Evaluating Extensions to IMS Session Setup for Multicast-based Many-to-Many Services

Ivan Vidal[a,*], Jaime Garcia-Reinoso[a], Ignacio Soto[a], Francisco Valera[a]

[a] *Departamento de Ingeniería Telemática, Universidad Carlos III de Madrid, Spain*

Abstract

Telecommunication networks are converging towards an all-IP paradigm that integrates a broad set of value-added services. In this context, the IP Multimedia Subsystem (IMS) is being developed by the 3GPP as a key element to achieve the convergence. Another development is that multiparty services are nowadays acquiring an increasing interest from the industry. In this respect, network multicast provides a cost-effective solution to deliver these services to the user. Nevertheless, although network multicast is being considered as an enabler for one-to-many services (e.g. IPTV) in the IMS, the specifications for many-to-many services still follow an unicast approach (e.g. push-to-talk and conference). This paper describes extensions to the session control procedures in the IMS, to support multicast based multi-user services. The idea was first described in a prior work, but this paper presents enhancements to provide a comprehensive solution and to improve the Grade of Service (GOS) perceived by the users. In addition, the GOS achieved by the proposal is evaluated. First, the bandwidth utilization for the multicast-based multi-user services is analyzed and compared against the unicast scenario. Next, the GOS is evaluated using an analytical approach, by obtaining the mathematical expressions for the session and user plane setup delays. Finally, the GOS is also evaluated using an experimental approach, and the results are compared with values recommended by the ITU-T.

Key words: IMS, Many-to-Many Multi-user services, Multicast, SIP

*Corresponding author

Email addresses: ividal@it.uc3m.es (Ivan Vidal), jgr@it.uc3m.es (Jaime Garcia-Reinoso), isoto@it.uc3m.es (Ignacio Soto), fvalera@it.uc3m.es (Francisco Valera)

1. Introduction

Telecommunication networks are evolving towards convergence using IP as the cornerstone technology. A key element to achieve the objective of having IP networks integrating any kind of services, including those traditionally provided through circuit switched networks, is the IP Multimedia Subsystem (IMS). IMS was initially developed by the 3GPP and then adopted by other standardization bodies (3GPP2, ETSI-TISPAN) for fixed and mobile networks. IMS is a signalling framework to provide multimedia services with QoS requirements over IP networks. It is important for operators because it enables a flexible framework to offer services in their networks and, at the same time, it allows operators to control those services and the resources used to provide them.

And among these services, multiparty applications, involving both communications one-to-many and many-to-many, are becoming increasingly important for operators. IP Television (IPTV), Video on Demand (VoD), videoconferencing, group communications, online gaming, or virtual worlds, are just some examples of services that are getting more and more prominent in telecommunication networks.

Network multicast is a bandwidth efficient way to provide multiparty services. Nevertheless network multicast solutions have been slowly introduced in commercial networks. The main reason is that the developed technical solution for network multicast did not consider issues that are important for a robust commercial implementation [1]. Nevertheless, due to the increasing interest in multiparty services and the need for efficient use of resources [2], operators have a strong motivation for the introduction of network multicast. In fact, this is the current trend with new multiparty services being offered through network multicast although in some cases multicast flows finish before reaching the customer premises. This is the case with experiences of IPTV over ADSL.

Multiparty services are also a topic of strong interest in IMS. Some of the early examples of services offered through the IMS, such as Push-to-Talk or conference are multiparty services. Nevertheless the model initially considered in the IMS for multiparty services is not based in network multicast. An application level server is placed in the middle of the communication, it receives media and duplicates it to the different destinations when needed. This seems to be unfortunate for two reasons:

- Multiparty applications in IMS could benefit from using network mul-

ticast to achieve significant bandwidth savings.

- IMS signalling provides a way to overcome the main issues with the deployment of network multicast solutions in commercial networks. Namely, it provides a way to control which users are authorized to send and receive traffic for particular applications, it allows controlling the resources that the network uses to serve each application, and it provides a rich event system that can be used to charge for services as required by the operator.

As a result, the use of IMS with one-to-many applications based on network multicast, in particular IPTV, has recently received a lot of interest [3, 4, 5]. This paper analyses the use of IMS in Many-to-Many services based on network multicast. In addition to the traditional multi-user services mentioned before, we envision an explosion of novel IMS-based Many-to-Many services that will be bandwidth intensive, such as shared virtual reality environments or shared augmented reality.

This article is organized as follows. Sect. 2 covers a brief overview on IMS and multi-user services. Section 3 presents extensions to the IMS session setup procedures to control many-to-many applications that use network multicast as transport. This idea was originally introduced in [6, 7], but here enhancements are presented to provide a comprehensive solution and to improve the grade of service perceived by the end users. Additionally, we have performed an evaluation of the proposed mechanisms from a theoretical and experimental perspective (Section 4). First, bandwidth savings for multicast-based multi-user services have been analyzed (in comparison with the unicast case). Next, the grade of service achieved by the proposed session establishment procedures is theoretically and experimentally estimated, and the obtained results are compared against a set of values recommended by ITU-T. Finally, Sect. 5 presents the main conclusions achieved along this work.

2. IMS and multi-user services

The development of Third Generation (3G) cellular networks has resulted in the deployment of new broadband wireless access technologies and enhanced terminals, paving the way towards a ubiquitous Internet. By means of a packet-switched domain, 3G terminals have IP connectivity to access to

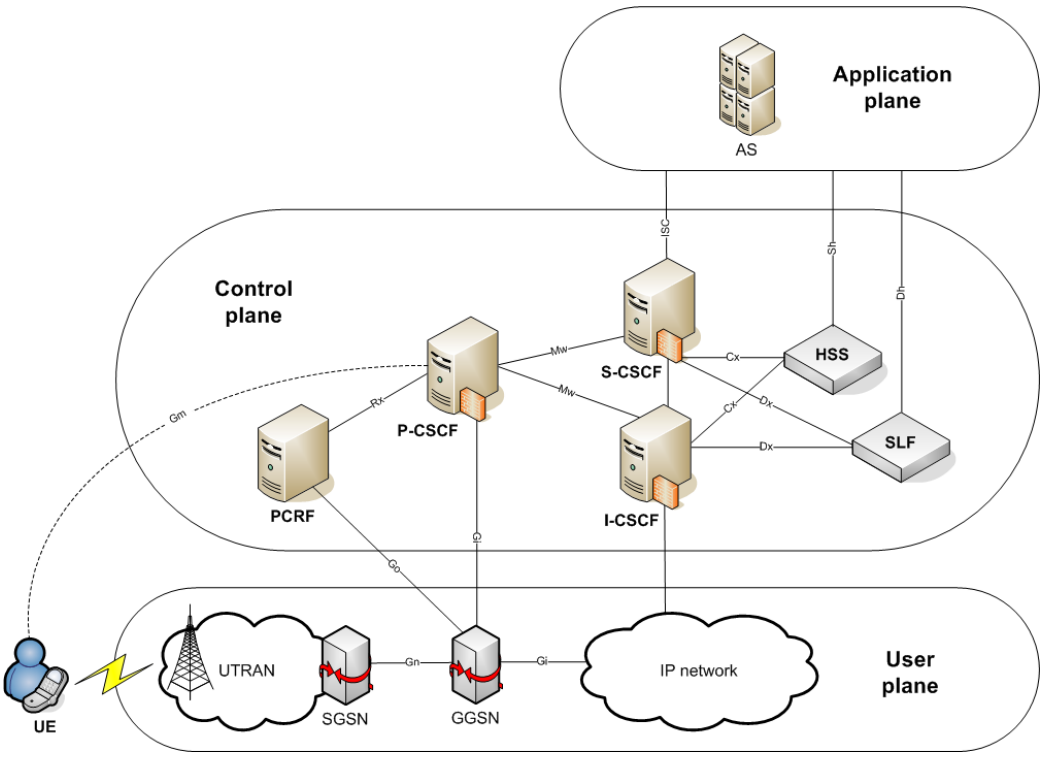


Figure 1: IMS architecture

the the broad set of services that are currently offered in the Internet, such as web browsing, email and video streaming. In this context, the IP Multimedia Subsystem (IMS) was introduced by 3GPP as part of the standardization process of the 3G UMTS technology. IMS is a key element in the UMTS architecture, that enables the delivery of the value-added multimedia services that are envisioned in the future of the Internet and the cellular worlds, by supporting facilities related to session control, QoS provision, charging, security and roaming. Figure 1 depicts a simplified overview of the IMS architecture [8], where an User Equipment (UE) is connected to the IMS by means of a UMTS access network, consisting of a UMTS terrestrial radio access network and the UMTS packet domain.

In this architecture, session control functionalities are provided by using the Session Initiation Protocol (SIP) [9]. In addition, the Session Description Protocol (SDP) [10] is used to describe multimedia sessions, and the Offer/Answer model with SDP [11] allows the different parties participating

in a session to reach an agreement on the session description to be established.

In this architecture, the Call Session Control Functions (CSCFs) acquire an special relevance. The CSCFs are the functional entities within the IMS in charge of processing the SIP signaling messages. The Proxy-CSCF (P-CSCF) is the entry point for the UE into the IMS network, and processes every SIP message that originates or terminates in the UE. The Interrogating-CSCF (I-CSCF) is the entry point in the user home network for every incoming session setup towards the UE. The Serving-CSCF (S-CSCF) performs functionalities related with session control and registration. In addition, this functional entity is in charge of routing the SIP signaling to one or more Application Servers (ASs) that provide services to the end user, such as Conference [12], Push-to-talk over Cellular (PoC) [13] or IPTV [5].

On the other hand, multi-user services are gaining attention from the industry and the standardization bodies. Nevertheless, the specifications related with the provision of many-to-many services over IMS still considers an unicast transmission in the user plane. In this approach, media is typically sent to a central node where it is replicated and forwarded to each UE that participate in the service. This solution can lead to an increment of the network load, and impact the scalability in terms of users and services. To address this issue, network multicast can be considered as a candidate alternative. As it has been pointed out in [7], the introduction of network multicast presents several advantages: better transmission efficiency in core and access networks, better scalability in terms of users and services, better fault tolerance (there is no central node to replicate the media) and augmented compatibility with multicast-based Internet services.

Nevertheless, delivering multicast-based many-to-many multimedia services requires to define session control procedures that, involving the participation of an arbitrary number of users, allow to establish a multicast multimedia session in the user plane among the participant users. These procedures should allow the participants to reach an agreement on the description of the multi-user session. In addition, as user terminals can support different capabilities (e.g. a terminal may or not integrate a video camera and support a restricted set of audio/video codecs), the session control procedures should enable each user to participate in the exchange of those media components that are supported by its terminal.

3. IMS signalling extensions to support Multicast-based Multi-user Services

This section briefly describes the session control procedures that have been defined to establish a multi-user multimedia session, and to setup the multicast based user plane. These procedures were presented in a prior work ([6] and [7]), but this section introduces enhancements to provide a comprehensive solution and to improve the grade of service perceived by the end users (these enhancements are covered in subsections 3.7, 3.8 and 3.9). In this section, it is assumed a 3GPP IP connectivity access network, where UEs need to perform a local resource reservation before exchanging media in the user plane (e.g. a UMTS terrestrial radio access network and the UMTS packet domain). Every UE that participates in the multi-user service has a dedicated bearer in the user plane, to support the exchange of SIP signaling messages with its corresponding P-CSCF. Considerations about other access network technologies are presented in Sect. 3.8.

As in a regular one-to-one IMS session, between two UEs, the first step in establishing a multi-user multimedia session is to setup a signaling relationship between the different parties that participate in the service. In this respect, a SIP dialog will be set with each UE. Jointly, these dialogs will provide this relationship, which will allow for negotiating the different parameters describing the multimedia session, performing session control functionalities (e.g. session establishment and release) and notifying the participant users about the session status.

3.1. Establishing a signaling relationship

To initiate the creation of the SIP dialogs, the initiator UE sends a SIP INVITE request towards the other UEs involved in the service (step 1 in Fig. 2). Following the routing procedures defined in IMS for SIP signalling (see [14]), the INVITE request reaches the P-CSCF and the S-CSCF. The S-CSCF matches the INVITE request against a set of filter criteria that correspond to the public user identity (this filter criteria are contained in the user profile that was obtained by the S-CSCF during the IMS registration process). As a result of this process, the S-CSCF verifies that the INVITE request must be processed by an Application Server that is specific to multi-user services, i.e. the Multiparty Application Server (**MAS**).

The MAS is a Back-to-Back User Agent, as it is defined in [9]. This Application Server is the core element of this proposal, being in charge of

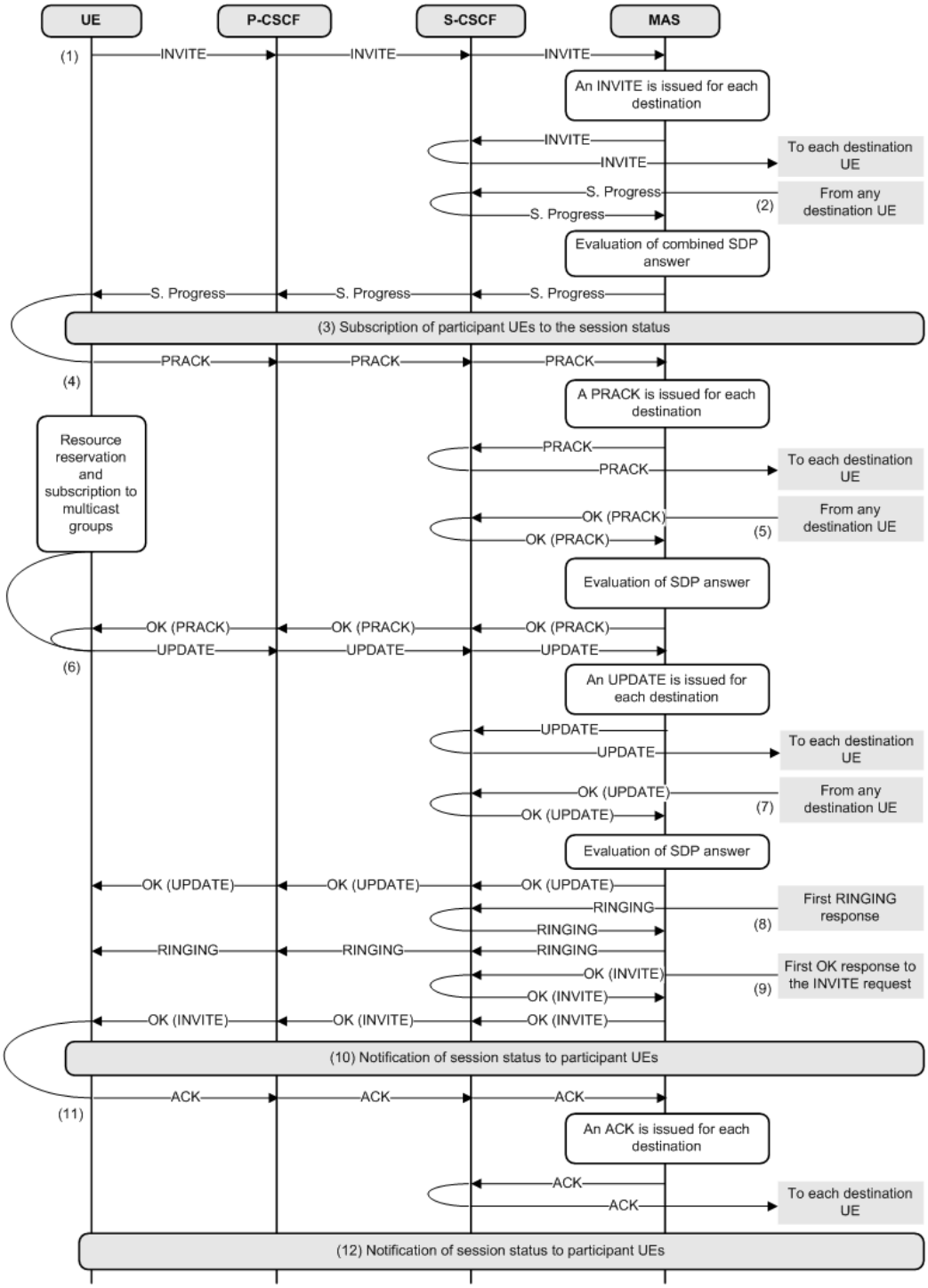


Figure 2: Session establishment, initiator side

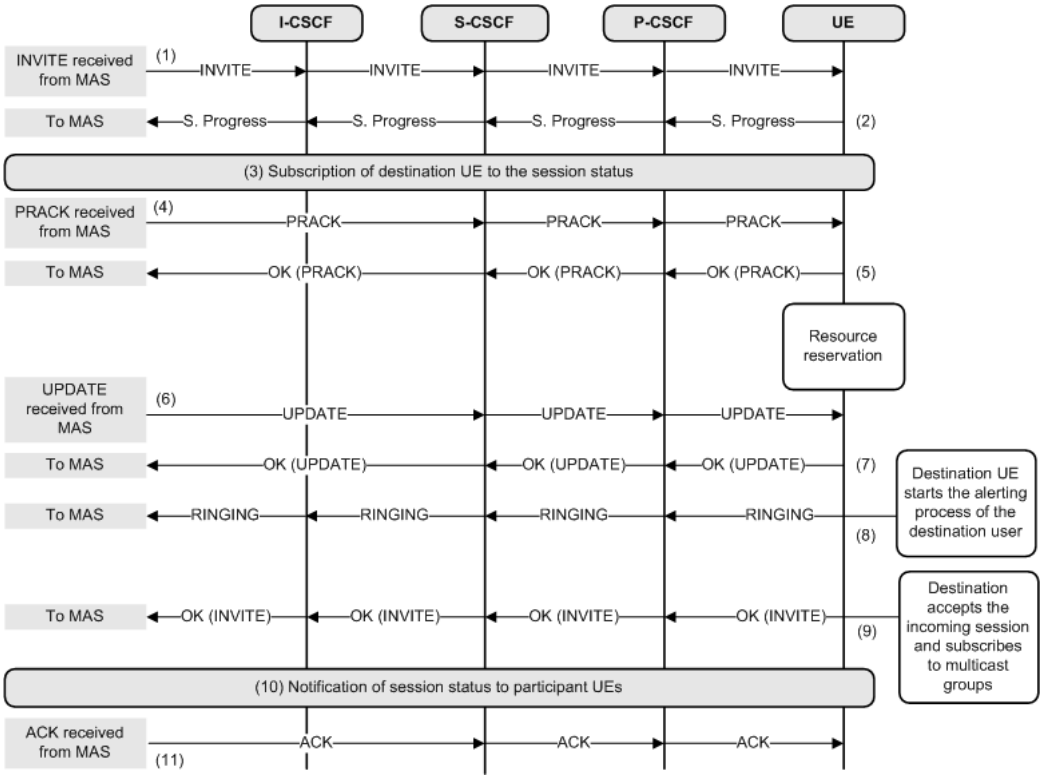


Figure 3: Session establishment, destination side

the following tasks (further detailed along this section):

- Extending the IMS control plane procedures for one-to-one sessions to the multi-user scenario.
- Handling the negotiation of the multi-user session description among the participant UEs.
- Administrating the multicast addressing space that will be used to deliver the different multi-user services.
- Transparently managing the QoS parameters that are exchanged within SDP payloads.
- Notifying session status to the participant users.

The MAS makes a copy of the INVITE request for each destination user, including in each copy the SIP header values that are necessary to receive further SIP requests and responses from the destination. Finally, each INVITE request is sent to the next hop towards the destination, i.e. the S-CSCF of the initiator user.

Upon receiving each of these INVITE requests, by following the procedures specified in [15] the S-CSCF obtains the addresses of a set of I-CSCFs in the home network of the destination user. The S-CSCF sends the request to one of them (step 1 in Fig. 3). Within the destination home network, the INVITE request is routed through the S-CSCF and P-CSCF corresponding to the receiver. Eventually, the request reaches the destination UE.

As every UE needs to perform a resource reservation in its local access network, the destination UE answers back the INVITE request with a SIP Session in Progress response (step 2 in Fig. 3). According to the regular SIP procedures, this response includes all the information needed to route it back to the MAS. As it will be explained next, the MAS waits to receive every Session in Progress responses from the destination UEs.The reception of each Session in Progress response implies the creation of a SIP dialog between the MAS and the responding UE. Eventually, all the responses corresponding to the INVITE request are received in the MAS[1], which sends a SIP Session in

[1]In case of an unreachable destination, a timer will fire within the INVITE transaction corresponding to the destination, and the B2BUA will be notified. This way, the B2BUA can continue the execution of the session setup procedures even in this case.

Progress response back to the initiator UE. Similarly, the reception of this response establishes a SIP dialog between the initiator UE and the MAS.

Therefore, after receiving the Session in Progress response at the initiator UE, a SIP dialog is established between the MAS and every participant UE. The MAS groups together all of these dialogs, that finally conform a signaling relationship between the different UEs that participate in the multi-user service. Any further SIP request, sent from a UE that participates in the session, will be sent within the dialog corresponding to the UE. Eventually, the request will reach the MAS, where it will be properly processed. This processing might involve sending new requests to a subset of participants in the session.

3.2. Negotiation of the session description

Before the session is established, the participant UEs must agree on which media components will be exchanged within the session (e.g. audio, video, etc), and on the different parameters that describe each of these media components. For this purpose, a procedure to negotiate the session description was designed, based on the SDP protocol [10] and on the Offer/Answer model of SDP [11]. The initial objectives that had to be fulfilled by the design of this procedure were the following:

- The initiator UE should be capable of indicating which media components are allowed in the context of the multi-user service.
- Each destination UE should be capable of indicating which media components, out of those indicated by the initiator, will exchange with the rest of participants during the execution of the service. The set of accepted media components does not need to be the same for every destination UE.
- Media will be transmitted in the user plane by means of network layer multicast. Therefore, each multimedia session must be provided with a set of multicast IP addresses. The assignment policy of multicast addresses should guarantee a coherent use of the available addressing space, enabling each participant UE to receive only the multicast media it has accepted during the negotiation phase.

As an example, suppose that a certain user initiates a videoconference, inviting some other users to participate. If the negotiation of the session

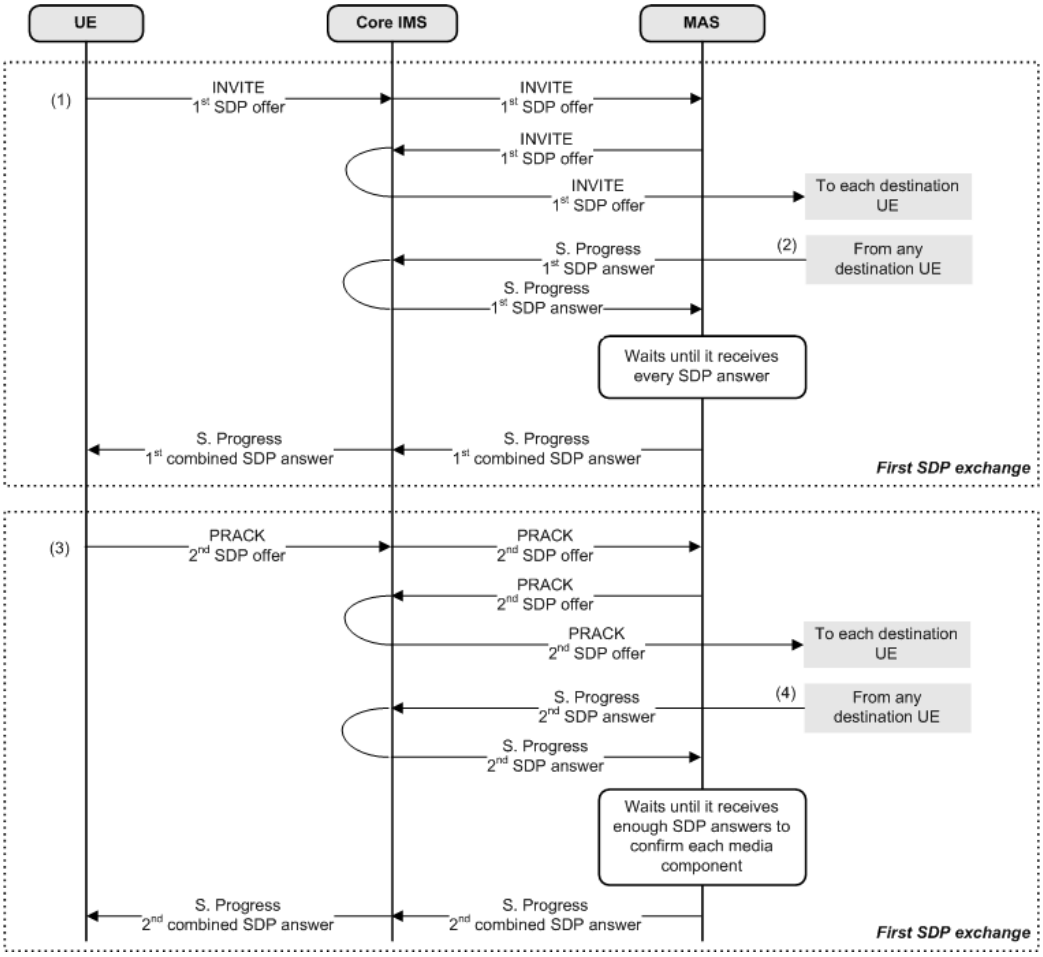


Figure 4: Session description negotiation

description taking place during the session establishment fulfills the previous objectives, then it is still possible for one UE that does not integrate a video camera to participate in the audio communication within the videoconference.

The procedure to negotiate the session description is schematized in Fig. 4, and consists of two SDP Offer/Answer exchanges:

(1) The initiator UE includes an SDP offer in the INVITE request. This offer contains the description of the different media components (e.g. audio or video) that the initiator wants to exchange within the multi-user multimedia session. This description includes, for each proposed media

component, its associated bandwidth requirements, addressing information (i.e. the transport port where the multicast media is to be received) and a set of formats (e.g. codecs) that are supported by the initiator side.

The MAS appends to each media component a specific multicast IP address. Therefore, a destination UE that accepts the exchange of a given media component can subscribe to its corresponding multicast group, and start receiving the data traffic associated with the component. The modified SDP offer is included in every copy of the INVITE request that is sent towards the set of destination UEs.

(2) Each destination UE answers back the SDP offer with an SDP answer, that is included in a SIP Session in Progress response. In this answer the UE can discard any proposed media component (e.g. a video component in case that the UE does not integrate video facilities). For each accepted component, the UE indicates the subset of supported formats out of those proposed in the received offer, and keeps unchanged the SDP parameters related with bandwidth requirements and addressing information.

Eventually, all the SDP answers are received by the MAS. At this point, the MAS generates a combined SDP answer for the initiator that reflects a consistent view of the multiparty session. In this answer, each media component proposed by the initiator will be accepted providing that it has been accepted by at least one destination UE. For each accepted media component, the MAS will include only the subset of formats accepted by all the destination UEs that have agreed to participate in the exchange of the component. If there are no formats in common for a specific media component, then the component is discarded from the SDP answer. Bandwidth requirements and addressing information are kept unchanged for every media component. The SDP answer is encapsulated in SIP Session in Progress response, that is finally sent to the initiator UE.

(3) After receiving the combined SDP answer, the initiator UE takes a decision about the specific format that will be used for each accepted media component. The UE indicates this information in a second SDP offer, that is sent in a SIP PRACK request towards the MAS (step 4 in Fig. 2).

This way, the negotiation requires a new SDP Offer/Answer exchange to take place. This is necessary because, if several formats are feasible, a resource reservation would be done to accommodate the most restrictive one, being possibly another format finally utilized.

Eventually, this SDP offer is received at the MAS, which in turn generates an SDP offer for each destination UE that is encapsulated in a new PRACK request (step 4 Fig. 3). Each new SDP offer proposes those media components that, having been proposed by the initiator, were accepted in the first SDP answer received from the destination. The bandwidth requirements and addressing information are left unchanged in the offer.

(4) Finally, each destination answers back the second SDP offer with a new SDP answer that is included in a SIP OK response (step 5 in Figs. 2 and 3). This answer accepts each proposed media component, leaving unchanged the bandwidth requirements and the addressing information.

The MAS waits until it receives enough SDP answers so as to confirm every media component that was proposed by the initiator in the second SDP offer. At this point, the MAS generates a second combined SDP answer, that is sent back to the initiator UE encapsulated in a SIP OK response. In case that after a predefined timeout the MAS cannot confirm every media component, it assumes that the communication path with the destination UEs that accepted the media component is broken, and the second combined SDP answer discards these media components. In addition, this subset of UEs is removed from the session status information.

As an improvement to this proposal, in case there are no formats in common for a given media component when generating the first combined SDP answer at the MAS, the MAS could select the subset of formats that allows maximizing the number of destination UEs capable of participating in the exchange of the component. Another possibility could be to introduce transcoding facilities in the user plane. This issue is, however, out of the scope of this paper.

3.3. Integrating the resource reservation

As a result of the negotiation of the session description, each UE obtains the different parameters of the media components it is going to exchange, such as the required bandwidth and the multicast addressing information. Nevertheless, to guarantee that each media component receives an appropriate end-to-end treatment in the user plane, some resource reservation process must be executed. In the scenario that has been considered in this section, it is assumed that this procedure is separately initiated by each UE, and results in the establishment of a set of transport bearers in its local access network (e.g. a set of PDP contexts in the case of UMTS). In order to achieve an efficient utilization of the QoS resources in the user plane, every UE needs to activate the following transport bearers per each media component:

- An uplink transport bearer, to transmit the multicast traffic associated with the media component from the UE to the access network.
- A downlink transport bearer, to deliver the multicast traffic associated with the media component from the access network to the UE. This transport bearer could be partially or totally shared by several UEs, depending on its location in the access network. For instance, in the case of UMTS, a shared PDP context would be activated and shared among all the UEs served by the same GGSN. This way, the multicast traffic would be efficiently transmitted in the downlink direction, from the GGSN to the UEs, by means of shared GTP tunnels and point-to-point and/or point-to-multipoint radio bearers[2].

Nevertheless, it must be taken into account that the process of establishing a transport bearer may fail, for instance due to resource availability constraints in the access network of the UE. That implies that the multimedia session cannot be established by a UE until it has successfully finished its local resource reservation process. In general, it can be stated that the activation of transport bearers must be achieved by the initiator UE and by at least one destination UE before alerting the corresponding destination user about the incoming session. This way it is guaranteed that, when a destination user is alerted, an adequate end-to-end resource reservation has been

[2]The use of a shared bearer plane in UMTS has been proposed in [16] for the Multimedia Broadcast/Multicast Service (MBMS)

configured between the initiator and destination UEs, and the multimedia session can be established between both UEs.

In a regular one-to-one IMS session this restriction holds as well, and it is enforced by the utilization of the precondition framework defined for the SIP protocol (see [17] and [18]). This framework will be used in the multi-user scenario, according to the following guidelines:

- The initiator UE includes QoS preconditions in the initial INVITE request, indicating that a resource reservation is needed in its local access for every proposed media component. These QoS preconditions are left unchanged in all the copies of the INVITE request that the MAS sends towards the destination UEs.

- Each destination includes QoS preconditions in the Session in Progress response, indicating that a resource reservation is also needed in its local access for every media component. In addition, these preconditions indicate that the destinations want to receive a confirmation when the resource reservation finishes in the local access. The MAS keeps these QoS preconditions in the first combined SDP answer.

- QoS preconditions are also included in the second SDP Offer/Answer exchange, according to the procedures specified in [17] and [18].

- Eventually, the initiator UE succeeds to activate the necessary PDP contexts. In this case, it generates a third SDP offer with QoS preconditions, indicating that the resource reservation has finalized within its local access. This SDP offer is included in a SIP UPDATE request (step 6 in Fig. 2) that is sent towards the MAS. The MAS sends a new UPDATE request for every destination UE that has previously sent an OK response to the PRACK request (step 6 in Fig. 3).

- Finally, each destination UE answers back the SDP offer with a new SDP answer, where QoS preconditions are included to indicate the status of its local resource reservation (that may or may not have finalized).

After receiving the UPDATE request, and once that the destination UE finishes its local resource reservation, it can resume the establishment of the multimedia session.

3.4. Alerting the destination UE

At this point, the destination UE can optionally start alerting its destination user about the incoming session (e.g. by playing some ringtone). In this case, the UE sends a SIP RINGING response towards the MAS (step 8 in Fig. 3). This RINGING response means that the destination UE has successfully finished the resource reservation process, and the destination user has been prompted to accept the session establishment. After receiving the first RINGING response (step 8 in Fig. 2), the MAS generates a new RINGING message that is sent to the initiator UE. From the point of view of the initiator, this response means that at least one destination UE has completed the resource reservation process, and that one destination user is capable of accepting the multiparty multimedia session. New RINGING responses cause no further processing at the MAS.

3.5. Accepting the session establishment

Finally, when the multimedia session is accepted by any destination user (e.g. by pressing the accept button in its IMS terminal), the destination UE answers back the INVITE request with an SIP OK response (step 9 in Fig. 3). After receiving the first OK response (step 9 in Fig. 2), the MAS generates a new OK response for the initiator UE. Eventually, this response reaches the initiator, which confirms the reception by means of a SIP ACK request (step 11 in Fig. 2). When the ACK request reaches the MAS, it generates and sends a new ACK request for every OK response that was received to an INVITE request.

3.6. Subscription to multicast groups

After finalizing the resource reservation, the initiator UE can subscribe to the multicast groups corresponding to those media components related to the user traffic it will receive. For this purpose, the UE utilizes the IGMP protocol [19]. This protocol supports the multicast group management between any UE and its corresponding GGSN. The exchange of IGMP messages is done by means of the transport bearer dedicated to signaling.

In addition, the GGSN needs to execute some multicast routing protocol (e.g. PIM-SM), to enable the reception of the media corresponding to the multicast groups subscribed by the UEs.

On the other hand, each destination UE subscribes to the multicast groups corresponding to the accepted media components after sending the

OK response to the INVITE request. This way, the UE prevents the reception of multicast traffic until its associated user has accepted to participate in the multimedia session.

3.7. Management of the session status

Once the session has been established between the initiator UE and a certain destination UE, the multicast traffic corresponding to the media components that were accepted by the destination can flow through the user plane between both entities. Nevertheless, as the OK response received by the initiator UE does not contain an SDP payload, at this point the initiator is not aware of which media components were accepted by the destination.

To address this problem, the SIP specific event notification framework is used [20]. This framework allows an entity to subscribe to the state information associated with a given resource, by means of a SIP SUBSCRIBE request. After receiving the request, the entity that keeps track of this information sends a SIP NOTIFY request, returning the required state information to the subscriber. If the resource state changes, a new NOTIFY request is sent from the notifier to the subscriber. For each type of state that may be associated with a given resource, an event package is defined. An event package defines the format and the semantics associated with state information that is included in the body of a NOTIFY request. In concrete, [21] describes an event package that allows the participants in a tightly coupled conference to receive state information associated with the conference.

In the solution presented so far, the MAS maintains updated information about the session state, such as the number of participants, their state (e.g. establishing the session, session established, disconnected from the session, etc.), and the media components that will be exchanged with all of them. Therefore, the MAS can provide state information about the multi-user multimedia session to every involved UE. In this respect, a simplified version of the event package defined in [21] is used, enabling three fundamental activities for the appropriate execution of a multi-user service:

- To initiate or maintain the transmission of multicast traffic in the user plane, for media components with active receivers (i.e. UEs that have accepted the session establishment and have agreed to receive the media component).
- To stop the transmission of multicast traffic in the user plane, for media components without active receivers.

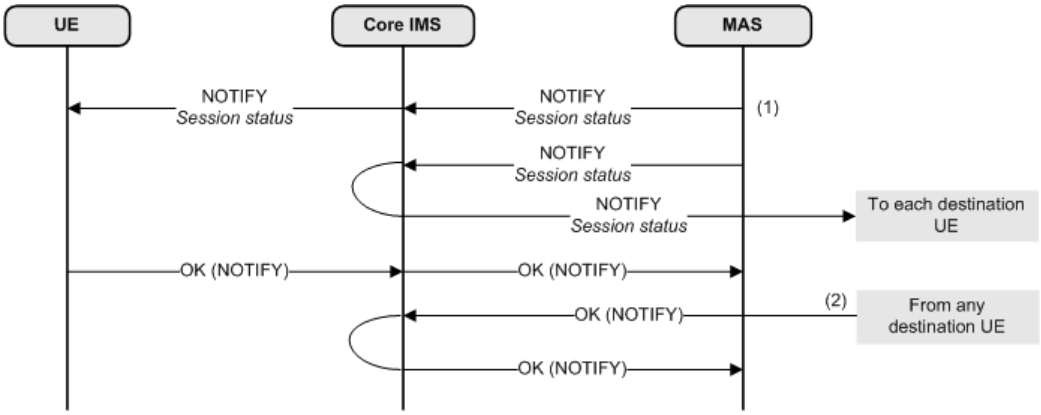


Figure 5: Notification of session status

- To free the QoS resources associated with a media component, if there are no participant UEs in the session that have agreed to exchange the component (e.g. after the session termination from all the UEs that accepted to exchange the media component).

This way, once that the MAS receives all the Session in Progress responses from the destination UE, it builds up the state information for the multi-user multimedia session (step 3 in Fig. 2). The reception of all of these SIP responses involve the subscription of the initiator UE and the responding UEs to the state information associated with the session. Consequently, the state information is included in a NOTIFY request that is sent to each participant UE (see Fig. 5). Further changes on the session status, such as those due to UEs accepting, rejecting or terminating the multimedia session, will be notified to the participants (these NOTIFY messages will only contain the state information that has changed since the last notification). The notification process corresponds to steps 3, 10 and 12 in Figs. 2 and 3. This way, UEs always have updated information about the session status, being capable of performing the activities that were previously presented.

In the prior proposal ([6] and [7]), only those UEs that had accepted the session establishment were subscribed to the status information. This approach had the inconvenient that UEs were notified for the first time about the session status after accepting its establishment. As the first NOTIFY request contains all the state information concerning the multimedia session, and this information may get a considerable size as the number of participants increases, the transmission of this SIP request may lead to significant

delays when obtaining the state information. This is particularly sensitive when notifying the initiator UE, as the time interval that elapses from the transmission of the OK response at the destination UE to the reception of the corresponding NOTIFY request at the initiator UE, is a parameter that measures the grade of service offered by the proposal (the values achieved for this parameter by the procedures here described are evaluated in the next section).

An additional benefit of the new scheme is that, by notifying every UE that participates in the session, application instances running on the UE can display to the user the session related information at any stage during its establishment. This information could be used by a destination user to decide whether to join or decline the invitation to participate in the service when being alerted of the incoming session.

3.8. Considerations about the access network technology

It this section, it has been assumed that each UE has a 3GPP IP connectivity access network, where each UE needs to perform a local resource reservation before completing the session setup.

If this is not the case (e.g. in case of a DSL access), initiator and destination UEs would indicate in the QoS preconditions that the resource reservation is not needed it their local accesses for every media component. As UEs are not responsible for executing the resource reservation, each terminal could send the SIP RINGING response towards the MAS after receiving the SIP INVITE request. Nevertheless, a second SDP Offer/Answer exchange is still necessary, to allow the initiator UE to choose a single format for each proposed media component. This way, the session description negotiation still follows the scheme depicted in Fig. 4.

Finally, after sending the OK response to the PRACK request, the UE can start alerting the destination user and can send the RINGING response. The UPDATE transaction is no longer necessary, as the initiator UE is not in charge of executing a resource reservation procedure, and then it does not need to confirm its finalization. Therefore, the session setup procedure would be the same indicated in Fig. 2 and 3, without the signaling flow corresponding to the UPDATE transactions.

3.9. Applicability of the proposal

This subsection analyses the scope and limitations of the procedures that have been presented so far. This proposal allows to establish a multicast-

based multi-user session among a set of participants, in order to execute a many-to-many multimedia service. Nevertheless, although the procedures have conceptually been designed to be valid for any number of participants, the peer-to-peer nature of many-to-many services may impose certain limits on the reasonable number of users. For example, in a videoconference service, the utility (from the end user perspective) decreases if the number of participants becomes too high, because only a limited number of users can speak at the same time and the number of video streams may become unmanageable within the end user display). This way, we envision that this proposal will be utilized to establish multimedia sessions that do not involve a large number of users.

On the other hand, one specific aspect that may limit the applicability of the proposal is related with the assignment of multicast addresses. The procedures described along this section assume that the MAS assigns a multicast address to every media component within the multimedia session. This requires a large range of available multicast addresses at the MAS, which may be dynamically assigned to end users for a limited period of time (i.e. the duration of the multimedia session). In addition, the multicast address assigned to an specific media component should be globally routable if the participants are located in different network domains.

According to IANA guidelines for IPv4 multicast address assignments [22], the blocks of addresses that could be used in this proposal are the *GLOP* block and the *administratively scoped* block. Out of these, the BLOCK block is the only one that contains globally scoped addresses. Nevertheless, the subset of multicast addresses from this block that can be assigned to a given domain is too small to support the address assignment procedures. Therefore, the administratively scoped address block must be used instead.

The challenge now is that, although this block of multicast addresses (239.0.0.0/8) can in principle be sufficient, it is for local use within a domain (i.e. addresses from this block are not globally routable). So, addresses within this block can be utilized providing that all the participant users are located in the same network domain, and new mechanisms must be designed in order to cover the scenario where participant users are located in different network domains. However, these mechanisms are out of the scope of this paper.

Finally, the procedures that have been presented in this section cover the establishment of ad-hoc multimedia sessions, which are initiated by one of the participants. Supporting multi-user multimedia sessions where authorized

participants gradually join the session is a matter of future research.

4. Evaluation of the proposed mechanisms

This section evaluates the main benefits derived from the presented proposal. First, the bandwidth savings in the user plane for multicast-based multi-user services multicast are analyzed. Next, the grade of service achieved by the session establishment procedures that have been described is estimated, and compared against a set of recommended values by ITU-T. It will be assumed along this section that every UE is connected to a 3GPP IP Connectivity Access Network, consisting of a UMTS terrestrial access network and the UMTS packet domain.

4.1. Analysis of bandwidth savings

This subsection analyses the bandwidth utilization corresponding to the multicast-based transmission for multiparty services that is proposed in this paper. The results achieved in the multicast case are compared against the bandwidth utilization results that can be obtained when a unicast-based transmission, that follows the current IMS related specifications for multiparty services, is implemented in the user plane. To address this theoretical analysis, the network model illustrated in Fig. 6 has been used.

In this network model, is is assumed that a certain number N of UEs participate in a given multiparty service. This service involves the exchange of a single media component (e.g. audio or video) among all the participant UEs. Under this scenario, it is possible for several UEs to be served by a common RNC, SGSN or GGSN. The GGSNs in the UMTS packet domain are interconnected by means of an IP network (from here on, the IP core), this way ensuring the global connectivity among all the participant UEs. It is also assumed that, out of the N participants in the service, N_s UEs behave as data sources, by sending the media components to all the UEs.

In the **unicast case**, it is assumed that the data traffic is sent from each source UE to a central node, where it is replicated and delivered by means of a unicast-based transmission to the rest of the UEs that participate in the multiparty service. In this scenario, the bandwidth consumption can easily be evaluated, by decomposing the theoretical calculation into two contributions that correspond to the IP core and the UMTS access network:

IP core. In this case, whenever a flow coming from a source enters into the network, it is routed towards the central node, where it is replicated for

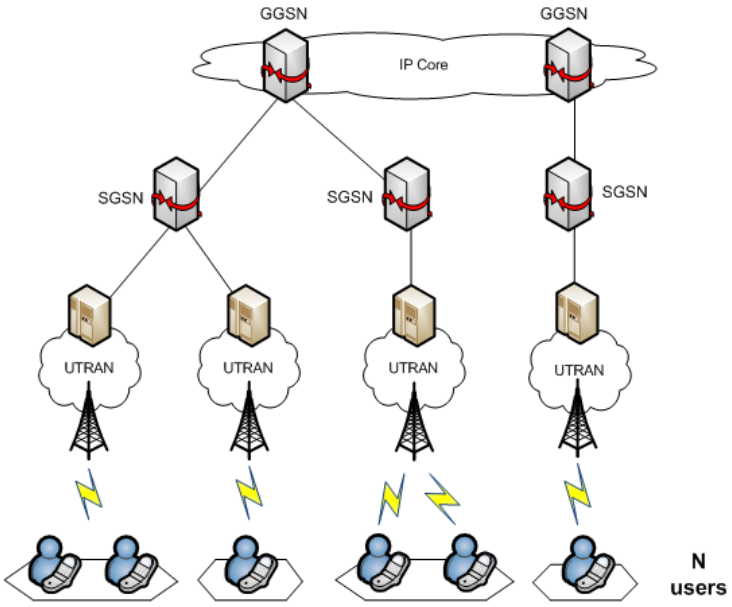


Figure 6: Network model

$N-1$ destinations (all the UEs but the source). Each replica is finally sent towards its corresponding destination. This way, the number of flows that are carried by the IP core for each source is $N-1$ plus the flow coming from the source. Assuming that the average length of the path between any of two edges of the IP core is L_p, and that the bandwidth that is required for each flow, from the media component perspective, is B_0, the total bandwidth utilization in the IP core can be expressed as:

$$BW_{unicast}^{core} = N \cdot L_p \cdot B_0 \cdot N_s \tag{1}$$

UMTS access network. In this case, the bandwidth consumption can be further decomposed into two new components, the bandwidth utilization in the uplink direction (from the UE to its serving GGSN) and in the downlink direction (from the GGSN to the UE):

- Uplink: each flow transmitted from any of the N_s sources will require bandwidth resources at the radio access (from the UE to the RNC), at the interface from the RNC to the SGSN (in the

form of a GTP tunnel) and at the interface from SGSN to GGSN (in the form of a GTP tunnel). This bandwidth requirements can be expressed, from the media component perspective, as $3 \cdot B_0$, leading to a total uplink bandwidth utilization of:

$$BW_{unicast}^{ANup} = 3 \cdot B_0 \cdot N_s \tag{2}$$

- Downlink: each source UE will receive $N_s - 1$ flows, while the $N - N_s$ remaining UEs (i.e. the destinations that do not behave as data sources) will receive N_s flows. Similarly to the uplink case, the bandwidth requirements for each downlink flow can be expressed as $3 \cdot B_0$. Therefore, the total downlink bandwidth utilization can be described as:

$$\begin{aligned} BW_{unicast}^{ANdown} &= 3 \cdot B_0 \cdot [(N - N_S) \cdot N_s + N_s \cdot (N_s - 1)] \\ &= 3 \cdot B_0 \cdot N_s \cdot (N - 1) \end{aligned} \tag{3}$$

Bandwidth utilization . Combining equations 1, 2 and 3, the total amount of bandwidth that is utilized in the unicast approach can be estimated as:

$$\begin{aligned} BW_{unicast} &= N \cdot L_p \cdot B_0 \cdot N_s + 3 \cdot B_0 \cdot N_s \cdot N \\ &= N_s \cdot B_0 \cdot N(L_p + 3) \end{aligned} \tag{4}$$

In the **multicast case** data traffic is sent from each source UE to its associated GGSN. From this point the traffic is efficiently distributed through the IP core by means of a multicast-based transmission. This way, traffic replication is performed when it is strictly necessary. In the UMTS access network, traffic will be sent to the destination UEs by means of shared GTP tunnels and radio bearers. Again, the theoretical calculation of the bandwidth consumption is decomposed into the two contributions corresponding to the IP core and the UMTS access network:

IP core. When a flow that comes from a source UE enters into the network, it will be routed towards all the GGSNs associated with the participant UEs, with the exception of the GGSN serving the source. Let

N_{GGSN} be the average number of users served by one GGSN. The relation N/N_{GGSN} corresponds to the average number of GGSNs that are utililized. The number of links, N_L, that are used in the IP core to deliver the flow will fulfill the following condition:

$$N_L \leq \left[\frac{N}{N_{GGSN}} - 1\right] \cdot L_p \tag{5}$$

That is, in the worst case there are no common links to route the flow from the source GGSN to all the destinations, and each data transmission must follow a separate path. This leads to a number of links that equals the number of destinations times the average length of the path between any two edges of the IP core. Typically, the number of links will be below this value, depending on the IP core topology. Taking into consideration the number of source UEs, N_s and the bandwidth requirements for each media component, B_0, the bandwidth consumption in the IP core can be expressed by the following inequality:

$$BW^{core}_{multicast} \leq \left[\frac{N}{N_{GGSN}} - 1\right] \cdot L_p \cdot B_0 \cdot N_s \tag{6}$$

UMTS access network. Again, the bandwidth consumption can be decomposed into two components, the bandwidth utilization in the uplink and downlink directions:

- Uplink: in this case, the bandwidth consumption is the same as in the unicast case, being determined by Eq. 2.
- Downlink: each GGSN that serves participant UEs will receive N_s flows. Let N_{SGSN} be the average number of participants that are served by a single SGSN. Under this assumption, the relation N/N_{SGSN} corresponds to the average number of SGSNs that are utilized. On the other hand, each SGSN will receive N_s flows, being all of them carried by means of a GTP tunnel[3] from its corresponding GGSN. The bandwidth requirements for each flow, from the media component perspective, can be expressed as B_0.

[3]A single GTP tunnel, from the GGSN to the SGSN, would be shared among all the UEs served by the SGSN (seeSect. 3.3)

Taking this into account, the downlink bandwidth utilization in the interface between the GGSNs and the SGSNs can be expressed as:

$$\frac{N}{N_{SGSN}} \cdot B_0 \cdot N_s \tag{7}$$

Now, let N_{RNC} be the average number of participants that are served by a single RNC. Analogously to the previous case, the relation N/N_{RNC} corresponds to the average number of RNCs that are used. Each of them will receive N_s flows, carried by N_s GTP tunnels from its associated SGSN. Again, the bandwidth requirements for each flow, at the interface between the SGSN and the RNC, can be expressed as B_0. Let k be the number of radio bearers that are necessary to transmit each flow to the users served by the RNC, with $1 \leq k \leq N_{RNC}$. In this case, the downlink bandwidth utilization, taking into account the radio interfaces and the interface between the SGSNs and the RNCs, is calculated as:

$$\frac{N}{N_{RNC}} \cdot B_0 \cdot (1+k) \cdot N_s \tag{8}$$

Therefore, by combining equations 7 and 8, the downlink bandwidth consumption in the multicast case can be estimated as follows:

$$BW_{multicast}^{ANdown} = N_s \cdot B_0 \cdot N \cdot \left[\frac{1}{N_{SGSN}} + \frac{1+k}{N_{RNC}}\right] \tag{9}$$

Bandwidth utilization. Combining equations 6, 2 and 9, the total amount of bandwidth that is utilized in the multicast approach follows this expression:

$$\begin{aligned} BW_{multicast} = {} & N_s \cdot B_0 \cdot N \cdot L_p \cdot \left(\frac{1}{N_{GGSN}} - \frac{1}{N}\right) \\ & + N_s \cdot B_0 \cdot N \cdot \left(\frac{3}{N} + \frac{1}{N_{SGSN}} + \frac{1+k}{N_{RNC}}\right) \end{aligned} \tag{10}$$

where $1 \leq k \leq N_{RNC}$

Comparing the resulting equations for the unicast and the multicast cases (Eqs. 4 and 10, respectively), it is clear that the bandwidth utilization will be lower in the multicast case providing that the following two inequalities are fulfilled:

$$\frac{1}{N_{GGSN}} - \frac{1}{N} < 1 \tag{11}$$

$$\frac{3}{N} + \frac{1}{N_{SGSN}} + \frac{1+k}{N_{RNC}} < 3 \tag{12}$$

Inequality 11 can be easily demonstrated, as $N_{GGSN} \geq 1$ and $N > 1$. Regarding to inequality 12, as $N_{SGSN} \geq 1$, $N_{RNC} \geq 1$, and $k \leq N_{RNC}$:

$$K \leq N_{RNC} \Rightarrow \frac{3}{N} + \frac{1}{N_{SGSN}} + \frac{1+k}{N_{RNC}} \leq \frac{3}{N} + \frac{1}{N_{SGSN}} + \frac{1+N_{RNC}}{N_{RNC}}$$

$$N_{SGSN} \geq 1, N_{RNC} \geq 1 \Rightarrow \frac{3}{N} + \frac{1}{N_{SGSN}} + \frac{1}{N_{RNC}} + 1 \leq \frac{3}{N} + 3$$

In the previous inequation, the term $3/N$ comes from the fact that in the mathematical equations presented for the multicast case, it has been considered, for the sake of simplicity, that the multicast traffic sent from each source is delivered to every participant UE, and no mechanism is provisioned to prevent a source from receiving back the traffic that it sends towards the GGSN. Under a strict implementation, this mechanism would nonetheless be implemented in the user plane, and therefore the following condition would be satisfied:

$$\frac{3}{N} + \frac{1}{N_{SGSN}} + \frac{1+k}{N_{RNC}} \leq 3 \tag{13}$$

Therefore, under a strict implementation Eq. 10 would always be lower than Eq. 4, and the utilization of the multicast based approach would always introduce benefits to the operator in terms of bandwidth consumption. Anyway, even if no mechanism is put in place to prevent a source UE from receiving its own data traffic, the bandwidth utilization would still be lower in the multicast case providing that several users share the same GGSN, SGSN and/or RNC (i.e. if values for N_{GGSN}, N_{SGSN} and N_{RNC} are greater than 1). This will be illustrated next with a set of examples. In all of them, it is assumed a population of 20 users ($N = 20$), where only three of them at

a time can behave as sources for data traffic ($N_s = 3$). In addition, a typical value of 2.5 has been considered for L_p:

Example 1. All the UEs are located in the same UMTS cell (this is the best case for the multicast approach).

- Unicast case: substituting the proposed values for N, N_s and L_p in Eq. 4, the following bandwidth utilization can be obtained:

$$BW = 330 \cdot B_0$$

- Multicast case: in this scenario, $N_{GGSN} = N_{SGSN} = N_{RNC} = 20$. Assuming that all the users share a radio bearer in the user plane, then $k = 1$, and according to Eq. 10:

$$BW = 18 \cdot B_0$$

  Comparing both values, it can be seen that in the multicast case the bandwidth consumption is 18.3 times lower than in the unicast case.

Example 2. Each UE is served by one GGSN.

This is the worst scenario for the multicast case (the results for the unicast case are the same as in example 1, as there are no resources shared by UEs in their UMTS access networks). In this scenario, $N_{GGSN} = N_{SGSN} = N_{RNC} = k = 1$. Subsituting these values in Eq. 10:

$$BW = 331.5 \cdot B_0$$

Comparing this value with the one obtained in the unicast case, it can be seen that the bandwidth utilization is slightly greater in the multicast case (the relation between both values is 1.0045). The reason for this is that the analysis does not consider that each source will not be interested in receiving its own data traffic, and the bandwidth consumption in the UMTS access network seems slightly greater for the multicast case. Nevertheless, it can be seen from the equations that the bandwidth utilization in the multicast case is lower within the IP core. Jointly, these effects make the bandwidth utilization be in practical terms identical in both cases, or slightly lower for the unicast case.

Example 3. A typical use case.

In this example, a typical scenario that has been envisioned in this proposal is considered. In this scenario, it is assumed that all the UEs are connected to the same GGSN, and certain degree of sharing will be allowed for SGSNs and RNCs. In concrete, it will be considered that $N_{SGSN} = 4$ and $N_{RNC} = 2$. A value of $k = 1.8$ will be used, this way allowing a minority of UEs in the same cell to share the same radio bearer. Under these assumptions, and according to Eq. 10, the bandwidth consumption in the multicast case is:

$$BW = 108 \cdot B_0$$

Comparing this value to the one obtained in the unicast case (the bandwidth utilization for the unicast case was obtained in example 1), it can be seen that the bandwidth consumption in the multicast case is approximately 3 times less than in the unicast case, representing significant resource savings in the user plane.

Sometimes, due to the nature of certain media components, it might be possible to perform mixing operations, so that flows coming from different sources can be combined into one single flow carrying the mixed media (for instance, this functionality could be provided by an RTP mixer that receives several audio flows in the same RTP session). If media mixing is feasible and is implemented in the central node of the unicast approach, then the number of flows traversing the IP core and the UMTS access networks can be decreased. This is because, the central node could combine the flows that are received from the N_s sources, and send a single flow to each UE that participates in the service. Similarly to the unicast and multicast cases presented so far, it can be easily deduced that, in this situation, the total bandwidth utilization in the IP core and UMTS access networks can be expressed by the following equation:

$$BW = L_p \cdot B_0 \cdot (N + N_s) + 3 \cdot B_0 \cdot (N + N_s) \tag{14}$$

Substituting the values indicated for the previous examples into this new equation, the bandwidth consumption under the assumption of media mixing can be estimated as:

$$BW = 126.5 \cdot B_0$$

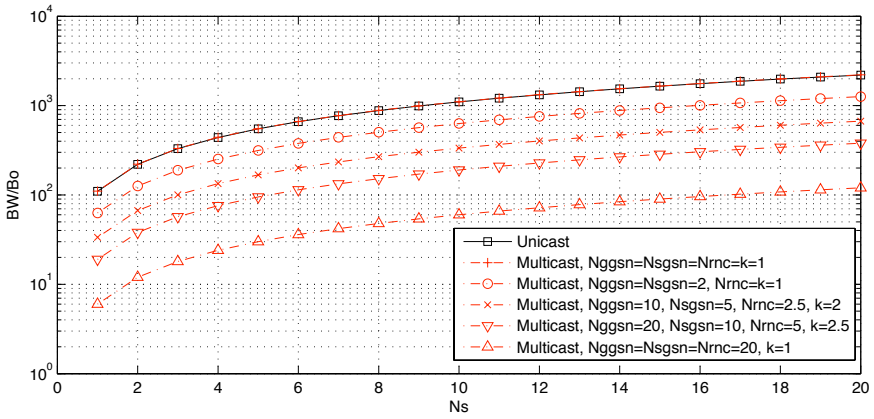


Figure 7: Bandwidth consumption comparative

Comparing this value with the values obtained in the different examples for the multicast case, it can be seen that for the applications envisioned in the proposal presented in this article, even by using media mixing in the user plane, a multicast-based transmission provides better bandwidth utilization results. Additionally, it is important to notice that media mixing is not always possible for every media component type, and even being feasible it might not be implemented in the user plane.

Figure 7 represents the bandwidth consumption (the relation BW/B_0) for the unicast and multicast cases (media mixing has not been considered for unicast), as the number of sources N_s increases. As it can be observed from the figure, the bandwidth utilization is practically identical in the unicast approach and in the worst case situation of the multicast approach, where each UE is served by one GGSN ($N_{GGSN} = N_{SGSB} = N_{RNC} = k = 1$). As the degree of sharing of the UMTS infrastructure increases, the bandwidth utilization significantly decreases for the multicast case. The best case for the multicast approach occurs when all the UEs are located in the same cell and receive the multicast media by means of a shared radio bearer ($N_{GGSN} = N_{SGSN} = N_{RNC} = 20, k = 1$).

4.2. Evaluation of session and user plane setup

In this section, the delays related to the session and user plane setup will be evaluated. The evaluation has been structured in two parts. First, a theoretical analysis is introduced pointing out the relevant delay parameters

that define the Grade of Service (GOS) associated with the session setup procedures presented in Sec. 3. These parameters are analytically estimated by a set of mathematical equations.

Next, an experimental evaluation is described, consisting of a software implementation of a simple multi-user service that utilizes the session setup procedures proposed in this article. This implementation has been used to obtain realistic values for the GOS parameters associated with the proposal. Finally, these values are compared against a set of recommended GOS values by the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T).

4.2.1. Theoretical analysis of session and user plane setup delays

Recommendation E.721 [23] of ITU-T defines a set of GOS parameters for call setup delays. Although this specification is focused on circuit-switched services in ISDN, the definitions included in the recommendation can also be applied to SIP-based call control procedures, such as the one specified in this paper for multiparty services. In this respect, two parameters defined in recommendation E.721 are specially relevant to evaluate the GOS achieved by the presented proposal: the *post-selection delay* and the *answer signal delay*. The definition of these parameters, adapted to the SIP session setup scenario, is indicated next:

Post-selection delay is defined as the time interval from the instant the UE requests a session establishment, until it is informed about the state of the request. In our case, this is the time elapsed between the first bit of the initial INVITE request is put in line by the initiator UE until the last bit of the first message indicating the session setup disposition is received by the initiator UE (RINGING response in case of successful resource reservation).

Answer signal delay is defined as the time interval from the instant that a destination user accepts the incoming session until the initiator user is notified about it. In our case, this is the time elapsed between the first bit of the OK response is put in line by the destination UE until the last bit of the corresponding NOTIFY request is received by the initiator UE.

Let T_1 be the maximum average time elapsed from the instant the initiator UE starts transmitting the INVITE request until the last destination UE receives the corresponding INVITE request). Then:

$$T_1 \approxeq T_{tr}^{UEinit}(INVITE) + T_{proc}^{MAS}(INVITE) + T_{tr}^{MAS \rightarrow UEdest}(INVITE) \quad (15)$$

Where:

- $T_{tr}^{UE_init}(INVITE)$ is the maximum average message-trip time of the INVITE request from the initiator UE to the MAS[1]
- $T_{proc}^{MAS}(INVITE)$ is the average time elapsed from the instant the MAS receives the INVITE request until it starts the transmission of the last INVITE request for the destination UEs.
- $T_{tr}^{MAS \rightarrow UEdest}(INVITE)$ is the maximum average message-trip time of the INVITE request from the MAS to the last destination UE[1].

Let T_2 be the maximum average time from the instant the destination UE, corresponding to the previous equation, receives the INVITE request until the Session in Progress response arrives to the initiator UE:

$$\begin{aligned} T_2 \approxeq & T_{proc}^{UEdest}(SessProg) + T_{tr}^{UEdest \rightarrow MAS}(SessProg) \\ & + T_{proc}^{MAS}(SessProg) + T_{tr}^{MAS \rightarrow UEinit}(SessProg) \end{aligned} \quad (16)$$

Where:

- $T_{proc}^{UEdest}(SessProg)$ is the the average time needed at the destination UE to process the INVITE request and to generate and start transmitting the Session in Progress response.
- $T_{tr}^{UEdest \rightarrow MAS}(SessProg)$ is the maximum average message-trip time of the Session in Progress response from the destination UE to the MAS.
- $T_{proc}^{MAS}(SessProg)$ is the average time that is needed in the MAS to process the session in Progress responses and to generate and start sending the combined Session in Progress that will be sent to the initiator UE.

[1]In order to obtain the maximum average time, it will be assumed from here on that the UMTS accesses that are available to both initiator and destination UEs, considered in this equation, have the lowest performance in terms of capacity among all the accesses that are available to the participants UEs.

- $T_{tr}^{MAS \rightarrow UEinit}(SessProg)$ is the maximum average message-trip time of the Session in Progress response from the MAS to the initiator UE.

Let T_3 be the maximum average time that elapses from the instant the initiator UE receives the Session in Progress response, until the corresponding PRACK request is received at the last destination UE:

$$\begin{aligned} T_3 \cong\ & T_{proc}^{UEinit}(PRACK) + T_{tr}^{UEinit \rightarrow MAS}(PRACK) \\ & + T_{proc}^{MAS}(PRACK) + T_{tr}^{MAS \rightarrow UEdest}(PRACK) \end{aligned} \tag{17}$$

Where:

- $T_{proc}^{UEinit}(PRACK)$ is the the average time needed at the initiator UE to process the Session in Progress response and to generate and start transmitting the PRACK request.
- $T_{tr}^{UEinit \rightarrow MAS}(PRACK)$ is the maximum average message-trip time of the PRACK request from the initiator UE to the MAS.
- $T_{proc}^{MAS}(PRACK)$ is the average time that elapses from the instant the MAS receives the PRACK request until it starts sending the last PRACK request to the destination UEs.
- $T_{tr}^{MAS \rightarrow UEdest}(PRACK)$ is the maximum average message-trip time of the PRACK request from the MAS to the destination UE.

Let T_4 be the maximum average time from the instant the destination UE, corresponding to the previous equation, receives the PRACK request until the OK response arrives to the initiator UE:

$$\begin{aligned} T_4 \cong\ & T_{proc}^{UEdest}(OK_{PRACK}) + T_{tr}^{UEdest \rightarrow MAS}(OK_{PRACK}) \\ & + T_{proc}^{MAS}(OK_{PRACK}) + T_{tr}^{MAS \rightarrow UEinit}(OK_{PRACK}) \end{aligned} \tag{18}$$

Where:

- $T_{proc}^{UEdest}(OK_{PRACK})$ is the the average time needed at the destination UE to process the PRACK request and to generate and start transmitting the OK response.

- $T_{tr}^{UEdest \rightarrow MAS}(OK_{PRACK})$ is the maximum average message-trip time of the OK response from the destination UE to the MAS.

- $T_{proc}^{MAS}(OK_{PRACK})$ is the average time that elapses from the instant the MAS receives the last OK response until it starts transmitting the combined OK response to the initiator UE. In general, the MAS does not need to wait for every OK response, but in the worst case it will need to receive the last OK response in order to progress on the session setup.

- $T_{tr}^{MAS \rightarrow UEinit}(OK_{PRACK})$ is the maximum average message-trip time of the OK response from the MAS to the initiator UE.

At this point, it can be stated that the initiator UE must send the UPDATE request after certain time T_{UPDATE}, measured from the initial instant, that can be expressed as:

$$T_{UPDATE} \cong \max(T_1 + T_2 + T_{proc}^{UEinit}(SessProg) + T_{PDP+IGMP}, \\ T_1 + T_2 + T_3 + T_4) \tag{19}$$

Where:

- $T_{proc}^{UEinitit}(SessProg)$ is the the average time elapsed from the instant the initiator UE receives the combined Session in Progress response until it finishes transmitting the PRACK request.

- $T_{PDP+IGMP}$ is the average time that is needed to establish the necessary PDP contexts and to send the IGMP reports corresponding to the multicast groups (this is a non blocking operation).

According to the proposal presented in Sect. 3, the RINGING response is sent to the initiator UE when the first RINGING response is received from the destination UEs. Calculating the instant when the first RINGING response is received at the MAS in the worst case is not a simple task. Instead of that, this instant will be overestimated by the arrival time of the last RINGING response at the MAS.

Let T_5 be the maximum average time from the instant T_{UPDATE} until the corresponding UPDATE request is received at the last destination UE:

$$T_5 \approxeq T_{proc}^{UEinit}(UPDATE) + T_{tr}^{UEinit \rightarrow MAS}(UPDATE) + T_{proc}^{MAS}(UPDATE) + T_{tr}^{MAS \rightarrow UEdest}(UPDATE) \quad (20)$$

Where:

- $T_{proc}^{UEinitit}(UPDATE)$ is the the average time from T_{UPDATE} until the the initiator UE starts transmitting the UPDATE request.
- $T_{tr}^{UEinit \rightarrow MAS}(UPDATE)$ is the maximum average message-trip time of the UPDATE request from the initiator UE to the MAS.
- $T_{proc}^{MAS}(UPDATE)$ is the average time that elapses from the instant the MAS receives UPDATE request from the initiator UE until it starts transmitting the corresponding UPDATE request to the last destination UE.
- $T_{tr}^{MAS \rightarrow UEdest}(UPDATE)$ is the maximum average message-trip time of the UPDATE request from the MAS to the last destination UE.

The destination UE, corresponding to the previous equation, will send a RINGING response after certain time $T_{RINGING}$, measured from the initial instant:

$$T_{RINGING} \approxeq \max(T_1 + T_2 + T_3 + T_{proc}^{UEdest}(PRACK) + T_{PDP}, T_{UPDATE} + T_5 + T_{proc}^{UEdest}(UPDATE)) \quad (21)$$

Where:

- $T_{proc}^{UEdest}(PRACK)$ is the average time from the instant the destination UE receives the PRACK request until it finishes transmitting the corresponding OK response.
- T_{PDP} is the average time that is needed to establish the necessary PDP contexts.
- $T_{proc}^{UEdest}(UPDATE)$ is the average time from the instant the destination UE receives the UPDATE request until it finishes transmitting the corresponding OK response.

Finally, the average *post-selection delay* (**PSD**) is overestimated as:

$$\begin{aligned} PSD \cong\, & T_{RINGING} + T_{proc}^{UEdest}(RINGING) \\ & + T_{tr}^{UEdest \rightarrow MAS}(RINGING) + T_{proc}^{MAS}(RINGING) \\ & + T_{tr}^{MAS \rightarrow UEinit}(RINGING) \end{aligned} \tag{22}$$

Where:

- $T_{proc}^{UEdest}(RINGING)$ is the time from $T_{RINGING}$ until the destination UE starts transmitting the corresponding RINGING response.
- $T_{tr}^{UEdest \rightarrow MAS}(RINGING)$ is the maximum average message-trip time of the RINGING response from the destination UE to the MAS.
- $T_{proc}^{MAS}(RINGING)$ is the average time elapsed from the instant the MAS receives the RINGING response until it starts transmitting the RINGING response to the initiator UE.
- $T_{tr}^{MAS \rightarrow UEinit}(RINGING)$ is the maximum average message-trip time of the RINGING response from the MAS to the initiator UE.

With respect to the *answer signal delay* (**ASD**), the worst case corresponds to the scenario where all the UEs have a UMTS access with the same performance in terms of bandwidth, and all the destination users accept the incoming session simultaneousy. Although improbable, this scenario provides the worst average value for the answer signal delay, which can be expressed by the following equation:

$$\begin{aligned} ASD \cong\, & T_{tr}^{UEdest \rightarrow MAS}(OK_{INVITE}) + T_{proc}^{MAS}(OK_{INVITE}) \\ & + T_{tr}^{MAS \rightarrow UEinit}(NOTIFY) \end{aligned} \tag{23}$$

Where:

- $T_{tr}^{UEdest \rightarrow MAS}(OK_{INVITE})$ is the average message trip-time of the last OK response to an INVITE request received from a destination UE

- $T_{proc}^{MAS}(OK_{INVITE})$ is the average time elapsed from the instant the last OK response is received at the MAS, until the MAS starts sending the NOTIFY request corresponding to the OK response. Notice that this time interval includes the necessary time to process all the previous OK responses, consisting of generating and putting in line an OK response for the initiator UE, and generating and putting in line a NOTIFY request for each participant UE for every previous OK response.

- $T_{tr}^{MAS \rightarrow UEinit}(NOTIFY)$ is the average message trip-time of the NOTIFY request from the MAS to the initiator UE.

Regarding to the average message-trip time from any UE to the MAS, this time is mainly determined by the delay imposed to the message in the UMTS access network of the UE, and the delays suffered at the set of traversed CSCFs. Assuming that the message-trip time from the GGSN to the P-CSCF and between CSCFs is negligible compared with those delays, the average trip time of a SIP message from the UE (initiator or destination) to the MAS can be expressed as:

$$T_{tr}^{UE \rightarrow MAS}(message) \cong T_{tr}^{access}(message) + \sum_{\forall CSCF} T_{proc}^{CSCF}(message) \quad (24)$$

Where:

- $T_{tr}^{access}(message)$ is the average delay experienced by the SIP message in the UMTS access, i.e. from the UE to the GGSN.

- $T_{proc}^{CSCF}(message)$ is the average processing time of the SIP message at a given CSCF (i.e. P-CSCF, S-CSCF or I-CSCF).

The term $\sum_{\forall CSCF} T_{proc}^{CSCF}(message)$ refers to the summation of the processing delays at the CSCFs traversed by the SIP message. For instance, for the case of the INVITE request transmitted from the initiator UE to the MAS, this summation includes the processing delays of one P-CSCF and one S-CSCF.

Analogously, the average message-trip time from the MAS to any UE (initiator or destination) can also be expressed by Eq. 24, although in this case the term $T_{tr}^{access}(message)$ refers to the average delay experienced by the SIP message in the UMTS access from the GGSN to the UE.

4.2.2. Evaluation of SIP signaling message delays

In this subsection, an experiment was designed in order to obtain realistic measures of the delays suffered by the different SIP signaling messages on a real UMTS access. For this purpose, a software implementation of the MAS and the UE was developed, capable of executing the session establishment procedures specified in Sect. 3. Further details on this implementation are provided in the next subsection.

By using this software, the sizes of the different SIP signaling messages, exchanged during a session setup, were obtained. For each message size, it is possible to measure its Round Trip Time (RTT) from a UE to its corresponding GGSN by using real UMTS access. Each RTT value represents an estimation of the delay of the SIP message corresponding to that size within the UMTS access, taking into account the transmission from the UE to the GGSN and in the reverse direction.

An acquisition process of RTT values was scheduled. In each execution of the process, five values of RTT were obtained for each message size. As the result, the execution produces five traces of time delays, each trace containing one RTT value for each of the SIP message sizes.The executions were daily planned with a period of fifteen minutes (from 00:00 to 23:45), and the acquisition process was maintained for over one month. Table 1 summarizes the average delay experienced by each message during the period of high load.

Message	**Average delay (ms)**
INVITE	134.37
Session in Progress	133.33
PRACK	133.52
OK_{PRACK}	121.26
UPDATE	136.62
OK_{UPDATE}	121.26
RINGING	78.04
OK_{INVITE}	78.04
ACK	80.87
NOTIFY	207.32
OK_{NOTIFY}	78.13

Table 1: Average delays in the UMTS access

Ignoring by the moment the processing delays at UEs and at the MAS, the average delays in the UMTS access network could be replaced in equations 22 and 23. Considering an average processing time of SIP signaling messages at each CSCF of 25 ms (see [24]), and asssuming that the resource reservation procedure finalizes in the initiator UE before receiving the OK response to the PRACK request, and in the destination UE before sending the OK response to the UPDATE request, then the *post-selection delay* can be calculated as:

$$PSD \approxeq 2299.28 \text{ ms} \tag{25}$$

Table 2 contains the target values of the *post-selection delay* and the *answer signal delay* for a local service, according to recommendation E.721 [23]. The local service corresponds to the most restrictive target values included in the recommendation, and is an appropriate category to represent the typical scenario considered in this proposal, where UEs are geographically close and may share the access network infrastructure, such as RNCs, SGSNs and/or GGSNs in the case of UMTS.

GOS parameter	**Normal load**		**High load**	
	Mean	**95%**	**Mean**	**95%**
Post-selection delay	3000 ms	6000 ms	4500 ms	9000 ms
Answer signal delay	750 ms	1500 ms	1000 ms	2000 ms

Table 2: target values for GOS parameters

Comparing the average value obtained for the PSD with the target value for high load (see table 2), it can be seen that the obtained value is significantly lower than the target (i.e. 4500 ms). Therefore, providing that the processing delays at UEs and MAS are below 2200.72 ms, the target value is satisfied. This restriction could be easily achieved in a real deployment, by running the software of UEs and MAS in hardware platforms with enough computing and memory resources, and by planning an appropriate distribution of UEs per MAS.

Regarding to the ASD, by evaluating Eq. 23, the following value is obtained:

$$ASD \approxeq 435.36 \text{ ms} \tag{26}$$

Again, the obtained value is lower than the target for high load (i.e. 1000 ms), and the same conclusions described for the PSD hold.

4.2.3. Experimental evaluation of session and user plane setup delays

The values obtained from Eqs. 25 and 26, correspond to theoretical estimations where the processing delays at UEs and the MAS have not been considered. In order to precisely evaluate the average values of *post-selection delay* and *answer signal delay*, as they evolve with respect to time and to the number of participants, taking also into account all the processing delays, a new experiment was designed.

For this experiment, a software implementation of the MAS was developed according to the specifications described in Sect. 3. In addition, a software implementation of an UE was addressed, capable of initiating and handle a session setup for a simple multiparty service, consisting of the exchange of a single media component. Both software prototypes were developed in Java (version 1.5.0), utilizing the JAIN-SIP API[4].

On the other hand, a set of ninety six files was generated, each one corresponding to a time between 00:00 and 23:45 with a 15 minutes period. These files were processed to contain all the delay traces acquired at that time during the acquisition process of RTT values. This way, each file was generated to contain a large number of traces, where each trace included a real RTT value for each SIP message exchanged in a session setup within an UMTS access.

For the experiment, a virtual interface system was set up on a quad-core computer with 12 GB RAM based on the ModelNet platform[5]. Under this platform, a number of UEs and one MAS can be executed, being assigned each of them to a virtual network interface. This way, the UEs and the MAS run as if they were installed on a single independent machine, being accessible to the rest of entities that participate in the service by means of their own interface. Assuming a number N of participant UEs and a given time in the interval from 0:00 to 23:45, then the experiment consisted of executing N instances of the UE and the MAS. In this scenario, one of the UEs assumes the role of the initiator and establishes 30 multi-user sessions, the next session setup starting after the previous completes. As a result,

[4]JAIN SIP Developer Tools, https://jain-sip.dev.java.net/

[5]ModelNet, https://modelnet.sysnet.ucsd.edu/

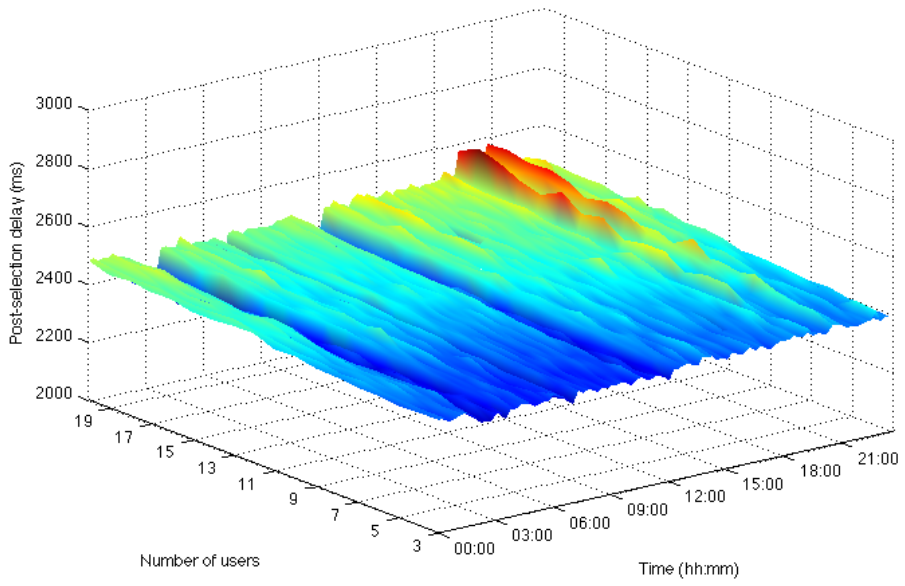


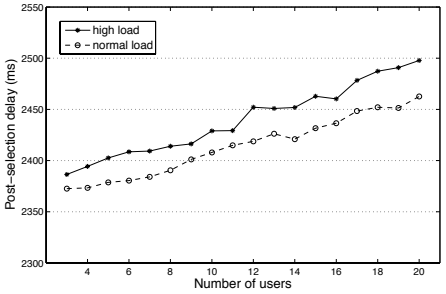
Figure 8: Evolution of the average value of the *post-selection delay*

the *post-selection delay* and the *answer signal delay* (in this case, for each destination UE) are experimentally measured.

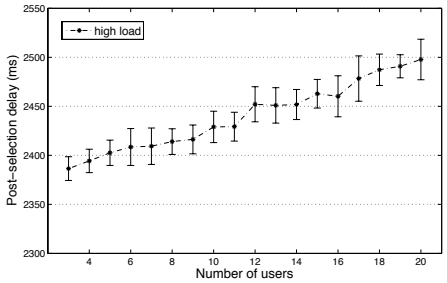
The experiment also takes into account the delays experienced by the different SIP messages in the UMTS accesses and the CSCFs. For each session setup, each UE is provisioned with a trace of RTTs from the chosen time, which includes a delay value to be applied to each SIP message. Therefore, each UE will apply to each SIP message its corresponding delay within the UMTS access. On the other hand, the UE also appends to this delay the different processing times that are suffered by the message when traversing the set of CSCFs (these processing times were obtained from [24]).

Finally, the proposed experiment was repeated for every value of N ranging from 3 to 20, and for all the times from 0:00 to 23:45 (with an interval of 15 minutes).

Figure 8 shows the results obtained for the case of the *post-selection delay* with respect to time and the number of users. As it can be observed from the figure, the values obtained for the PSD are always kept under the target value for normal load (i.e. 3000 ms). The graph does not show a significant variation of the average PSD with respect to time and as the number of



(a) HIgh and normal load



(b) Confidence intervals for high load

Figure 9: Evolution of the *answer signal delay* with the number of users

users increases. The maximum values are obtained for larger number of users during the period of high load (i.e. afternoon and evening)

Figure 9(a) shows a detailed view of the variation of the average PSD with respect to the number of users. The graph includes the results for high and normal load. As it can be observed, there is a similar variation of PSD for maximum and normal load, being the average values corresponding to the maximum load greater than those obtained for normal load. This is due to the fact that the PSD corresponds to the time elapsed since the initiator sends the INVITE request until it receives the RINGING response. This time cumulates the delay corresponding to several SIP messages, and these delays are greater for the case of maximum load. On the other hand, the average PSD for high and normal load are significantly lower that the target values defined in table 2 (4500 ms and 3000 ms respectively). Figure 9(b) shows the confidence intervals for the average PSD with respect to the number of users, in the high load period.

Figure 10 shows the results obtained for the *answer signal delay* with respect to time and the number of users. Similarly to the case of the PSD, the average values of the ASD are always kept below the target limit for normal load (i.e. 750 ms).

In order to get a more detailed view of the ASD evolution, Fig. 11(a) shows the values obtained for the ASD with respect to the number of users. The graph includes the results for high and normal load. As it can be seen from the graph, the average values for normal load are very close to the values obtained for maximum load. This is because, the ASD is the time elapsed from the instant the destination UE sends the OK response to the INVITE

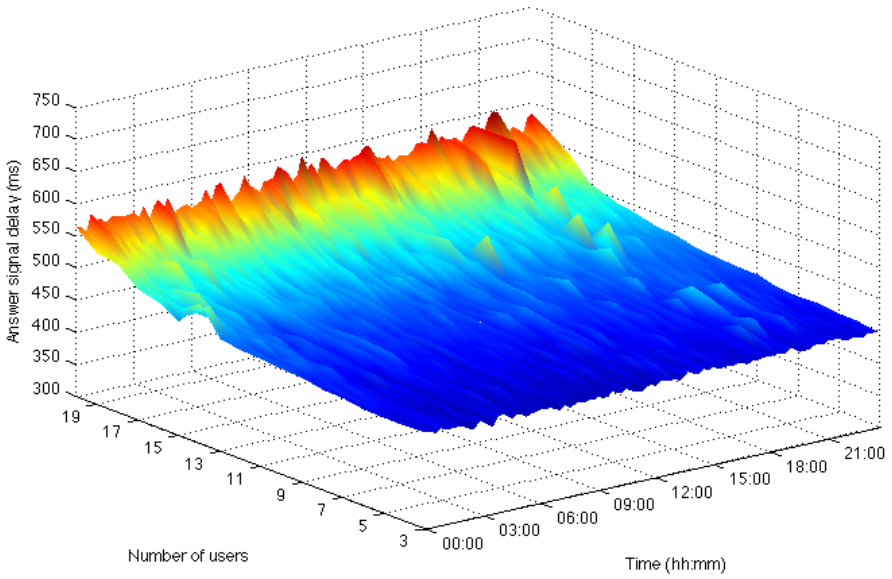
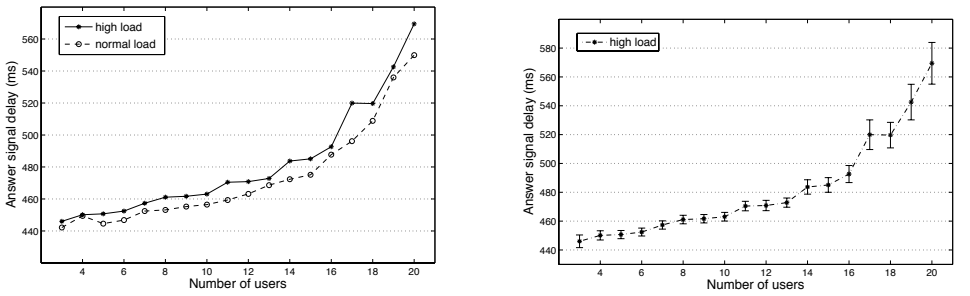


Figure 10: Evolution of the average *answer signal delay*

request until the corresponding NOTIFY request arrives to the initiator UE. This time only includes the delays corresponding to the OK and NOTIFY request. This way, the differences are not too accentuated between high and normal load. On the other hand, whenever one user accepts the incoming session, a SIP OK response is received at the MAS. The processing of this response implies generating and sending one NOTIFY request for each participant user. Consequently, the load at the MAS increases as the number of users increases. This load has an impact on the results obtained for the ASD, that increases with the number of users. However, for the number of users that has been considered, the average ASD for high and normal load is significantly lower than the corresponding target values recommended by the ITU-T (i.e. 1000 ms and 750 ms respectively). Figure 11(b) shows the confidence intervals for the average ASD with respect to the number of users, in the high load period.



(a) High and normal load (b) Confidence intervals for high load

Figure 11: Evolution of the *answer signal delay* with the number of users

5. Conclusions

This paper describes extensions to the IMS session setup procedures, so as to support multicast-based many-to-many services. The basic procedures were first described in a prior work, but in this article several enhancements are presented, related with notifying the session status to the participant users, analyzing the scope and the applicability of the solution and considering access network technologies where UEs do not need to perform resource reservation. These enhancements allow to provide a comprehensive solution and improve the Grade of Service (GOS) perceived by the end users.

On the other hand, the proposed mechanisms were theoretically and experimentally evaluated. A theoretical analysis demonstrate that network multicast outperforms the unicast-based transmission in terms of bandwidth consumption, as the degree of sharing of the network infrastructure increases among the UEs that participate in a multi-user service. On the other hand, the GOS achieved by the proposal has been theoretically evaluated, and a test-bed has been deployed to address the evaluation from a experimental perspective. In both cases, the results have been compared with a set of target values recommended by the ITU-T, verifying that the obtained values are aligned with this recommendation.

Future work will include extending this proposal to cover the scenario where participant users are located in different network domains, and developing new session setup mechanisms to allow users to gradually join an ongoing multicast-based multi-user session.

References

[1] C. Diot, B. N. Levine, B. Lyles, H. Kassem, D. Balensiefen, Deployment Issues for the IP Multicast Service and Architecture, IEEE Network 14 (1) (2000) 78–88.

[2] S. Ratnasamy, A. Ermolinskiy, S. Shenker, Revisiting IP Multicast, in: ACM SIGCOMM'06, Pisa, Italy, September 2006.

[3] C. Riede, A. Al-Hezmi, T. Magedanz, Session and Media Signaling for IPTV via IMS, in: Mobilware'08, Innsbruck, Austria, February 2008.

[4] Open IPTV Forum Functional Architecture V1.1, `http://www.openiptvforum.org` (2008).

[5] ETSI, IPTV Architecture, TS 182.027, Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN) (Jul. 2009).

[6] I. Vidal, I. Soto, F. Valera, J. Garcia, A. Azcorra, IMS signalling for multiparty services based on network level multicast, in: 3rd EURO-NGI Conference on Next Generation Internet Networks, Trondheim, Norway, May 2007.

[7] I. Vidal, I. Soto, F. Valera, J. Garcia, A. Azcorra, Multiparty Services in the IP Multimedia Subsystem, in: M. Ilyas, S. A. Ahson (Eds.), IP Multimedia Subsystem (IMS) Handbook, CRC Books, 2009.

[8] 3GPP, IP Multimedia Subsystem (IMS); Stage 2, TS 23.228, 3rd Generation Partnership Project (3GPP) (Sep. 2008).
URL `http://www.3gpp.org/ftp/Specs/html-info/23228.htm`

[9] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, SIP: Session Initiation Protocol, RFC 3261 (Proposed Standard), updated by RFCs 3265, 3853, 4320, 4916, 5393, 5621, 5626, 5630 (Jun. 2002).
URL `http://www.ietf.org/rfc/rfc3261.txt`

[10] M. Handley, V. Jacobson, C. Perkins, SDP: Session Description Protocol, RFC 4566 (Proposed Standard) (Jul. 2006).
URL `http://www.ietf.org/rfc/rfc4566.txt`

[11] J. Rosenberg, H. Schulzrinne, An Offer/Answer Model with Session Description Protocol (SDP), RFC 3264 (Proposed Standard) (Jun. 2002).
URL http://www.ietf.org/rfc/rfc3264.txt

[12] 3GPP, Conferencing using the IP Multimedia (IM) Core Network (CN) subsystem; Stage 3, TS 24.147, 3rd Generation Partnership Project (3GPP) (Mar. 2008).
URL http://www.3gpp.org/ftp/Specs/html-info/24147.htm

[13] 3GPP, 3GPP enablers for Open Mobile Alliance (OMA) Push-to-talk over Cellular (PoC) services; Stage 2, TR 23.979, 3rd Generation Partnership Project (3GPP) (Jun. 2007).
URL http://www.3gpp.org/ftp/Specs/html-info/23979.htm

[14] 3GPP, Internet Protocol (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) (Sep. 2008).
URL http://www.3gpp.org/ftp/Specs/html-info/24229.htm

[15] J. Rosenberg, H. Schulzrinne, Session Initiation Protocol (SIP): Locating SIP Servers, RFC 3263 (Proposed Standard) (Jun. 2002).
URL http://www.ietf.org/rfc/rfc3263.txt

[16] 3GPP, Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description, TS 23.246, 3rd Generation Partnership Project (3GPP) (Jun. 2008).
URL http://www.3gpp.org/ftp/Specs/html-info/23246.htm

[17] G. Camarillo, W. Marshall, J. Rosenberg, Integration of Resource Management and Session Initiation Protocol (SIP), RFC 3312 (Proposed Standard), updated by RFCs 4032, 5027 (Oct. 2002).
URL http://www.ietf.org/rfc/rfc3312.txt

[18] G. Camarillo, P. Kyzivat, Update to the Session Initiation Protocol (SIP) Preconditions Framework, RFC 4032 (Proposed Standard) (Mar. 2005).
URL http://www.ietf.org/rfc/rfc4032.txt

[19] B. Cain, S. Deering, I. Kouvelas, B. Fenner, A. Thyagarajan, Internet Group Management Protocol, Version 3, RFC 3376 (Proposed Standard), updated by RFC 4604 (Oct. 2002).
URL http://www.ietf.org/rfc/rfc3376.txt

[20] A. B. Roach, Session Initiation Protocol (SIP)-Specific Event Notification, RFC 3265 (Proposed Standard), updated by RFC 5367 (Jun. 2002).
URL http://www.ietf.org/rfc/rfc3265.txt

[21] J. Rosenberg, H. Schulzrinne, O. Levin, A Session Initiation Protocol (SIP) Event Package for Conference State, RFC 4575 (Proposed Standard) (Aug. 2006).
URL http://www.ietf.org/rfc/rfc4575.txt

[22] Z. Albanna, K. Almeroth, D. Meyer, M. Schipper, IANA Guidelines for IPv4 Multicast Address Assignments, RFC 3171 (Best Current Practice) (Aug. 2001).
URL http://www.ietf.org/rfc/rfc3171.txt

[23] ITU-T, Network grade of service parameters and target values for circuit-switched services in the evolving ISDN, Recommendation E.721, Telecommunication Standardization Sector of International Telecommunication Union (ITU-T) (May 1999).

[24] D. Pesch, M. Pous, G. Foster, Performance evaluation of SIP-based multimedia services in UMTS, Computer Networks 49 (3) (2005) 385–403.