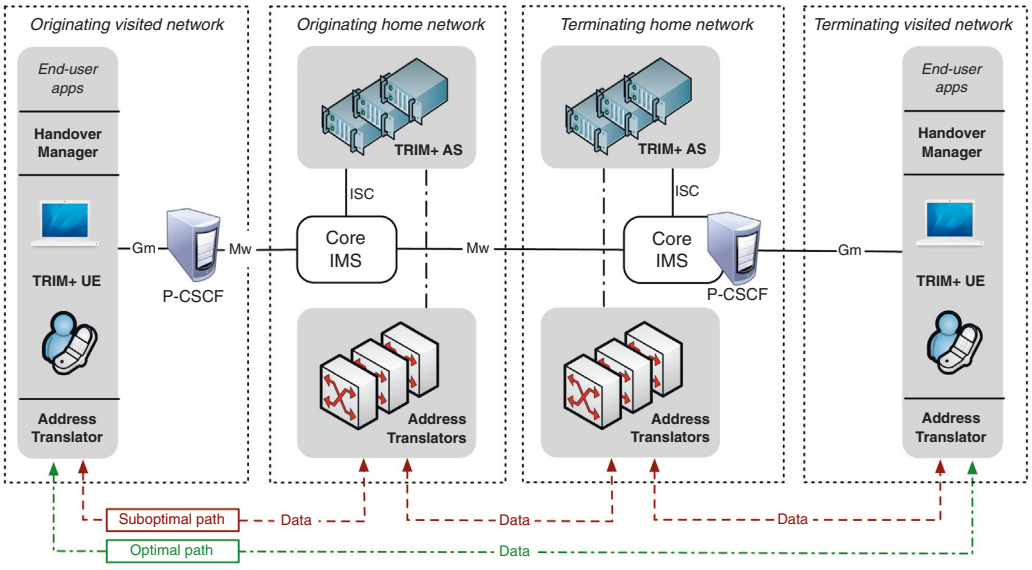

Fig. 2. Overview of the TRIM+ architecture.

In the rest of this work, we use the following terminology: MN refers to an end user device that is mobile, UE refers to an end user device that can be either mobile or fixed (therefore, a MN is a UE), and CN denotes a UE acting as a correspondent peer in a communication.

3.1. Overview of the TRIM+ architecture

Fig. 2 outlines the main components of the TRIM+ architecture. The figure represents two TRIM+ enabled UEs, where an originating UE has established an IMS session with a terminating UE. Both UEs have acquired network connectivity in a visited network, and are accessing the IMS by means of a P CSCF. The figure covers both scenarios where a UE can use a P CSCF in the visited and in the home network, as both of them are supported in our proposal.

The key elements in the TRIM+ architecture are the TRIM+ Application Server (AS), the address translator in the home network, and the address translator in the compliant UE. The TRIM+ AS is the anchor point for supporting the mobility of a MN. The TRIM+ AS is located in the home network of the MN (its operator's network) and it is always kept in the signaling path of the MN. A communication between a TRIM+ MN and other node (the correspondent node, CN) creates two signaling sessions, one between the MN and the TRIM+ AS, and another between the TRIM+ AS and the CN. The TRIM+ AS is responsible for configuring, during session establishment, an address translator in the network. The easiest deployment configuration is to have the address translator located in the same network as the TRIM+ AS (in the MN home network). The address translator is kept in the data path of the communication. Once configured, the MN sends the data to an address belonging to the address translator. The address translator sits in the middle of the communication, changing the addresses of the data packets, so both the MN and CN receive packets with source addresses belonging to the address translator. Regarding the destination address of the packet, the address translator sends the packets to the respective addresses of the MN and CN. Note that the identification of the end points of the communication is through the SIP URIs, not the IP addresses that are acting just as locators. The association is done during the establishment of the session, the addresses of the address translator are given as the addresses of the MN, or the CN, to the communication peer. A movement of the MN is completely transparent to the CN, both in the data plane and in the signaling plane, as only the leg of the communication between the MN and the TRIM+ AS has to be updated. Therefore, the handover process requires the MN to execute a new IMS registration and a new INVITE transaction. The INVITE message arrives at the TRIM+ AS that reconfigures the address translator with the new IP address of the MN, hence updating the location of the MN in the data path.

Even then, the MN is changing its address with each movement and this would affect the applications running in the MN. Here is where the address translator in the MN comes into play (see in Fig. 3 the architecture of a TRIM+ enabled UE). This address translator basically presents always the same address (an internal address) to user applications, and translates between this address and the address in use by the MN in the current access network. A handover manager in the MN takes care of keeping the MN address translator updated with the addressing information. The IMS stack will be aware of the IP address in use by the MN in the

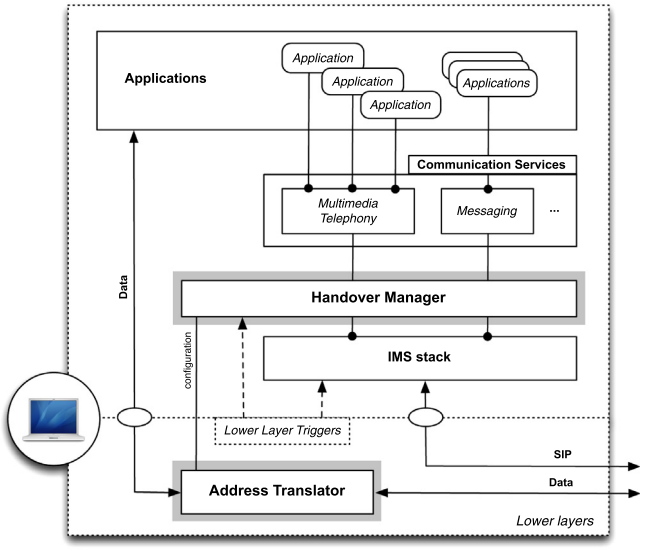

Fig. 3. Architecture of a TRIM+ enabled UE.

current access network, and that is the IP address that will be used for SIP signaling.

With this arrangement, TRIM+ provides transparent mobility support, but using a suboptimal data path. The problem is quite similar in nature and origin to the suboptimal route in Mobile IP. In both cases the data path is anchored in the home network of the MN. In Mobile IP, it is anchored in the Home Agent that does signaling and data functions. In TRIM+, it is anchored in the address translator in the network for the data path, and in the TRIM+ AS for the signaling path. This creates an inefficient data path, that has to go through the home network, which depending on the topology can significantly increase the communication end to end delay.

Similarly to the route optimization procedure in Mobile IPv6, TRIM+ is able to optimize the data path to allow a direct communication between the MN and the CN, while keeping the signaling anchored in the TRIM+ AS. This optimization is only possible if both participants in the communication are able to handle it, i.e., they implement the required procedures while keeping the mobility transparent to the applications.

The bottom part of Fig. 2 highlights the entities involved in the exchange of data traffic between the originating and terminating UEs, both without optimization and with optimization. In TRIM+ without route optimization, any data exchange requires an address translator in the UE home network. The route optimization procedure of TRIM+ allows each address translator in the network to be individually removed from the media path. This way, TRIM+ supports the exchange of data traffic between two compliant end points using both, one or none of the address translators (the latter case corresponds to a direct communication between end points). Route optimization procedures are described in subsequent sections.

TRIM+ requires support in the MN, i.e. we need TRIM+ specific software in the MN. TRIM+ also requires support in the network, but it adds elements that are IMS compliant (an IMS AS and an intermediate element in the data path), and their introduction does not require modifications to the IMS specifications or infrastructure. The important advantage of TRIM+ is that it neither requires modifications to the CNs, they can be standard IMS enabled nodes, nor to the user applications, which are kept unaware of the mobility. In TRIM+ an unmodified CN can communicate with MNs without being aware of their mobility, but a CN with TRIM+ functionality can participate in the TRIM+ route optimization procedure enjoying better performance.

3.2. Session setup procedures

Assume that for the execution of a given service, an application running at MN A needs to establish a multimedia session with another application running at MN B. For instance, because a user at MN A uses a multimedia telephony application to initiate a video call towards another user who is currently at MN B. The local application at MN A would generate an SDP offer describing the multimedia session to be established. This offer indicates, for each media component (e.g. audio or video), the IP address and ports where user traffic will be received. As MN A supports TRIM+ enhancements, any call from the local application to the lower layers to retrieve the IP address of the MN returns back an internal IP address, only meaningful within the scope of the MN (e.g. a private or a loopback address). This is a permanent address that is not affected by eventual changes in the network connectivity of the MN (e.g. after a handover to a new access network), and that will be used by the local application to exchange data traffic with any peer application, such as the one located at MN B.

After generating the SDP offer, the local application triggers the appropriate IMS communication service in MN A, in order to initiate the service execution (e.g. to initiate a video call in the case of a multimedia telephony application). The communication service implements the service logic, using a set of IMS enablers provided by an IMS stack.

Eventually, the IMS communication service uses the IMS stack to initiate the session setup towards the destination user. This involves sending a SIP INVITE request encapsulating the SDP offer. The Handover Manager (HM) in the TRIM+ architecture, is located between communication services and the IMS stack, thus it ensures that any SIP message sent towards the network includes real addressing information belonging to the MN. In particular, the HM changes the internal IP addresses included in the SDP offer by real IP addresses of the MN. Therefore, the IMS service is always initiated with real addressing information. Fig. 4 shows the different scenarios corresponding to session establishment. The figure outlines the session setup procedures in the originating and terminating sides where, for simplicity, we assume that communicating parties do not need to perform a resource reservation.

In the first scenario (case a in Fig. 4), the CN does not support TRIM+. For instance, the CN may be a UE connected to a fixed access network (e.g. a video streaming server). As the originating user is subscribed to the transparent mobility management service provided by TRIM+, the evaluation of initial filter criteria within the core IMS results in the routing of the INVITE request to a TRIM+ AS (AS_A). The TRIM+ AS becomes the anchor point at the signaling level to support the mobility of MN A, and remains on the path of any subsequent SIP requests and responses exchanged between the MN and the terminating side (the CN). During the session establishment, the TRIM+ AS configures an address translator in the MN home network, to forward the data traffic exchanged between the MN and the remote side. The address translator in the network anchors the media session, this way guaranteeing session continuity (for any transport protocol, for example TCP or UDP) in the event that MN A changes its IP address due mobility. On the other hand, as the terminating user is not subscribed to the service provided by our architecture, no TRIM+ AS (and consequently no address translator) is involved in the home network of the CN.

Once the IMS session is established, the CN sends the traffic in the session to an IP address of the address translator in the home network of the MN. This is because, during the session setup, AS_A provides the terminating party with an updated SDP offer, which reflects addressing information corresponding to the address translator AT_A. This address translator, using the information configured by the AS_A in the session setup procedure, translates the destination address to the temporal address used by the MN in its current visited network, and the source address to an address belonging to the address translator in the network. Following a similar approach, traffic in the reverse direction (originating at the MN) goes through the address translator in the network. Finally, we want to note that, although case a in Fig. 4 represents the scenario where the IMS session is originated by the TRIM+ MN, the procedures would be analogous if the session is initiated from the UE that does not support TRIM+ (i.e. CN_B).

Hence, a home network deploying TRIM+ allows the regular operation of legacy terminals, i.e., 3GPP terminals that do not implement the mechanisms proposed in this paper. Following the aforementioned procedures, IMS sessions established between regular 3GPP devices do not involve the utilization of the new functional entities defined by TRIM+. Additionally, our architecture supports the interoperation between a MN that implements TRIM+ and a regular 3GPP UE, following the procedures described in case a in Fig. 4. In this case, media traffic is always anchored in the home network of the TRIM+ MN, and our architecture can offer transparent mobility support (for MN movements) although not route optimization. The latter is not possible in this scenario because TRIM+ is unsupported by the peer communicating with the MN.

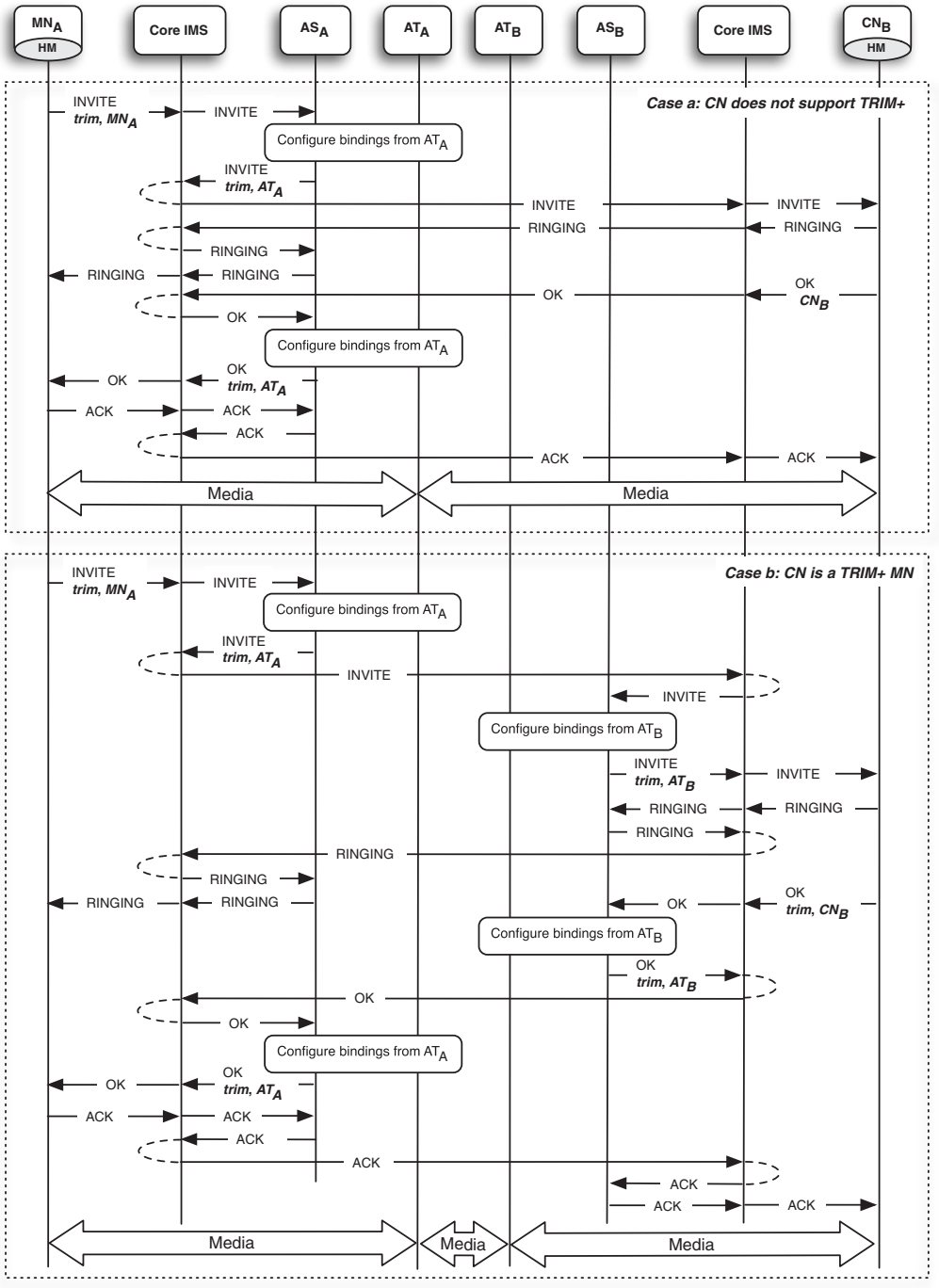

Fig. 4. TRIM+ processing of IMS session establishment.

TRIM+ defines a media feature tag that indicates that the SIP user agent (e.g. the MN) supports TRIM+, and will handle mobility transparently to its end user applications. This feature tag allows each party involved in the session establishment to determine if TRIM+ is supported by the other end of the SIP communication. This information is needed to decide whether route optimization procedures can or cannot be triggered. According to [29], the MN adds the feature parameter *trim* to the Contact header field value of the INVITE request, thereby indicating that it supports TRIM+. This feature parameter is also appended to the contact header field value of the INVITE request sent by the TRIM+ AS to the remote side. If the entity that terminates an INVITE transaction (i.e. the TRIM+ AS or the CN) supports TRIM+, it adds the feature parameter *trim* to the Contact header field value of the response to the INVITE. Note that the use of this feature tag does not prevent interoperabil ity with non compliant UEs, as the new header parameter will be considered unknown and ignored according to SIP regular procedures.

In the second scenario (case b in Fig. 4), the CN is a TRIM+ en abled MN. Therefore, a TRIM+ AS in the terminating network will configure an address translator, and data traffic will initially be ex changed between the originating and terminating MNs through their corresponding address translators. In this scenario, TRIM+ of fers transparent mobility and route optimization.

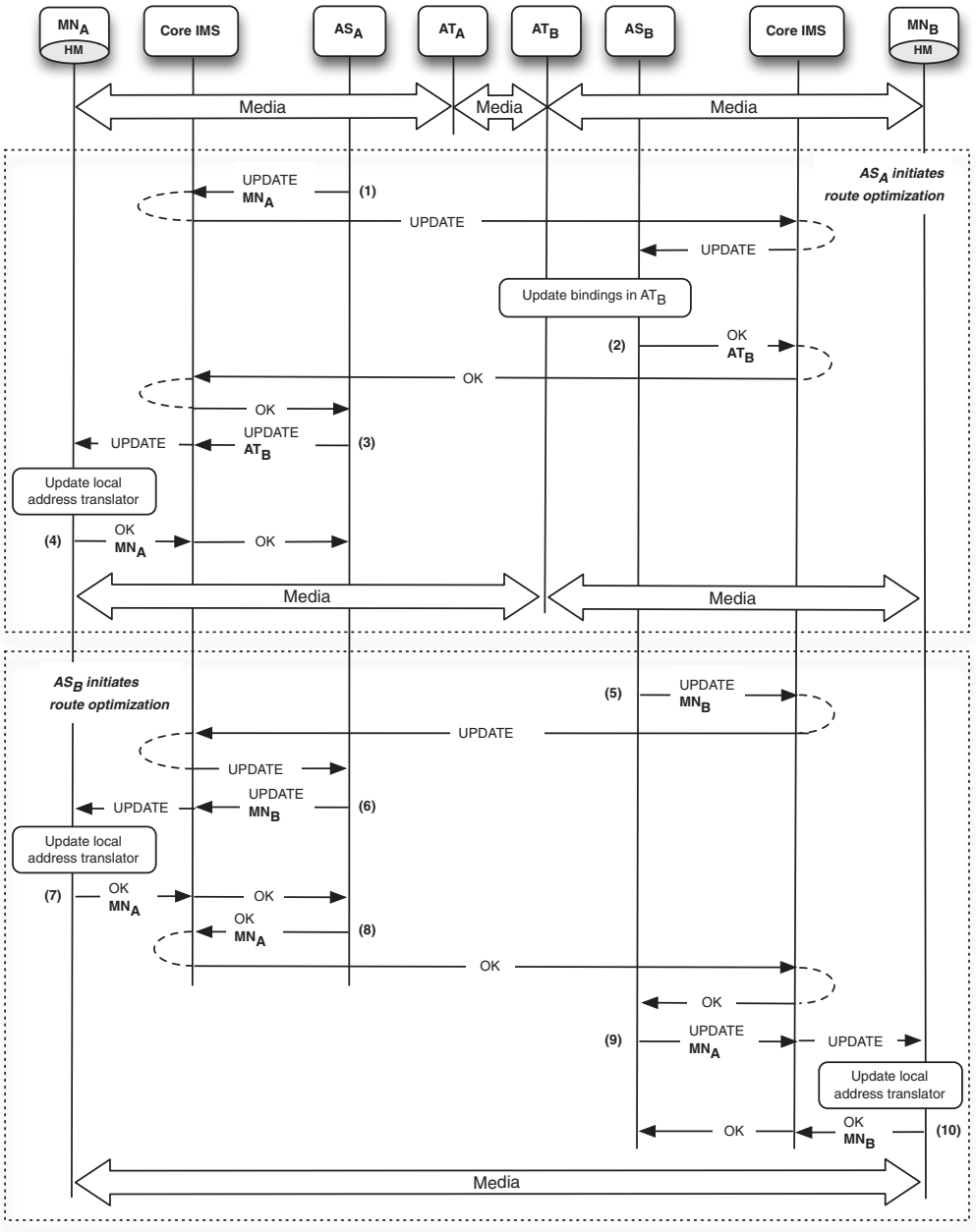

Fig. 5. Route optimization procedure.

In addition to these two cases, the design of TRIM+ considers a third scenario where the CN is connected to a fixed access network, and route optimization is still desirable. TRIM+ supports this func tionality but it requires upgrading the architecture of the CN as shown in Fig. 3. Note that a TRIM+ AS and an address translator in the network side of the CN are not needed, as the CN does not change the access network. Thus, by upgrading the software in its server equipment to make them TRIM+ compliant, a server pro vider (e.g. an IPTV or video streaming provider) can enhance the quality of the experience perceived by its TRIM+ mobile subscrib ers. Next subsections describe in detail the transparent route opti mization and mobility management procedures in TRIM+.

3.3. Transparent route optimization

Once the session has been established between the originating and terminating UEs, route optimization procedures can be exe cuted. In TRIM+, these procedures are started individually by each TRIM+ AS. However, it is important to understand that a TRIM+ AS can only initiate route optimization procedures if its SIP peers in the IMS session have indicated that they support TRIM+, by means of the *trim* feature parameter.

Fig. 5 shows the route optimization procedures that have been defined in this paper. The figure assumes that two TRIM+ MNs, A and B, have completed the IMS session setup procedures, therefore the media session is initially anchored by the address translators of both MNs.

In the figure we assume, without loss of generality, that the TRIM+ AS corresponding to MN A initiates the route optimization procedures. To do that, the AS generates a new SDP offer to update the session that the AS has established with its SIP peer in the ter minating side. This offer contains, for each media component in the session (e.g. audio or video), the addressing information (i.e. IP ad dress and port) where MN A is willing to receive the component

(the TRIM+ AS knows this information as a result of the session set up). Finally, the TRIM+ AS includes this SDP offer in a SIP UPDATE request, and sends the request towards the terminating side (step 1 in Fig. 5).

Eventually, the UPDATE request is received at the TRIM+ AS of MN B. The AS verifies that the media session is anchored in the address translator of MN B (AT B), and re configures this address translator to update the addressing information corresponding to the originating side of the media session. From this moment, data traffic coming from MN B to the address translator is directly addressed to MN A. The AS answers back the UPDATE request with a SIP OK response (step 2 in Fig. 5). This response carries an SDP answer, although it does not include new changes to the session description.

Note that the change in the configuration of AT B may lead to transient packet reordering. This is because packets in transit that were transmitted before the change will follow the route via AT A, which will typically impose a higher delay than the direct route from AT B to MN A (i.e., the new route after the change). This situation may potentially arise after changing the configuration of any of the address translators used in our proposal. However, we want to highlight that packet reordering is a common issue in the execution of applications that operate over packet networks, and this situation is typically handled at the transport layer (in case that TCP is used) or at the application layer (e.g. if UDP is used). Additionally, we want to emphasize that our solution does not introduce duplicate packets or packet loss due to any change on the configuration of an address translator. Packets arriving to an address translator are simply forwarded to the appropriate destination after the change.

When the OK response reaches the TRIM+ AS of MN A, the AS generates a new SDP offer to update the session leg corresponding to MN A. In this case, each media component in the offer includes the addressing information (i.e. IP address and port) corresponding to the address translator in the home network of MN B (the TRIM+ AS learned this information from the previous SIP signaling exchange with the terminating side). The TRIM+ AS encapsulates the SDP offer in a new SIP UPDATE request, which is sent within the dialog maintained with MN A (step 3 in Fig. 5). Note that this request could have been sent in parallel to the first SIP UPDATE, as the TRIM+ AS of MN A has already obtained the addressing information corresponding to the terminating side during the session setup procedures. Nevertheless, the design of TRIM+ processes these two UPDATE transactions sequentially, one after the other, this way ensuring that the remote side has processed the route optimization before updating the leg of the MN. This prevents undesirable situations where one of the peers of the TRIM+ AS fails to do the route optimization while the other succeeds. This may happen, for instance, when the TRIM+ AS of B receives the UPDATE request but has already started the execution of its own route optimization mechanisms and thus, it has already sent a SIP UPDATE request to the TRIM+ AS of MN A. According to [30], as both UPDATES contain an SDP offer, the TRIM+ AS of B must reject the request with a 491 failure response.

At some point, MN A receives the UPDATE request from the TRIM+ AS. Within the MN, the Handover Manager (HM) receives the session update request along with the corresponding SDP offer. The HM verifies that the SDP offer only specifies a change in the addressing information corresponding to the media components, and thus the session update can be done without the intervention of any IMS communication service or application. Consequently, the HM updates the configuration of the local address translator at the MN, to reflect the new addressing information where the media should be delivered (i.e. the address translator in the home network of MN B). Then, the HM generates an SDP answer to the offer, which does not include new changes to the session description. This answer is encapsulated in a SIP OK response to the UPDATE request, which is routed back to the TRIM+ AS (step 4 in Fig. 5).

At this point, the TRIM+ AS of MN A has concluded the route optimization procedures, which have been executed transparently to end user applications at MN A. Data traffic is now directly exchanged between MN A and the address translator of MN B. Therefore, the media session is no longer anchored by the address translator in the home network of MN A, which has been extracted from the data path.

The route optimization procedures executed by the TRIM+ AS of MN B proceed in a similar way. Consequently, AT B is removed from the data path. Data traffic is now directly exchanged between mobile nodes, following the shortest communication path between them.

The proposed mechanisms for route optimization allow each home operator to implement its own policies to independently decide if these mechanisms should be triggered and, in case they are, the specific instant to begin their execution. Nevertheless, it may always happen that both operators initiate the route optimization procedures within the same time period. This can lead to situations where a TRIM+ AS receives an UPDATE request while it is still waiting for a response to its own UPDATE request. According to [30], as both UPDATES contain an SDP offer, the AS must reject the UPDATE received from the remote peer with a SIP 491 failure response. After receiving 491 response, the AS starts a timer and, when this timer fires, attempts the UPDATE transaction again. In [30], a specific mechanism is described to choose the values for the timers corresponding to both ends of a SIP dialog. Therefore, even in the case where both operators initiate the route optimization procedures within the same time period, and both TRIM+ ASs reject the UPDATE request received from their SIP counterpart, timers eventually allow route optimization to be successfully completed.

A common case where this may happen is when the operator policy dictates to start the route optimization procedures when the SIP ACK request, corresponding to the session setup, is received at the TRIM+ AS (i.e. just after finishing the session establishment). In this case, both TRIM+ ASs start route optimization synchronously, thus creating a possibility for parallel UPDATE rejection that would increase the signaling overhead and delay the optimization process. To deal with this specific situation, in the design of TRIM+ we include a set of recommendations related with the initiation of route optimization procedures. In particular, the TRIM+ AS in the originating side of the session can start the route optimization procedures right after sending the SIP ACK request towards the terminating side. When the TRIM+ AS located in the terminating side receives this ACK request, it starts a timer with a value randomly chosen between 0 and T seconds. When this timer expires, the terminating AS can start the route optimization procedures. This recommendation aims at preventing both ends from synchronously beginning the execution of the optimization mechanisms. Values of T can be chosen by operators according to their own policies

Finally, the route optimization procedures described in this section are still valid when the communication takes place between a TRIM+ MN and a TRIM+ enabled UE connected to a fixed access network. The only difference in this scenario is that route optimization does not involve a TRIM+ AS in the home network of the fixed UE.

3.4. Transparent handover management

Fig. 6 illustrates the different scenarios of handover management, assuming that a MN has previously established a multimedia session with a CN. For simplicity, we assume that communicating parties do not need to perform a resource reservation.

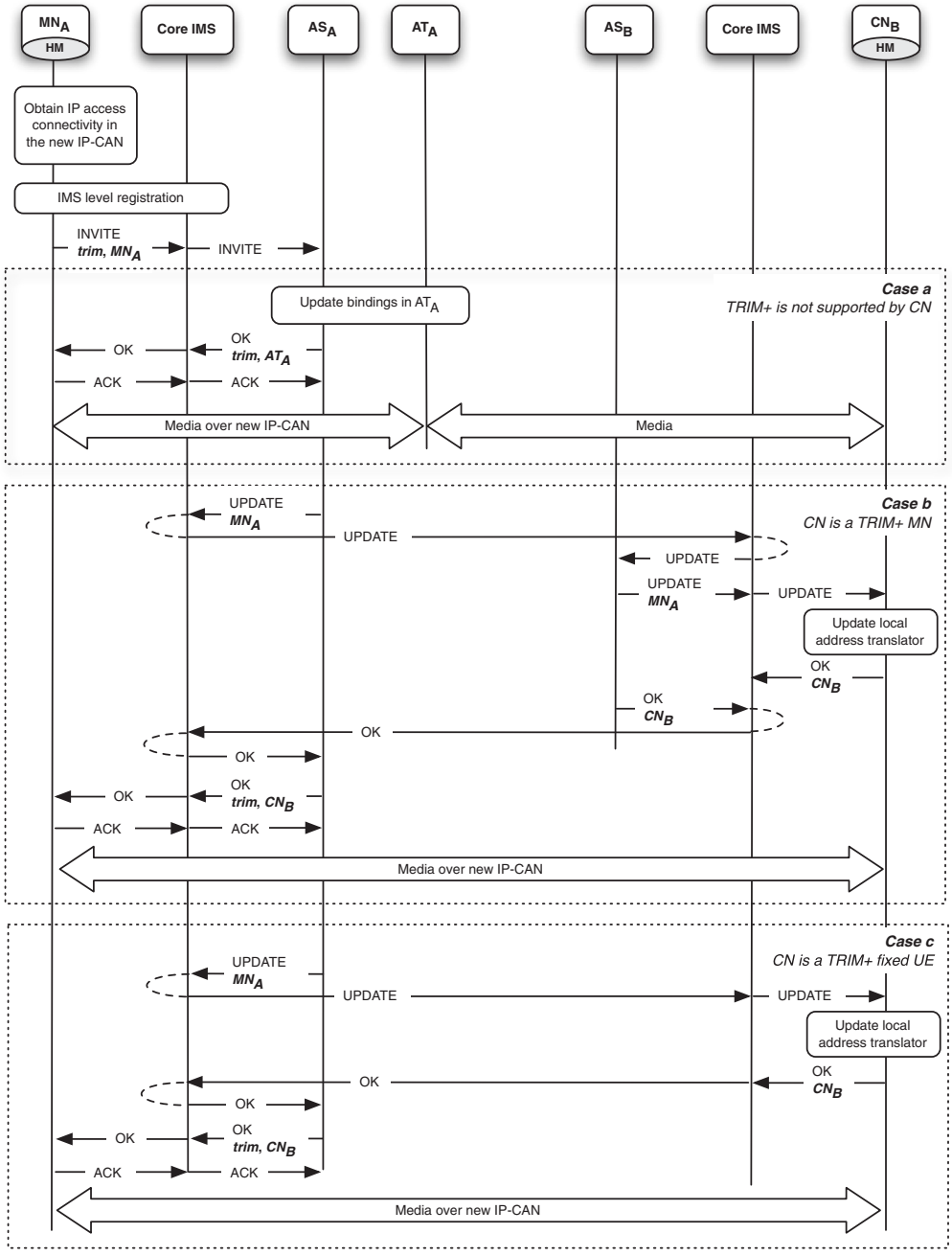

Fig. 6. Transparent mobility management procedure.

Independently of the handover scenario, the MN obtains IP access connectivity in the new IP Connectivity Access Network (IP CAN). As a result of this, it acquires an IP address in the new network, as well as the IP address of a P CSCF (if a new P CSCF is to be used). Then, the IMS stack and the HM are triggered from the lower layers[3] (see Fig. 3). This way, the IMS stack can register the new contact URI of the user, reflecting the current IP address, in the core IMS. Next, the HM retrieves the information corresponding to the active IMS sessions that are exchanging user traffic through the interface involved in the handover. For each of these sessions, the HM generates a new SDP offer containing the new addressing information associated with the interface (i.e. its current IP address). Then, the HM uses the IMS stack to update each affected multimedia session according to the new SDP offer, by means of an INVITE request. This request is routed to AS A, as this is the anchor point at the signaling level to support the mobility of MN A. Handover management procedures differ from this point, depending on the functionalities supported by the CN.

In the first scenario (case a in Fig. 6), TRIM+ is not supported by the terminating side. In this situation, the TRIM+ AS of MN A updates the new addressing information corresponding to the MN in the address translator (AT_A), and completes the SIP signaling exchange with the mobile node. The terminating leg of the communication is not affected in the signaling plane (TRIM+ AS to CN) nor in the data plane (address translator to CN).

In the second scenario (case b in Fig. 6), the CN is a TRIM+ MN. In the figure, we assume that the route between the MN and the CN

[3] Note that the direct support of link layer triggers is already present in technologies such as IEEE 802.11 as defined in the IEEE 802.11u amendment.

has already been optimized, following the procedures presented in the previous section. Therefore, the handover of the MN requires the IMS session in the terminating side to be updated with the new addressing information corresponding to MN A. This guaran tees that the communication between MN A and CN B can continue using the optimal route between them. Therefore, the TRIM+ AS of MN A generates a new SDP offer to update the remote leg of the multimedia session at CN B. The offer indicates, for each media component (e.g. audio or video), the current addressing informa tion (i.e. IP address and port) where MN A will receive the compo nent. The SDP offer is encapsulated in a SIP UPDATE request, which is routed to the TRIM+ AS in the terminating side.

As the session is not anchored in the terminating home net work, this AS propagates the UPDATE request to the CN. The HM at CN B receives the request to update the session according to the SDP offer. As the session description uniquely modifies the addressing information corresponding to the media components, the HM takes the responsibility of updating the multimedia session transparently to end user applications at the CN. Thus, it accesses to the configuration of the local address translator at CN B, and up dates it to reflect the current addressing information of MN A. Then, the HM generates an SDP answer that does not include fur ther changes to the session, and encapsulates it in a SIP OK re sponse, which is routed to the TRIM+ AS of CN B. In turn, the AS answers back the received UPDATE request with a new OK re sponse, indicating the successful update of the session in the ter minating side. When this response is received by the TRIM+ AS of MN A, the outstanding INVITE transaction is completed accord ing to the regular IMS procedures.

The scenario where the correspondent node is a TRIM+ fixed UE (case c in Fig. 6) proceeds similarly, although a TRIM+ AS is not in volved in the terminating side.

3.5. Management of packet loss during handovers

In the presented mobility management procedures, if the MN is temporarily able to receive media from the old access network dur ing the handover procedure, then the disruption of the media ex change can be minimized. Nonetheless, there are cases where maintaining the media exchange during a handover via the old network is simply not possible. In these cases, commonly encom passed under the denomination of *hard handovers*, packet loss due to MN mobility may significantly affect the performance of fered to end user applications.

The mobility management solution described in this paper ad dresses packet loss management in hard handover scenarios. To this end, TRIM+ enables the MN to request buffering capacity in its home network to temporarily store incoming media traffic dur ing the handover delay, which would otherwise be dropped. This solution, hereafter referred to as *in home buffering*, allows storing the user traffic transmitted from the CN while the media path is not established through the new access network. Once the MN suc cessfully completes the handover procedures, the multimedia ses sion with the CN is re established, and buffered media is delivered to the MN.

In home buffering is a general technique that could be applied to other architectures to improve handover performance (see Sec tion 4), but it fits nicely in our solution because we use existing nodes and signaling to implement the new functionality. This tech nique is supported by TRIM+ transparently to any end user appli cations running at the MN and the CN.

In particular, this solution may be particularly useful in the case of multimedia applications. Although the majority of these applica tions are tolerant to occasional packet loss, which can normally be concealed to the user, long interruptions causing packets loss may negatively impact the user experience. Networked multimedia applications typically use an application level buffer to store incoming media (e.g. RTP packets). Arriving media is removed from the buffer and rendered to the user after certain playout delay. This delay permits accommodating network jitter and ensuring that the majority of the packets are received in time for playout.

In TRIM+, traffic addressed to the MN during the handover is buffered in its corresponding home network (this traffic would otherwise be dropped). However, a multimedia application in the MN can still maintain the continuous playout of media using the data stored in its application level buffer. If this buffer is not emp tied by the application during the handover, data stored in the home network may be quickly transferred to the MN once the handover completes. This data will be placed in the application le vel buffer, which will be rapidly restored to an appropriate state, and will be available to the application for playout.

Of course, there is always the possibility of longer handover de lays that cannot be accommodated with the application level buf fer. Even in this case, data packets stored in the home network will promptly be delivered to the MN after a successful handover. So, packet loss is reduced, as TRIM+ still preserves media that would otherwise be lost in the absence of in home buffering.

Our buffering mechanism is illustrated in Figs. 7 and 8. I n Fig. 7, we assume a hard handover scenario, where a MN with a single ac tive interface needs to change the IP Connectivity Access Network (IP CAN) used for the communication with a CN. The picture shows the procedures that are necessary in TRIM+ to set up in home buf fering prior to the handover of the MN. Before the handover is ini tiated, the Handover Manager (HM) in the MN is triggered from the lower layers (see Fig. 3). As a consequence of this trigger, the HM issues a MESSAGE request that is addressed to the Public Service Identity (PSI) of the TRIM+ service (step 1 in Fig. 7). The HM indi cates in the body of the request that a handover is to be initiated, and includes the information that is necessary to activate in home buffering. In particular, this information contains the IP address of the MN in the current IP CAN. Additional mobility management data may also be included, such as information about the current access network and a description of the reasons triggering the handover. Eventually, as a result of evaluating initial filter criteria at the S CSCF, the MESSAGE request is delivered to the TRIM+ AS allocated to the user of the MN, which immediately answers back the request with a SIP OK response.

Our architecture uses MESSAGE requests to transport the infor mation that is necessary to activate in home buffering.[4] Although other mechanisms could be used, such as the SIP event notification framework [32], the utilization of MESSAGE requests brings impor tant advantages. In first place, these type of requests do not establish dialogs and do not require the periodic exchange of SIP messages to activate in home buffering. This way, resource utilization on IMS CSCFs and TRIM+ application servers, caused by a potentially large number of MNs, is reduced. Secondly, MESSAGE request are stand alone messages, and thus they can be sent over any of the registered interfaces of the MN.

The TRIM+ AS is an intermediate element that participates in every session setup originating and terminating in the MN. There fore, it maintains up to date information about all the multimedia sessions involving the MN. Using this information and the content of the MESSAGE request, the TRIM+ AS determines the media com ponents received by the MN that will be affected by the handover (i.e. those using the IP address of the MN indicated in the MESSAGE request). Then, it configures the address translator assigned to the MN within its home network to buffer incoming data packets from these media components (step 2 in Fig. 7).

[4] The SIP MESSAGE method is defined in [31], and its implementation is mandatory in IMS terminals.

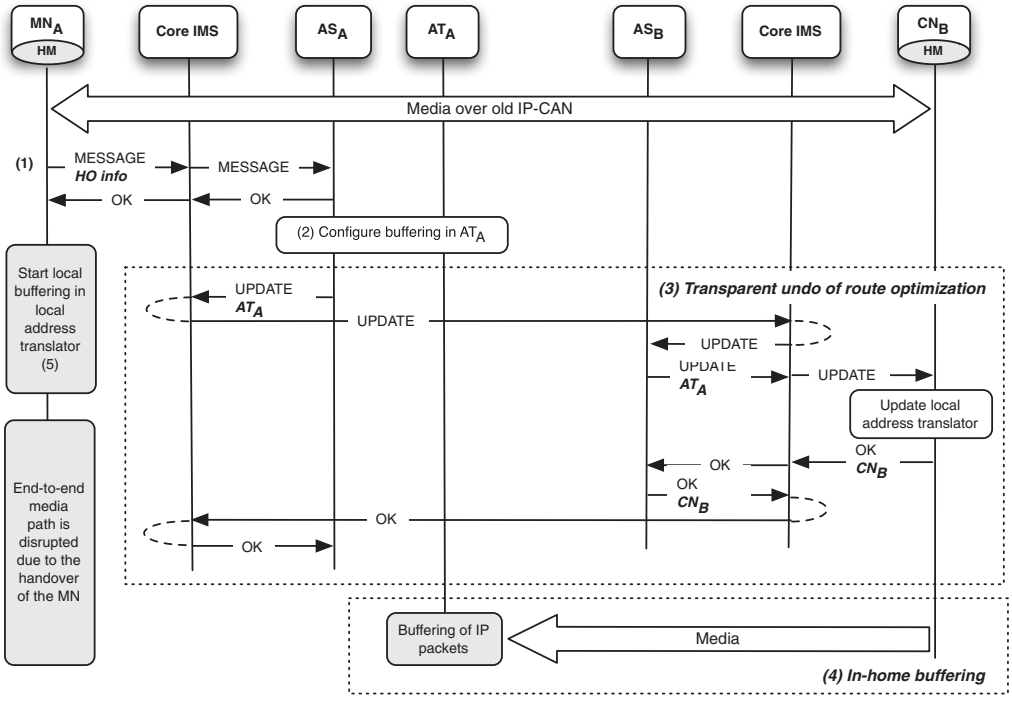

Fig. 7. Activating in-home buffering prior to handover.

In the example shown in Fig. 7, we assume that the communi cation between the MN and the CN is using the optimal media path. Therefore, the TRIM+ AS contacts the CN to undo the route optimization (step 3 in Fig. 7). As a result of this procedure, any media packets transmitted from the CN, which may otherwise be dropped during the handover of the MN, are temporarily buffered in its home network (step 4 in Fig. 7). It is important to note that the address translator buffering media traffic is an intermediary entity, which simply stores incoming IP packets until the MN com pletes the handover and the media path is re established through the new access network. Thus, it does not terminate the media path at the transport level (e.g. it does terminate a TCP connection in the data plane). Finally, we want to highlight that procedures comprising step 3 are not necessary in case that the media path was already anchored in AT_A.

We argue that in home buffering provides a scalable solution to mitigate packet losses in hard handover scenarios. As the commu nication with the MN is temporarily interrupted during the hand over, the TCP media flows from the CN will eventually stop the transmission of new packets, alleviating the demand of buffer space in the home network of the MN. This is because the address translator in the home network simply stores TCP media coming from the CN, and does not acknowledge the reception of TCP seg ments, as this can only be done by the other end of the TCP connec tion (i.e., the MN), which is unavailable during the handover. On the other hand, UDP flows originating from inelastic applications (e.g. real time audio and video) will not reduce their sending rate. Even in this case, the usage of buffer space in the home network is limited, as storage capacity is only needed for a short time period, i.e. the handover delay. Additionally, we want to emphasize that in the TRIM+ architecture, as it is illustrated in Fig. 2, the home net work may include a number of address translators (for scalability and redundancy reasons), each serving a number of TRIM+ termi nals. Therefore, buffering demands caused by the handover of MNs are distributed.

Anyway, in those cases where this distribution is insufficient, and a number of concurrent handovers leads to buffer space exhaustion in an address translator, a set of operator policies, configured in the TRIM+ AS and enforced in the address translator, may govern an appropriate distribution of the available buffer re sources among competing MNs.

On the other hand, once the MN receives the SIP OK response to the MESSAGE request, it can initiate the handover procedure to move to the new IP CAN. These procedures are illustrated in Fig. 8. In addition, and depending on the resources available at the MN, the HM may request from the local address translator to temporarily store the data packets transmitted by the MN during the handover (step 5 in Fig. 7). This would enable to mitigate pack et loss for outgoing traffic, which is generated by the applications running at the MN, while the multimedia session with the CN is reestablished via the new IP CAN.

Assuming that in home buffering has been triggered (resulting in the *initial setup* shown in Fig. 8), the MN obtains IP connectivity in the new IP CAN and completes the IMS level registration from the new access network (steps 1 and 2 respectively in Fig. 8). Then, it executes the access transfer procedures that are necessary to re establish the multimedia session with the CN via the new IP CAN (step 3 in Fig. 8). These procedures involve executing an INVITE transaction with the TRIM+ AS, to update the addressing informa tion of the MN in the signaling and data paths. As result of this transaction, the TRIM+ AS reconfigures the address translator as signed to the MN in the home network with the new addressing information, and media exchange between the MN and the CN is resumed through the address translator.

In TRIM+, once the access transfer procedures have been com pleted, buffered packets in the home network are transmitted at the maximum rate available in the address translator (AT_A in the example). As data transmission from the CN is buffered during a reasonably short time (i.e. the handover delay), we assume that the new IP CAN is able to process the traffic load corresponding to the buffered packets after re establishing the session. Although we believe that this will be the common situation, there may be scenarios where this assumption does not hold. In particular, inelastic UDP applications (e.g. real time audio and video), facing high handover delays, may introduce a high load of buffered traffic that exceeds the capacity allocated in the access network. Even in

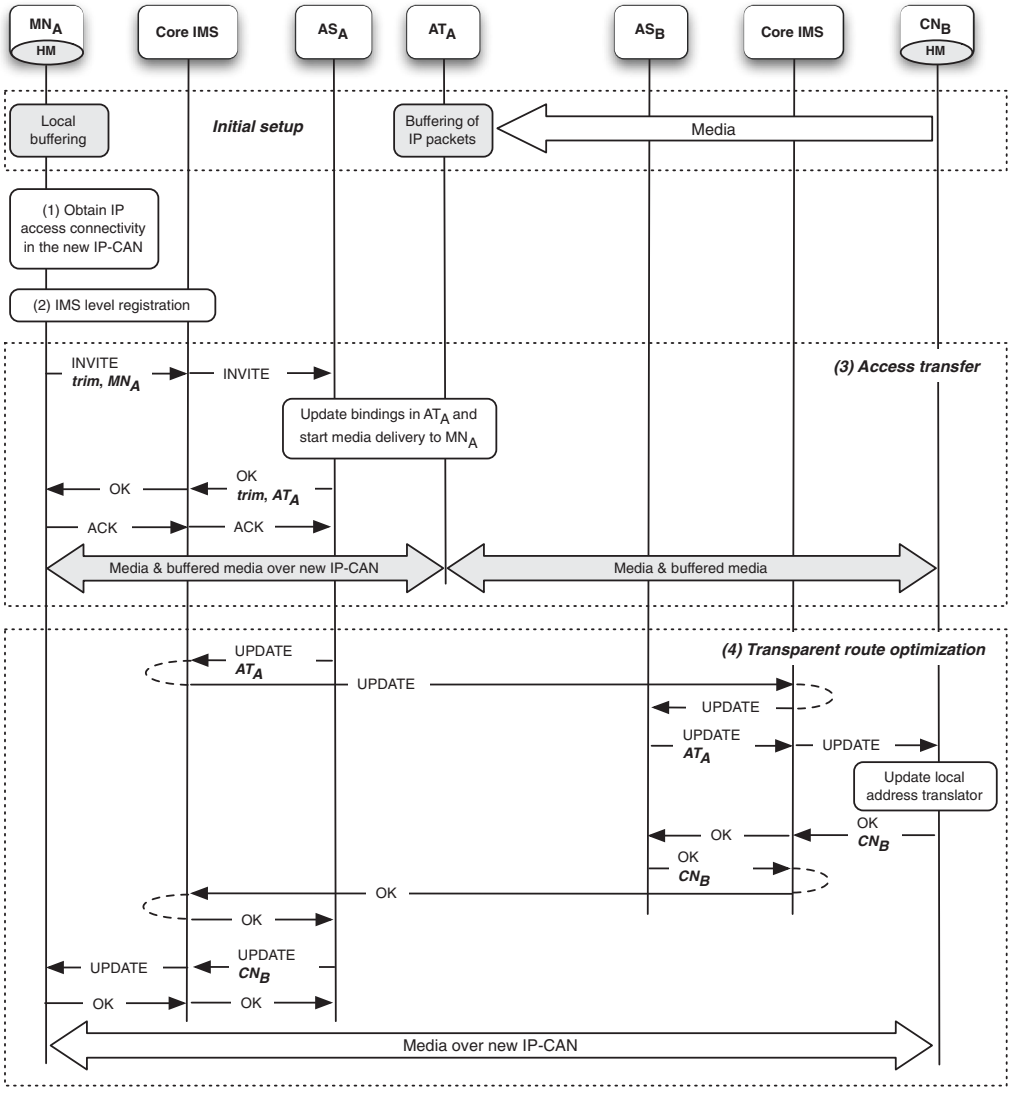

Fig. 8. Mobility management and recovery of buffered media.

this case, where packets may be dropped, our solution can mitigate packet loss to the extent enabled by the underlying IP CAN.

Finally, if TRIM+ is supported by the CN, the TRIM+ AS of the MN can re establish the optimal route via the new IP CAN (step 4 in Fig. 8).

4. Evaluation of the proposal

This section focuses on the evaluation of the mobility manage ment solution that has been presented in this paper. As a proof of concept, we have implemented a software prototype of the TRIM+ architecture. This prototype has been deployed over an evaluation testbed, which has been built using virtual machines. By using this testbed, we conducted several experiments allowing to validate the proposal presented in this paper.

In Section 4.1 we describe the evaluation testbed. In Section 4.2, we demonstrate the appropriate operation of our architecture, assuming UDP media communications in the data plane. The vali dation considering TCP media communications is included in Sec tion 4.3. In Section 4.4 we show the performance advantages of using in home buffering both for UDP and TCP traffic. Finally, in Section 4.5, we provide some considerations about the perfor mance that can be achieved by TRIM+, emphasizing the advantages of this proposal with respect to the architecture developed by 3GPP for the delivery of IMS Service Continuity [26].

4.1. Evaluation testbed

To validate TRIM+ and to demonstrate a real implementation of our proposal, we have implemented a software prototype of the TRIM+ architecture including all the functional entities presented in Fig. 2, i.e., a TRIM+ Application Server (AS), an address translator and a TRIM+ UE (including the modules depicted in Fig. 3). The implementation was written in Java, using the Java JAIN SIP API[5] for the SIP modules, and includes the different functionalities related with session setup (Section 3.2), transparent route optimization (Section 3.3) and and handover management (Section 3.4). To imple ment address translation functionalities, both in the network and in the UE, we used the Click! modular router platform.[6]

We deployed the software prototype of the TRIM+ architecture in a testbed of virtual machines, using a physical machine running the VMWare ESXi[7] virtualization software. The evaluation testbed

[5] JAIN SIP developer tools, http://jain-sip.dev.java.net/.
[6] http://www.read.cs.ucla.edu/click/click/.
[7] http://www.vmware.com/.

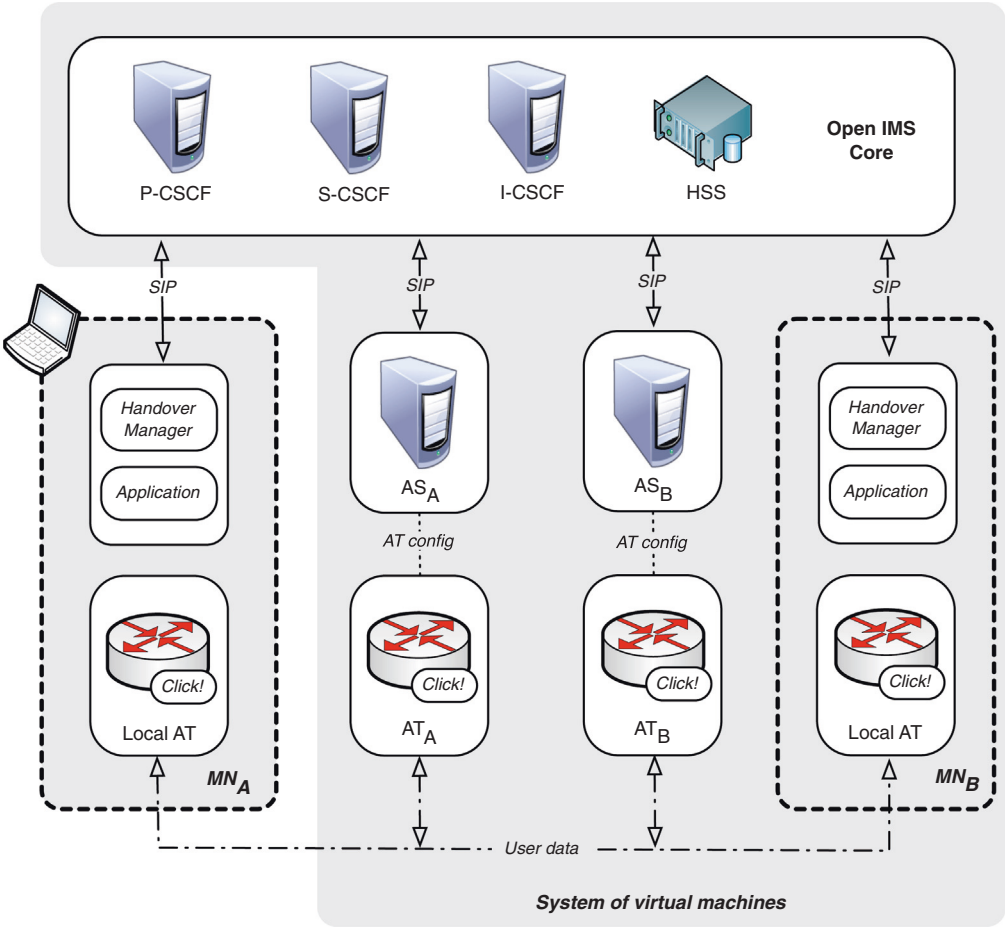

Fig. 9. Evaluation testbed.

includes a real IMS core, in particular, the Fokus Open IMS core,[8] which allowed us to verify the appropriate execution of all the IMS signaling procedures of TRIM+. Note that no modifications were needed to the IMS core, as they are not required in our proposal. We also included two TRIM+ mobile nodes, MN_A and MN_B. In our experiments, AS_A is the TRIM+ AS that anchors the signaling sessions of MN_A, while AT_A is the address translator that will be used in MN_A home network. Similarly, AS_B and AT_B are respectively the TRIM+ AS and the address translator that will be utilized for MN_B.

The testbed is shown in Fig. 9. To execute handover procedures over real IP Connectivity Access Networks (IP CAN), we deployed MN_A in an external laptop with both Ethernet and WLAN network interface cards (MN_A). The remaining entities in the testbed (namely $AS_A, AT_A, AS_B, AT_B, MN_B$ and the open IMS core) were de ployed as virtual machines. The evaluation scenario also includes a DNS server (that is omitted in the figure) and an HSS (Home Sub scriber Server). The HSS can be considered as the database of the IMS core, and we configured it with the identities of the users hold ing MN_A and MN_B and the TRIM+ service profiles.

In our experiments, a media communication is established be tween MN_A and MN_B. This communication can use different media paths, due to the execution of the route optimization and handover management procedures of TRIM+. Fig. 10 illustrates the data paths that are used in this evaluation, both for UDP and TCP media, according to the different optimizations enabled by our architec ture. To reflect the performance offered by the different media paths, each was configured with a given round trip delay within the system of virtual machines (these delays are also shown in the figure). Nonetheless, it is important to note that the actual end to end delay, for any of the considered media paths, also de pends on the network delay from the physical machine running the VMWare to the laptop deploying MN_A (that may be via Ether net or WLAN).

The round trip delays were enforced in our testbed using the Click! software. The configuration of Click! is based on the inter connection of different elements, which are pieces of code that implement a certain functionality. In our case, the configuration of each address translator (AT_A, AT_B and the local translators at MN_A and MN_B) has three different parts: first, the incoming traffic is classified based on the transport ports (source or destination) contained in the packets; the selected output of such classifier is connected to a *DelayShaper* element, which delays the transmis sion of incoming packets by a configurable time; finally, the packet headers are rewritten in order to properly change the addressing information.

4.2. Validation of UDP media communications

In this section, we assume that communications taking place in the user plane are based on UDP. In our experiment, MN_A and MN_B first register in the IMS core. Then, MN_A starts a session setup to wards MN_B using the IMS core, following the signaling exchange presented in case b of Fig. 4. When the session is established, a cli ent application based on *iperf*, running at MN_B, starts sending UDP traffic towards an *iperf* server application, which runs at MN_A at a rate of 1 Mbps. According to TRIM+ procedures, UDP packets are

[8] http://www.openimscore.org/.

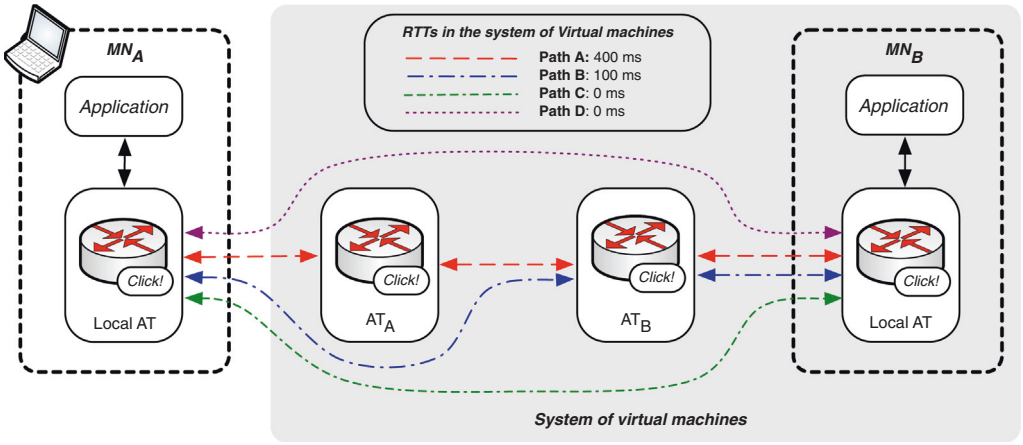

Fig. 10. Media paths in the communication between MN_A and MN_B.

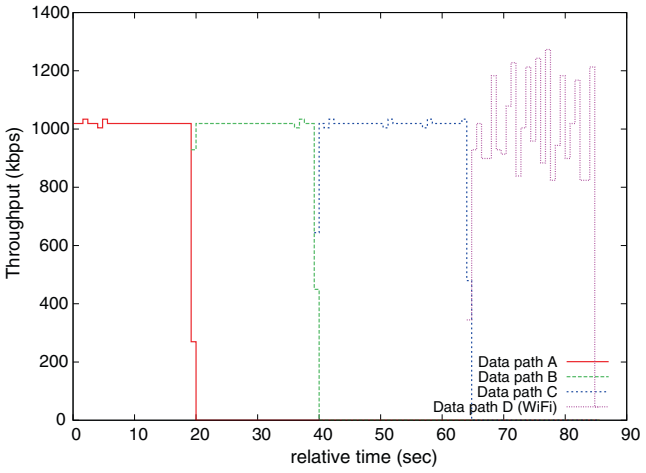

Fig. 11. Route optimization and handover of an UDP flow.

initially delivered to MN_A via a suboptimal data path that traverses AT_B and AT_A (i.e., data path A in Fig. 10). In our software prototype, we configured AS_A to start the route optimization procedures approximately 20 s after the session setup, using the signaling ex change illustrated in Fig. 5. As a result of these procedures, AT_A is extracted from the data path and UDP packets are delivered to MN_A through AT_B (i.e. using data path B in Fig. 10). On the other hand, AS_B waits for 40 s to start route optimization, which take place according to Fig. 5. Once the data path optimization con cludes, traffic is delivered directly from MN_B to MN_A (i.e. using data path C in Fig. 9).

After some time, we manually initiate a handover from the Ethernet to the WLAN access network. As a result of this, the IMS stack and the handover manager (HM), implemented at MN_A are triggered. This way, the IMS stack initiates a new IMS level regis tration, updating the location of the MN in the core IMS (now it is available in the WLAN access network). Then, the HM at MN_A re establishes the multimedia session via the WLAN IP CAN, by executing the signaling exchange shown in case b of Fig. 6. To avoid packet losses, the configuration of MN_A is changed to start using the wireless interface while the Ethernet is still receiving media. Once the session is re established, MN_A starts receiving UDP pack ets via de WLAN interface (path D in Fig. 5).

It is important to remark that the *iperf* applications running at MN_B and MN_A are launched when the IMS session is established, and route optimization and handover management are completely transparent to these applications. This is enabled by the handover manager running at each mobile node, which changes the configu ration of its local address translator to conceal any change on the addressing information to the end user application (i.e. *iperf* client and server).

Fig. 11 shows the throughput of the UDP traffic, transmitted from the *iperf* server at MN_B and measured at MN_A. The figure shows the traffic that is received through paths A, B, C an D. Note that although TRIM+ makes route optimization and handover pro cedures transparent to the end user application running at MN_A, i t was possible to distinguish these sub graphs using the Wireshark[9] network packet analyzer listening at the *any* interface of MN_A.

4.3. Validation of TCP media communications

This section focuses on validating the proposed solution assum ing that media communications are based on TCP. To provide a realistic scenario, where TCP traffic exchanged between MN_A and MN_B competes with other traffic flows for the available bandwidth, we limited the bandwidth on the network path connecting MN_A with the system of virtual machines to 10 Mbps. In addition, we configured a continuous background traffic through that network path, by setting a TCP bulk data transmission with the *iperf* tool. This allowed measuring the benefits of the optimization in terms of throughput.

For the experiment, we followed the same steps as in the vali dation of UDP media communications (see Section 4.2): we estab lished an IMS session between MN_A and MN_B; we set a TCP transmission from MN_B to MN_A using the *iperf* application; we con figured the execution of route optimization procedures after the same time periods; and we executed a handover of MN_A from the Ethernet to the WLAN access network. In order to present the re sults, in Fig. 12 we show the Sequence number and Throughput vs Time graph obtained as the result of the complete experiment.

Focusing on the graph of the sequence number, it can be seen how the slope of the curve is different for each of the data paths taken by the flow (i.e., paths A, B, C and D). This behavior is due to the inverse relation of TCP throughput with the round trip delay of the path. Following this idea, it can be seen how the part of the graph corresponding to path C (the one with smaller delay) is the one with a stepper slope. It is worth to notice, that the handover and the changes of path for this experiment do not cause packet

[9] http://www.wireshark.org/

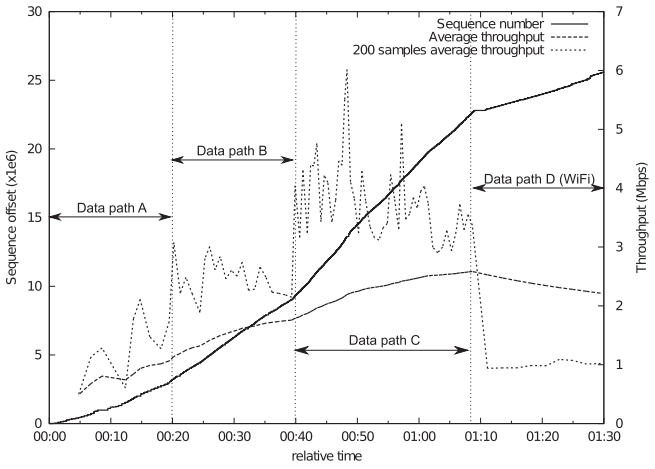

Fig. 12. Sequence number and Throughput vs Time.

loss, as old paths are maintained active to deliver outstanding traf fic in transit.

This analysis shows experimentally how the use of our optimi zation procedures has a positive impact on the throughput achiev able by TCP, while also reducing the RTT, without requiring the re establishment of TCP connections. Hence its use is beneficial, specially in environments where the delay between the ATs is very high or when the MNs are far away from their corresponding AT.

4.4. Performance analysis of in home buffering

Once we have validated our mobility management proposal and its route optimization mechanism, in this section we turn our attention to analyze the benefits in terms of performance that can be achieved with in home buffering in the data plane. In order to show the benefits of using a buffering mechanism instead of dis carding packets, two different scenarios have been tested: (i) a con stant bit rate flow generated by means of *iperf* and (ii) a file transfer. While in the former, both UDP and TCP transport proto cols where used, in the latter just TCP was tested. Both scenarios share the same testbed already used in Section 4.1 and depicted in Fig. 9, maintaining the bandwidth limitation of 10 Mbps in the network path from the system of virtual machines to MN_A. In all these experiments, the MN_A acts as the receiver so, AT_A will be con figured to buffer packets, when the buffering is activated or to dis card packets if it is not. We always use the suboptimal path via AT_A with a configured RTT of 30 ms. The Click! modular router that we use to implement address translation also allows us to introduce the buffering functionality. In particular, the buffering function is implemented as a *Queue* connected to an *Unqueue* element, used to pull packets from the queue: when the buffering is active, the *Unqueue* element is inactive, so all received packets are stored at the buffer. As soon as the buffering receives the proper signal, the *Unqueue* element is activated again, so it starts pulling the queue at the maximum available rate. In case of discarding, the *Un queue* element sends its output to a discard element.

In the first scenario, an instance of *iperf* is configured as a server at MN_A, so it can receive a 1 Mbps traffic generated by an *iperf* cli ent running at MN_B, through AT_A. In case of UDP traffic, *iperf* allows configuring the transmission rate. In the TCP case, this rate limita tion was enforced at MN_B using the *tc* tool.[10] In order to simulate a handover, after approximately 4 s,[11] AT_A is configured to buffer or discard packets for 1 s. The results for TCP are shown in Fig. 13, while Fig. 14 presents the results for UDP. In the TCP experiments, the fig ures show the segments received versus time and, as it can be seen in Fig. 13(a), the gap when the buffering is not active is higher than one second. This is because the RTO (retransmission timeout) at MN_B has increased as no ACKs have been received for one second of inter ruption. The zoom area inside the figure, depicting the first packets received just after the interruption, shows that MN_B restarts trans mission using the TCP slow start algorithm. On the other hand, as shown in Fig. 13(b), when the buffering mechanism is active, the ex tra interruption time is lower. The explanation is this behavior is the following: as soon as the AT_A receives the order to transmit the pack ets stored at the buffer, it uses the maximum available bandwidth to transmit all of them. Eventually, those packets will be received by MN_A, which generates the corresponding ACKs packets, allowing MN_B to continue the paused transfer (TCP pauses the transmission during the interruption because it is not receiving ACKs). In this case, the first packets received after the interruption are the ones stored at the AT_A buffer, as shown in the zoom area of Fig. 13(b). The packets in the buffer are not only recovered packets, they also allow to boot strap the TCP connection opening the TCP transmission window in MN_B.

The main difference between UDP and TCP is that in the former, all packets discarded do not arrive to the destination, as shown in Fig. 14(a). In case buffering is used, all packets stored at the queue are transmitted at the maximum available bandwidth after the

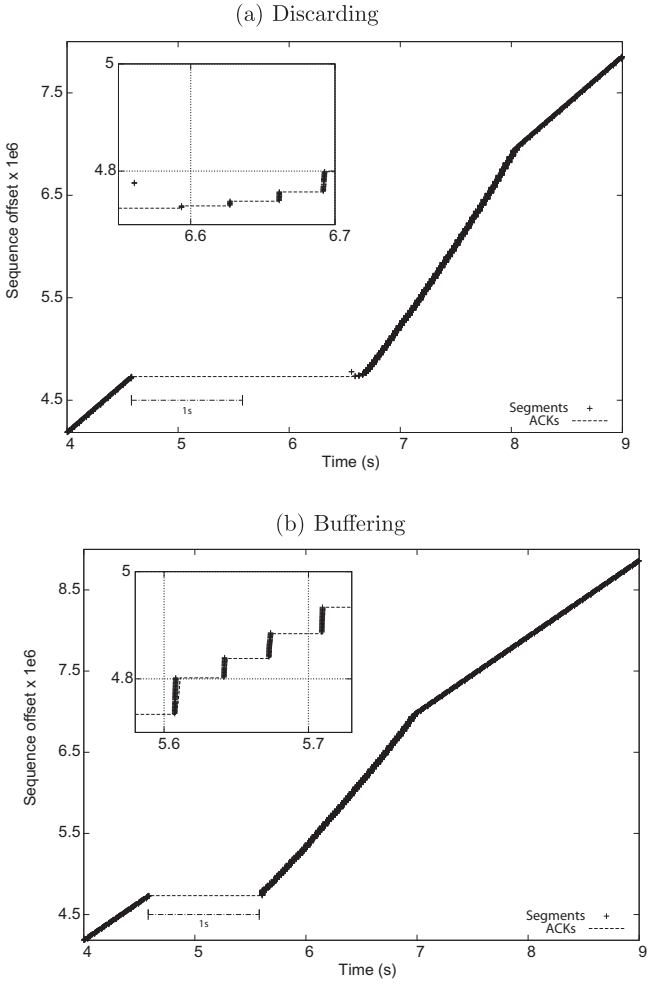

Fig. 13. TCP results.

[10] http://www.linuxfoundation.org/collaborate/workgroups/networking/netem.

[11] This value was chosen so that the time period from the beginning of the experiment was enough to reach the target rate of the iperf source, i.e., 1 Mbps.

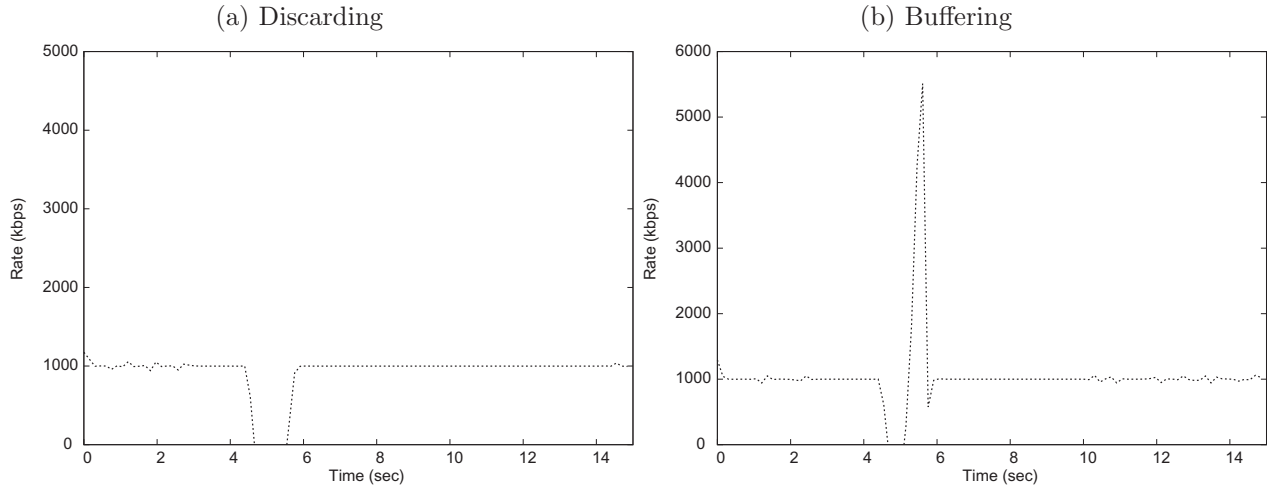

Fig. 14. UDP results.

interruption so, as it can be seen in Fig. 14(b), they are recovered very quickly after the interruption with no packet loss.

The second scenario, as stated before, includes a file transfer between the two mobile nodes. This scenario is intended to provide an idea about how long it takes to recover after an interruption. In this case, MN_A is downloading a 2 MBytes file stored at MN_B, through AT_A. Similarly to the previous scenario, the flow corresponding to the file transfer is limited to 1 Mbps. Four different interruption times, representing handover delays, are provided: 0 s (i.e., no interruption), 400 ms, 1 s and 3 s. 400 ms has been chosen because it is the maximum end to end delay for voice recommended by ITU T [33]. One and three seconds are used as upper bounds for real interruptions during a handover. For every single interruption time, 60 runs were executed, where the initial time for the interruption follows a random variable with uniform distribution between 1 and 15 s. The results obtained are shown in Fig. 15, where the symbol marks represent the mean value of the executions and the arrows centered at the mean show the 95% confidence interval for each value. These results demonstrate the benefits of the buffering mechanism, specially for low interruptions, where the transfer time is similar to the values obtained with the normal download process. Moreover, the buffering reduces the variance of the transfer time compared with the discarding process, where the transfer time strongly depends on the moment of the interruption.

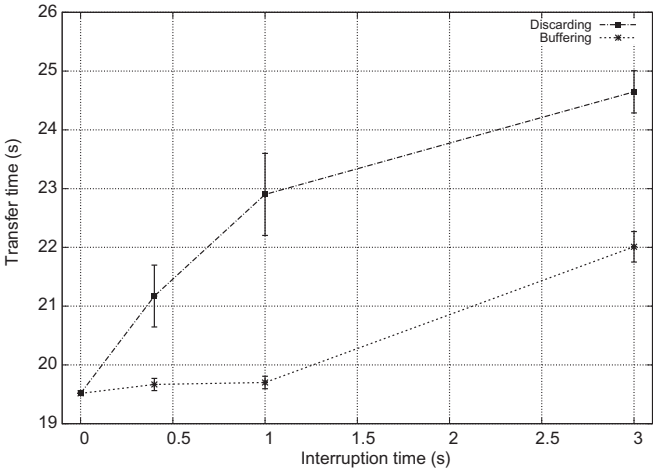

Fig. 15. Transfer time.

4.5. Considerations about performance

In this section, we provide a set of considerations about the performance that can be achieved by TRIM+, and we analyze the main benefits of this proposal with respect to the architecture defined by 3GPP to maintain IMS Service Continuity (SC) in the event of terminal mobility [26]. Our analysis does not cover other solutions for mobility management that we describe in Section 2, as these solutions cannot be directly integrated in IMS networks [8,10].

One of the main metrics that can be used to evaluate the performance of a mobility management solution is the **handover delay**. In the following, we provide an estimation of the handover delay that can be achieved by TRIM+, using an approach similar to [2]. Assuming that a mobile node (MN) is using a given access network to exchange media traffic with a correspondent node (CN), we define the handover delay HO_d as the time from the instant the MN starts the attachment procedure to a new access network until it can receive media traffic in that network. Fig. 16 represents the different contributions corresponding to the handover delay. The figure presents a simplified view of the handover management procedures of TRIM+, encompassing cases a, b and c of Fig. 6. The contributions to the handover delay are described next:

- $IPcan_d$ is the delay that is necessary to get IP access connectivity in the new IP Connectivity Access Network (IP CAN). This delay includes the time required to acquire an IP address in the new network, as well as the IP address of a P CSCF (if a new P CSCF is to be used).
- *Reg* is the time consumed by the IMS level registration from the new network.
- *AccessTransfer* comprises the time that is required to re establish the multimedia session via the new access network. This procedure involves a SIP signaling exchange between the MN and the TRIM+ AS. Fig. 16 shows a simplified overview of this SIP exchange. The detailed procedures are illustrated in Fig. 6 for the three possible cases: (a) TRIM+ is unsupported by the CN, (b) the CN is a TRIM+ MN and (c) the CN is a TRIM+ fixed UE.

The handover delay can then be expressed as the sum of the aforementioned contributions:

$$HO_d \quad IPcan_d + Reg + AccessTransfer \tag{1}$$

If TRIM+ is not supported by the CN (case a in Fig. 6), TRIM+ anchors the data path in an address translator located in the home network

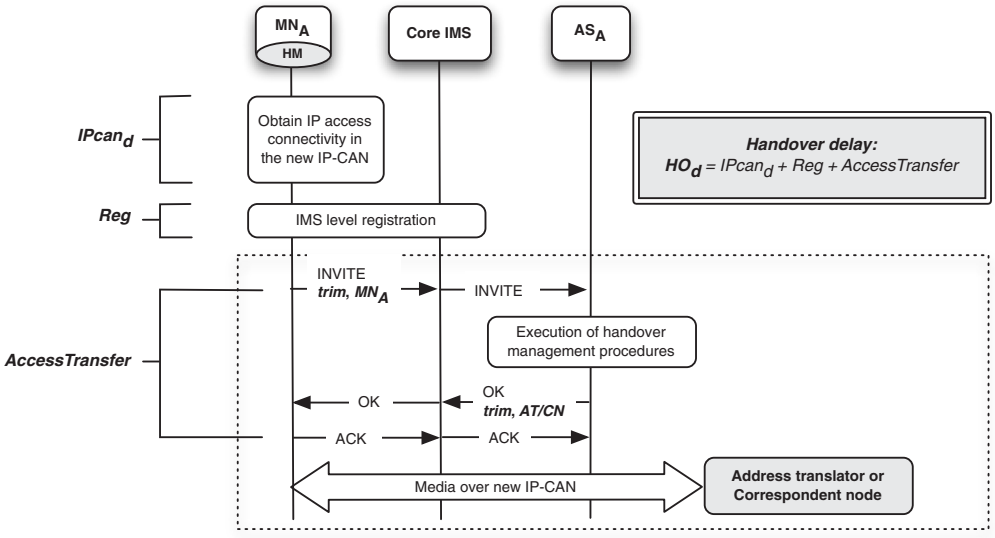

Fig. 16. Contributions to the handover delay.

of the MN. In this scenario, the access transfer is initiated by the MN by means of a single INVITE transaction with the TRIM+ AS. As a result of this transaction, the AS reconfigures the address translator with the new addressing information of the MN. Thus, media traffic is forwarded to the new location of the MN. The handover delay corresponding to this scenario, given by Eq. 1, is the same that can be achieved with our previous proposal (i.e. TRIM [2]), which does not integrate any route optimization procedures. We denote this handover delay as $HO_d^{case\ a}$.

On the other hand, if the CN is a TRIM+ MN (case b in Fig. 6), then route optimization is feasible. Assuming that the media communication between the MN and the CN is using the optimal data path, the access transfer incurs in an additional delay, as the CN must be updated with the the new addressing information corresponding to the MN. The handover delay achieved by TRIM+ in this scenario will typically exceed $HO_d^{case\ a}$. The reason for this is that updating the multimedia session in the CN requires the exchange of SIP signaling messages through the set of IMS entities that serve the CN, which are not necessarily located in the home network of the MN. On the contrary, if TRIM+ is not supported by the CN, access transfer only requires the reconfiguration of an address translator from the TRIM+ AS, being both of them located in the same operator domain.

However, we want to emphasize that the handover delay achieved by TRIM+ in this scenario (case b in Fig. 6), given by Eq. 1 and hereafter denoted $HO_d^{case\ b}$, is comparable to the handover delay provided by the solution defined by 3GPP for IMS SC. This is because access transfer in IMS SC also comprises connecting to the new IP CAN (and configure a new IP address), registering to the S CSCF via the new IP CAN, initiating an INVITE transaction to an intermediate element that anchors the IMS session in the home network of the MN (i.e. the Service Centralization and Continuity Application Server), and updating the session in the CN.

If the CN is a TRIM+ fixed UE (case c in Fig. 6), route optimization is also possible. Assuming that the communication between the MN and the CN is using the optimal media path, then the handover delay obtained by TRIM+, given by Eq. 1 and hereafter referred to as $HO_d^{case\ c}$, will typically be lower than $HO_d^{case\ b}$, as updating the session in the CN does not involve a TRIM+ AS in the terminating side.

With respect to the **length of the data path**, IMS SC uses the optimal media path between the MN and the CN, as the multimedia session is always updated in the remote side when the MN gets IP connectivity in a new network. Nevertheless, the handover management procedures triggered by the MN are not transparent to the end user applications running at the MN and the CN, which must be programmed with mobility in mind. In particular, TCP media flows will definitely be affected, and application specific mechanisms will be needed to recover TCP connections. In addition, UDP applications are also likely to be affected, needing to reconfigure or to open new network sockets.

On the contrary, mobility management procedures in TRIM+ are executed transparently to any end user applications, which can communicate using UDP or TCP in the data plane. If TRIM+ is supported both by the MN and the CN, the proposed solution can use the optimal route between them, achieving the same performance as IMS SC in terms of length of data path, but adding the additional advantage of **transparency**. In case that TRIM+ is not supported by the CN, our architecture still provides transparent handover management of the MN, anchoring the data path in an address translator located in the home network of the MN. This way, the CN is provided with stable addressing information for the duration of

Table 1
Performance comparison between TRIM+ and IMS SC.

	HO delay[a]	Data path	Transparent management	Packet loss management
TRIM+ case a	$HO_d^{case\ a}$	Suboptimal	Supported	Hard/Soft handover
TRIM+ case b	$HO_d^{case\ b}$	Optimal	Supported	Hard/Soft handover
TRIM+ case c	$HO_d^{case\ c}$	Optimal	Supported	Hard/Soft handover
3GPP IMS SC	$HO_d^{case\ b}$	Optimal	Unsupported	Soft handover
TRIM [2]	$HO_d^{case\ a}$	Suboptimal	Supported	Soft handover

[a] Note that, as it was previously indicated, the value of $HO_d^{case\ b}$ will typically exceed the values of $HO_d^{case\ a}$ and $HO_d^{case\ c}$.

the multimedia session established with the MN. However, achiev ing transparency in this scenario, which is the main objective of TRIM+, comes with the cost of utilizing a suboptimal data path in the user plane.

Additionally, TRIM+ supports **packet loss management** in hard handover scenarios. In these scenarios, the end to end media path is disrupted during the handover delay, as the MN cannot use the previous access network while it re establishes the multimedia session via the new one. TRIM+ addresses packet losses in hard handover scenarios by means of in home buffering. As we have shown in the previous section, this scheme can improve the perfor mance of media delivery under these scenarios, independently of the transport protocol being used in the user plane (i.e. UDP or TCP). On the contrary, IMS SC can only tackle packet losses in soft handover scenarios, where the MN is able to receive media traffic from the old access network during the handover delay.

Table 1 summarizes the different considerations presented along this section, with respect to the performance achieved by TRIM+ and IMS SC. As a reference, the table also shows the perfor mance that may be achieved with our previous solution, TRIM [2], which does not include route optimization nor specific mecha nisms to address packet losses in hard handover scenarios.

5. Conclusion

This paper describes TRIM+, an architecture to support trans parent mobility management and route optimization procedures in the IMS. TRIM+ supports a route optimization procedure, allow ing compliant UEs to use a direct communication path to exchange user traffic. TRIM+ also addresses packet loss management in hard handover scenarios, where the media path is disrupted during the handover of the MN. The designed approach, referred to as *in home buffering*, enables the MN to temporarily store incoming traffic in its home network during the handover (this traffic would other wise be dropped). As a proof of concept, we have implemented a software prototype of our architecture and we have deployed it in a real IMS testbed. Our evaluation includes a comparison of per formance with the 3GPP IMS Service Continuity and a set of exper iments, with UDP and TCP based applications, which validate our design and highlight the main benefits of our solution.

In summary, TRIM+ has a set of properties that are not available together in the alternative solutions in the literature:

(1) Its mobility support procedures are transparent to applica tions and correspondent nodes.
(2) It is compatible with legacy terminals, i.e., a TRIM+ terminal can communicate with a standard 3GPP terminal or other ter minals without TRIM+ functionality, and still have mobility support.
(3) It does not require modifications to the IMS infrastructure.
(4) Its performance is comparable to that of the 3GPP IMS Ser vice Continuity in terms of communications end to end delay and handover delay, and it is better in terms of packet loss man agement during hard handovers.

Our future work includes exploring the potential of TRIM+ to handle the transfer of a media session to a different terminal, and to support transparent mobility to legacy circuit switched networks.

Acknowledgements

This article has been partially granted by the Madrid Commu nity through the MEDIANET project (S 2009/TIC 1468), and by the European Community through the CROWD project (FP7 ICT 318115). The work of Ignacio Soto has been partially sup ported through the I MOVING project (TEC2010 18907).

References

[1] 3GPP, IP Multimedia Subsystem (IMS); Stage 2, TS 23.228 version 11.7.0 Release 11, TS 23.228, 3rd Generation Partnership Project (3GPP), January 2013.
[2] I. Vidal, A. de la Oliva, J. Garcia-Reinoso, I. Soto, TRIM: an architecture for transparent IMS-based mobility, Computer Networks 55 (7) (2011) 1474–1486.
[3] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, SIP: Session Initiation Protocol, RFC 3261, Internet Engineering Task Force, June 2002.
[4] M. Handley, V. Jacobson, C. Perkins, SDP: Session Description Protocol, RFC 4566, Internet Engineering Task Force, July 2006.
[5] 3GPP, IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3, TS 24.229, 3rd Generation Partnership Project (3GPP), June 2013.
[6] C. Perkins, IP Mobility Support for IPv4, RFC 3344, Internet Engineering Task Force, August 2002.
[7] C. Perkins, D. Johnson, J. Arkko, Mobility Support in IPv6, RFC 6275, Internet Engineering Task Force, July 2011.
[8] I. Vidal, J. Garcia-Reinoso, A. De La Oliva, A. Bikfalvi, I. Soto, Supporting mobility in an IMS-based P2P IPTV service: a proactive context transfer mechanism, Computer Communications 33 (14) (2010) 1736–1751.
[9] T. Chiba, H. Yokota, A. Dutta, D. Chee, H. Schulzrinne, Performance analysis of next generation mobility protocols for IMS/MMD networks, International Wireless Communications and Mobile Computing Conference, IWCMC'08, IEEE, 2008, pp. 68–73.
[10] T. Renier, K.L. Larsen, G. Castro, H.-P. Schwefel, Mid-session macro-mobility in IMS-based networks, Vehicular Technology Magazine, IEEE 2 (1) (2007) 20–27.[11] X. Chen, J. Wiljakka, Problem Statements for MIPv6 Interactions with GPRS/UMTS Packet Filtering, Internet-draft, Internet Engineering Task Force, work in progress, 2007.
[12] M.M.A. Khan, M.F. Ismail, K. Dimyati, Seamless handover between WiMAX and UMTS, IEEE 9th Malaysia International Conference on Communications (MICC), IEEE, 2009, pp. 826–830.
[13] N.M. Alamri, N. Adra, Integrated MIP-SIP for IMS-based WiMAX-UMTS vertical handover, 2012 19th International Conference on Telecommunications (ICT), IEEE, 2012, pp. 1–6.
[14] A. Udugama, K. Kuladinithi, C. Gorg, F. Pittmann, L. Tionardi, NetCAPE: enabling seamless IMS service delivery across heterogeneous mobile networks, Communications Magazine, IEEE 45 (7) (2007) 84–91.
[15] M. Boutabia, E. Abd-Elrahman, H. Afifi, A hybrid mobility mechanism for heterogeneous networks in IMS, 2010 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2010, pp. 1570–1575.
[16] A. García-Martínez, M. Bagnulo, I. Van Beijnum, The Shim6 architecture for IPv6 multihoming, Communications Magazine, IEEE 48 (9) (2010) 152–157.
[17] S.J. Koh, M.J. Chang, M. Lee, mSCTP for soft handover in transport layer, Communications Letters, IEEE 8 (3) (2004) 189–191.
[18] A. Achour, K. Haddadou, B. Kervella, G. Pujolle, A SIP-SHIM6-based solution providing interdomain service continuity in IMS-based networks, Communications Magazine, IEEE 50 (7) (2012) 109–119.
[19] N.H. Thanh, N.T. Hung, T.N. Lan, T.Q. Thanh, D. Hanh, T. Magedanz, mSCTP-based proxy in support of multimedia session continutity and QoS for IMS-based networks, Second International Conference on Communications and Electronics, ICCE 2008, IEEE, 2008, pp. 162–168.
[20] 3GPP, General Packet Radio Service (GPRS); GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface, TS 29.060, 3rd Generation Partnership Project (3GPP), March 2013.
[21] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, B. Patil, Proxy Mobile IPv6, RFC 5213, Internet Engineering Task Force, August 2008.
[22] H. Schulzrinne, E. Wedlund, Application-layer mobility using SIP, Service Portability and Virtual Customer Environments, IEEE, 2000, pp. 29–36.
[23] W. Wu, N. Banerjee, K. Basu, S.K. Das, SIP-based vertical handoff between WWANs and WLANs, Wireless Communications, IEEE 12 (3) (2005) 66–72.
[24] P. Bellavista, A. Corradi, L. Foschini, An IMS vertical handoff solution to dynamically adapt mobile multimedia services, IEEE Symposium on Computers and Communications, ISCC 2008, IEEE, 2008, pp. 764–771.
[25] D. Vali, S. Paskalis, A. Kaloxylos, L. Merakos, An efficient micro-mobility solution for SIP networks, Global Telecommunications Conference, GLOBECOM'03, vol. 6, IEEE, 2003, pp. 3088–3092.
[26] 3GPP, IP Multimedia Subsystem (IMS) Service Continuity; Stage 2, TS 23.237 version 11.6.0 Release 11, TS 23.237, 3rd Generation Partnership Project (3GPP), January 2013.
[27] K.O. Marungwana, N. Ventura, Performance evaluation of IMS session continuity signaling with heterogeneous access, Wireless Communications and Networking Conference (WCNC), IEEE, 2010, pp. 1–6.
[28] I. Vidal, I. Soto, M. Calderon, J. Garcia-Reinoso, V. Sandonis, Transparent network-assisted flow mobility for multimedia applications in IMS environments, Communications Magazine, IEEE 51 (7) (2013) 97–105.
[29] J. Rosenberg, H. Schulzrinne, P. Kyzivat, Indicating User Agent Capabilities in the Session Initiation Protocol (SIP), RFC 3840, Internet Engineering Task Force, August 2004.

[30] J. Rosenberg, The Session Initiation Protocol (SIP) UPDATE Method, RFC 3311, Internet Engineering Task Force, October 2002.
[31] B. Campbell, J. Rosenberg, H. Schulzrinne, C. Huitema, D. Gurle, Session Initiation Protocol (SIP) Extension for Instant Messaging, RFC 3428, Internet Engineering Task Force, December 2002.
[32] A. Roach, SIP-Specific Event Notification, RFC 6665, Internet Engineering Task Force, July 2012.
[33] ITU-T, ITU-T Recommendation G.114, One-Way Transmission Time, Standard G.114, ITU-T (International Telecommunication Union, Telecommunication Standardization Sector), February 1996.