Performance-driven Evaluation for Deploying IMS-based Interoperability Scenarios

Jose Oscar Fajardo, Fidel Liberal
Dept. Communications Engineering
University of the Basque Country (UPV/EHU)
Bilbao, Spain

Fudong Li, Nathan Clarke, Is-Haka Mkwawa, Lingfen Sun
School of Computing & Mathematics
Plymouth University
Plymouth, U.K.

***Abstract*— This paper deals with the performance evaluation of deploying an IMS-based media plane interoperability framework. The 3GPP standards describe two possible operating modes based on either a reactive or a proactive approach. We show that both approaches entail some advantages and drawbacks in terms of signaling overhead and call setup times. In order to use experimental individual delay contributions, a prototype implementation of the required elements was carried out for which transmission and processing times and the average ratio of incompatible calls where evaluated. Additionally, the evaluation focuses on the impact of deploying the interoperability solutions over current UMTS and LTE radio access networks.**

Keywords—IMS interoperablity, performance evaluation

I. Introduction

IP Multimedia System (IMS) has become a prevailing architectural framework for delivering information-rich multimedia services over next generation IP networks (e.g. Voice over LTE (VoLTE)) since the original proposal of the 3rd Generation Part Project (3GPP) in 2002 [1]. With nearly one billion expected subscribers in 2013, many Internet users have already started to utilize the cost-effective and convenient IP multimedia service to form their normal daily communications [2]. In addition, Internet users could also capitalize on the IP multimedia service to summon emergency services when they are in critical situations (e.g. car accident) [3]. It is envisaged that some multimedia services (e.g. video conferencing, picture transmission and location) contain additional critical information that can be utilized by emergency services to build a better view of incidents and therefore more lives could be potentially saved.

In order to successfully establish an IP multimedia service session, User Equipment (UE) is required to be compatible with each other at both signaling and media planes of the IMS framework according to [1] and [4]. It is a common practice that UEs are dependent upon the Session Initiation Protocol (SIP) and Session Description Protocol (SDP) for call managements (e.g. call setups) in the signaling plane and the Real-time Transport Protocol (RTP) and/or Secure RTP (SRTP) for media transmissions in the media plane [5]. However, UEs can utilize various voice/video codec schemes to encode and decode the media content and different security mechanisms (i.e. the combination of the key exchange method and the crypto suite) to secure the media transmission. Incompatibility issues will occur whenever UEs utilize different codec schemes and/or security mechanisms resulting into early call terminations in the signaling plane or into the transmission of unencrypted media contents in the media plane.

Interoperability for any incompatible UEs may be provided by the inclusion of two components in the standard IMS framework (as shown in Fig. 1): a control element at the end users' signaling path, i.e. a Third Party Call Control (3PCC) and an interoperability enabler at the end users' media path, i.e. the Media Resource Function (MRF). The 3PCC is originally proposed for handling advanced features in a multimedia communication, e.g. voice continuity and multiparty calls. The MRF is a media plane element initially designed to provide UEs with various media support, e.g. controlling media stream resources and processing media streams. In order to provide interoperability for media mismatched UEs, the 3PCC should be able to detect if a media mismatch occurred at the signaling plane and the MRF should be able to provide appropriate media plane support (i.e. transcoding and cross-ciphering).

Two distinct signaling approaches (i.e. proactive and reactive) that are designed to offer interoperability in the signaling plane can be utilized by the 3PCC to invoke the MRF for the purpose of transcoding and cross-ciphering during a call setup session between UEs [1]. For the proactive approach, the 3PCC is configured to assume that all UEs are not compatible with each other; hence, it invokes the MRF for additional media supports before the call offer reaches the callee UE. Meanwhile, for the reactive approach the 3PCC is programmed to consider that all UEs are compatible with each other by default. As a result, the invocation on the MRF only occurs after the call offer is received and rejected by the callee UE. In both signaling cases, media capability of the MRF must be acquired in the form of SIP messages to ensure the communication is established between incompatible UEs.

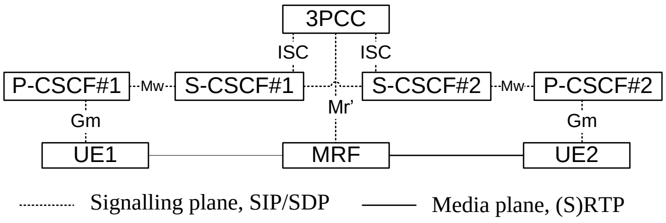

Fig. 1. A simplified IMS framework with 3PCC and MRF enabled

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 284863 (FP7 SEC GERYON).

The amount of additional information required is dependent upon individual approaches and their impact upon the IMS signaling plane is arguably different. Therefore, this paper presents a performance-driven study with the aim of evaluating the impact of the proactive and reactive signaling approaches upon the IMS signaling plane in terms of system overload and call setup time. As general preconditions for the study, it is assumed that (i) the MRF can always provide interoperability support for UEs at the media plane; (ii) there is no a priori knowledge about the ratio of calls with incompatible endpoints.

This paper begins by introducing the need of interoperability support for establishing real-time multimedia communication between incompatible UEs and analyzing two interoperability approaches that can be utilized in the IMS signaling plane. In section II, both the proactive and reactive signaling approaches are explained in detail. Section III describes the performance analysis for both approaches in terms of number SIP messages and total processing time. Section IV details the comparative performance results in terms of signaling overhead and call setup times, by using individual experimental delay values from a prototype implementation. Performance results concerning LTE and UMTS access networks are especially stressed. The paper finishes by highlighting the conclusions and future research directions.

II. Description of the Interoperability Modes

As mentioned in the introduction, two approaches can be adopted by the 3PCC to provide interoperability for incompatible UEs during the call setup session in the IMS signaling plane: proactive and reactive. Details of how the 3PCC operates under these two possible operating modes are fully described in the following sections.

A. *Proactive Approach*

When 3PCC operates in the Proactive Approach, it includes the media capabilities of the MRF into the session request of the calling UE and forwards the modified request to the called UE regardless of the UEs media capabilities. A high-level SIP message flow of the “Proactive Call” between the UEs, the 3PCC and the MRF is illustrated in Fig. 2.

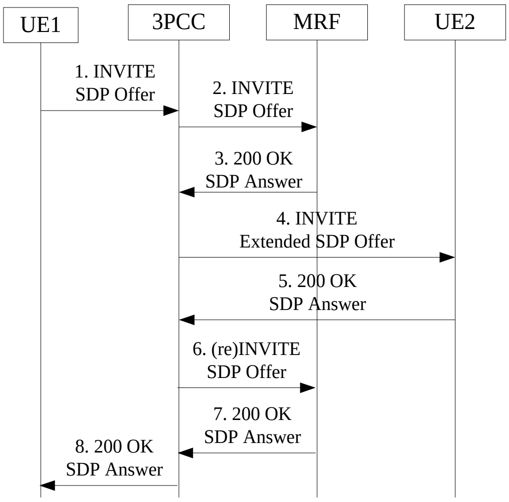

Fig. 2. Proactive Approach

- 1. UE1 sends an INVITE SDP offer including its media capabilities to UE2
- 2. 3PCC forwards the INVITE SDP offer to the MRF as it is programmed to invoke the MRF
- 3. The MRF replies with its media capabilities
- 4. The 3PCC sends the modified SDP offer (including the media capabilities of UE1 and MRF) to UE2
- 5. UE2 replies with its preferred media capabilities
- 6. Depending upon the reply of UE2, the 3PCC decides whether the media support of the MRF is required. The 3PCC re-invites the MRF if the MRF is needed; otherwise, the MRF is released.
- 7. The MRF sends a 200 OK Answer if it is required; otherwise it sends a BYE message.
- 8. 3PCC replies a 200 OK message for the INVITE SDP offer of the UE1 and communication can be started at the media plane.

B. *Reactive Approach*

When the 3PCC works reactively, it adds the media capability of the MRF to a second request towards the called UE only after the original session request of the caller UE has been sent to and rejected by the callee UE. A high-level SIP message flow of the reactive approach between the UEs, the 3PCC and the MRF is illustrated in Fig. 3.

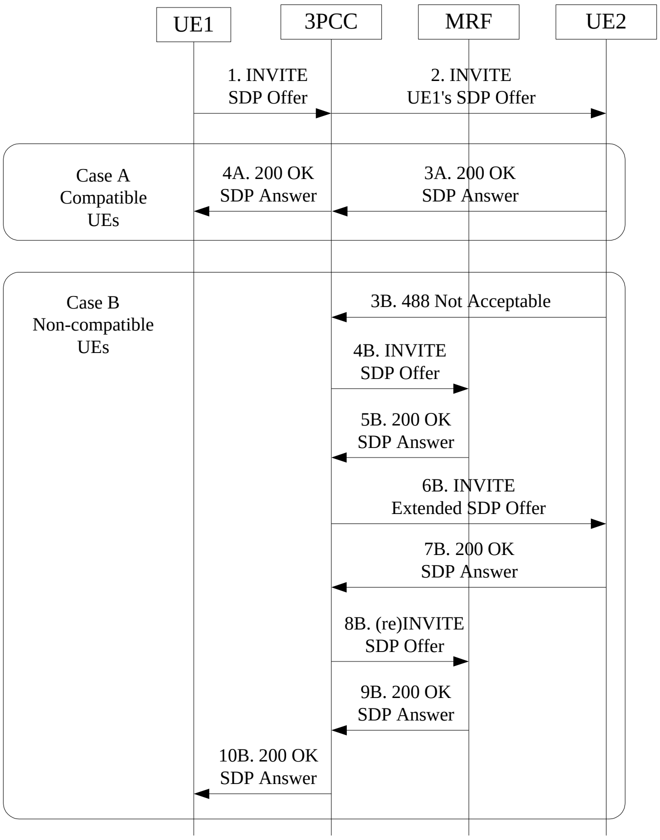

Fig. 3. Reactive Approach

- 1. UE1 sends an INVITE SDP offer including its media capabilities to UE2
- 2. 3PCC forwards the INVITE SDP offer to UE2 as it is programmed not to invoke the MRF unless two UEs are incompatible with each other.

When two UEs are compatible with each other, the call setup is denoted as a "Direct Call" and the following SIP messages occur:

- 3A. UE2 replies with 200 OK message.
- 4A. The 3PCC forwards the 200 OK message to UE1 and two UEs start the media communication without the support of the MRF.

When two UEs are not compatible with each other, the following SIP messages occur to proceed with the "Reactive Call":

- 3B. UE2 replies with 488 Not Acceptable message due to their incompatibility.
- 4B) The 3PCC detects incompatibility issue occurred and it invokes the MRF for additional media supports by sending the INVITE SDP offer of the UE1 to the MRF.
- 5B. The MRF replies with its media capabilities.
- 6B. The 3PCC sends the modified SDP offer (including the media capabilities of UE1 and MRF) to UE2.
- 7B. UE2 replies with its preferred media capabilities.
- 8B. The 3PCC re-invites the MRF with the preferred media capabilities of both UEs.
- 9B. The MRF sends a 200 OK message to the 3PCC.
- 10B. The 3PCC sends a 200 OK to UE1 and the communication can be started at the media plane.

C. Preliminary Discussion of Operating Modes

As illustrated above, in the Proactive Approach, the 3PCC always includes the additional media support of the MRF before the call request reaches the callee UE. In this way, the communication can be guaranteed regardless of the UEs media capabilities. However, additional delays will be added for UEs that are compatible with each other.

For the Reactive Approach, the 3PCC only invokes the MRF for its media capability if and only if the the callee UE confirms that its media not compatible with the caller UE. In this context, additional delays will only occur when two UEs are not compatible with each other. However, the amount of added delays is higher than that of the proactive approach in terms number of SIP messages.

It is clear that both the proactive and reactive approaches can provide interoperability for media incompatible UEs with some compromise of the overall system performance. Therefore, a number of factors that may influence the IMS signaling plane performance under each aforementioned approach are analyzed in the next section.

III. PERFORMANCE ANALYSIS

Two metrics will be taken into consideration to evaluate the performance of the proactive and reactive signaling approaches: the number of SIP messages required as an indicator for the signaling overload of the system and the estimated call setup time required to establish the media sessions. In both cases, the performance metrics are compared between the two alternative operating modes from an overall system standpoint, considering average values for all the call setup procedures.

For the number of SIP messages, it is straight forward for the Proactive Approach. All the call setup procedures experience the same amount of SIP messages regardless of the media capability of the UEs. Since no a priori information about the compatibility of the devices is given, every call setup will be handled with the proactive addition of media attributes. In comparison, the number of SIP messages for the Reactive Approach varies depending upon the compatibility of the UEs. The "Direct Calls" require less number of SIP messages than the "Proactive Calls", while the "Reactive Calls" result on the highest signaling overhead. As a result, the expected performance in the Reactive Approach depends on the ratio of sessions requiring additional call management procedures due to interoperability of UEs, i.e. the ratio of "Reactive Calls".

Although this latency value should be generally related to the number of required SIP messages, the actual values are not directly proportional. First of all, some signaling messages (e.g. 100 trying) are used for progressing feedback to the previous hop, and thus the SIP state machines do not need to wait for the reception of every message. Additionally, the signaling messages may entail different transmission delays in function of the technology at each hop. Different wireless technologies (e.g. LTE or UMTS) will add higher delays in the messages involving the UEs, while signaling messages within the core network would usually require lower processing time. Finally, processing time must be considered at the different nodes with especial focus on the UEs, the 3PCC and the MRF.

Based upon the aforementioned considerations, an analytical expression for computing the call setup time for a single session (T0) is illustrated in (1).

$$T_0 = N_{AN.}T_{AN}+N_{CN.}T_{CN} +N_{3PCC.}T_{3PCC} +N_{MRF.}T_{MRF} +N_{UE.}T_{UE}(1)$$

The different contributions to the total delay considered are:

- The delay due to the number of transmissions through the access networks (N_{AN}) according to each individual access network delay (T_{AN}). In this case, the same time is considered for the two UEs and in both directions (i.e. uplink and downlink).
- The delay due to the number of transmissions within the core network (N_{CN}). A single core network delay (T_{CN}) is considered as an average of the core segments.
- The delay associated to the processing of the signaling messages at the 3PCC (T_{3PCC}) and the number of required 3PCC operations (N_{3PCC}).

- The delay due to the processing of the signaling messages at the MRF (T_{MRF}) and the number of required MRF operations (N_{MRF}).
- The processing times at the UEs (T_{UE}) multiplied by the number of times that the endpoints need to apply some processing logic (N_{UE}).

A detailed analysis of the call setup procedures is required in order to compute the actual overall delay that is contributed by each individual network node. The high-level procedures provided in Fig. 2 and Fig. 3 for the two operating modes can be used as a reference for the analysis. Moreover Fig. 4 illustrates a detailed call setup procedure for the case of a "Reactive Call", including the whole set of messages involved in the case study.

In this case, the UEs belong to the same IMS domain and have different P-CSCF nodes as entry points. Meanwhile, they are served by the same S-CSCF node and thus no additional signaling between different S-CSCF nodes is considered.

From the whole set of signaling messages, and processing operations, Fig. 4 identifies those messages that must be actually taken into account for the evaluation of the call setup times. For example, the "100 Trying" messages are used in a per hop basis to feedback the reception of the message. These messages have to be considered for accounting the overall number of SIP messages, but not for the entire call setup times. Similarly, the "200 OK (INVITE)" is triggered after the UE has processed the SDP content, but there is no need to wait for the arrival of the "200 OK (PRACK)". For clarity purposes, the core network messages relevant to the setup time are not marked. However, the total number of significant core messages are those associated to the SIP messages marked in the access networks, and those related to interactions between the 3PCC and MRF nodes.

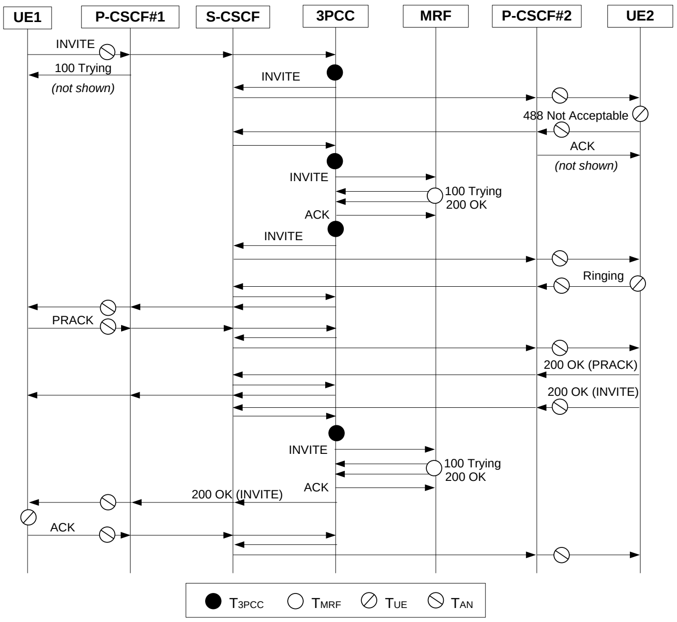

Fig. 4. Detailed sequence diagram of "Reactive Calls" setup procedure

IV. Performance Results

Based upon foundation that is laid by the theoretical analysis, a practical performance study on the impact of proactive and reactive signaling approaches upon the IMS signaling plane is conducted in this section.

A. *Performance in terms of signalling overhead*

As mentioned before, the total number of signaling messages required to complete the session establishment procedures is a key parameter that can be utilized for the performance comparison of the aforementioned signaling approaches. It is envisaged that a higher ratio of signaling overhead implies a larger utilization of network resources, which may lead to possible performance degradations of the network segments and nodes.

From Fig. 4, we can determine the number of SIP messages required for the establishment of a "Reactive Call". The total number of SIP messages is 62, taking into account the whole procedure. Considering only the SIP messages within the core network, so thus not including the messages involving the endpoints, the number is reduced to 44.

Table I provides the values obtained from similar analyses for the different types of session described in Section II. As can be observed, the "Direct Call" requires the smallest number of signaling messages (i.e. 42) to complete a call session setup; the "Reactive Call" needs the largest number of SIP messages (i.e. 62) to achieve the same goal; and the "Proactive Call" experiences an intermediate value of 50 SIP messages..

In order to compare performance in terms of SIP messages under the two signaling approaches, the traffic mix between sessions requiring interoperability and sessions with compatible endpoints needs to be analyzed. The ratio of compatible and non-compatible UEs determines the performance of the Reactive Approach as it contains a mixture of "Direct Calls" and "Reactive Calls"; while the performance of the Proactive Approach should always be static as it entails a fixed number of messages for every call setup.

Fig. 5 illustrates the required signaling messages (i.e. the total and core messages) for both proactive and reactive approaches under different traffic patterns. As demonstrated in Fig. 5, the ratio of non-compatible calls has not impact upon the performance of the proactive signaling approach; while there is a linear relationship between the ratio of non-compatible calls and the performance of the reactive signaling mode: when the ratio increases, the performance decreases and vice versa.

TABLE I. NUMBER OF SIP MESSAGES REQUIRED

Required mode	**Required number of SIP messages**	
	Total number of messages	***Number of core messages***
Direct	42	28
Reactive	62	44
Proactive	50	36

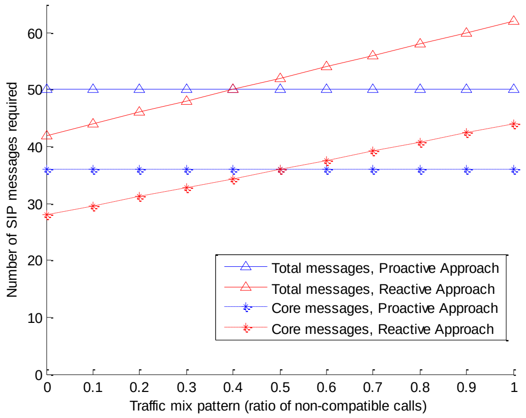

Fig. 5. Required signaling messages for different traffic patterns

Also, the performance comparison of both approaches can be obtained by setting thresholds to the ratio of non-compatible calls of 40% and 50% for the number of total SIP messages and core messages respectively: the reactive approach outperforms the proactive mode when the ratio is under threshold and the proactive mode outclasses the reactive approach when the ratio exceeds the threshold. Therefore, the threshold can be utilized as a guideline to determine which signaling approach should the 3PCC be implemented in a real network situation to fulfill the purpose of providing interoperability and at the same time minimizing the impact of these signaling approaches upon the network performance.

B. Performance in terms of call setup times

In order to compare the performance of the proactive and reactive approaches in terms of call setup time, the overall time of how a call is setup under each approach should be obtained. Following (1) and the sequence diagram provided in Fig. 4, we can estimate the call setup time for the sessions of type “Reactive Call” over a particular network. Also, the processing time of each individual network nodes (e.g. the 3PCC, the MRF and two UEs) should also be taken into consideration.

With the aim of estimating the aforementioned delays, a modest scale prototype IMS system was deployed. The 3PCC was implemented based on the SIP Express Media Server (SEMS) project, including the required logic to handle the different types of calls and the specific interface to the MRF. Also, during a first set of tests a dummy MRF was developed based on the Open Source SIPp test tool. For the testbed scenario, both the 3PCC and the MRF are deployed as Application Servers in an IMS infrastructure based on the FOKUS Open Source IMS Core project. Moreover, UE1 and UE2 are based on Android smartphones running the imsdroid software as IMS clients.

Table II provides a set of illustrative values for the proposed case study. The delay for the transmission of the SIP messages at each hop within the core network is averaged to 5 ms. In order to include different wireless access networks into the analysis, the T_{AN} parameter is defined as a variable between 10 ms and 80 ms. The average values for the UE’s processing times (those marked in Fig. 4) is measured to 10 ms.

TABLE II. DELAY VALUES CONSIDERED FOR SIMULATIONS

T_{AN}	T_{CN}	T_{3PCC}	T_{MRF}	T_{UE}
10-80 ms	5 ms	10-100 ms	15 ms	10 ms

Regarding the 3PCC, different delays are introduced in function of the traffic load ranging from 10 ms to 100 ms in the performed experiments. Finally, the dummy MRF provides an average processing time of 15 ms.

As an illustrative example, Fig. 6 shows the comparison of call setup times of the two alternative operating modes considering that 40% of the sessions in the traffic mix are “Reactive Calls”. The figure gathers the subtraction of the fixed call setup time for the Proactive Approach and the average of call setup times for the Reactive Approach. Positive values in Fig. 6 represent that the call setup times are lower in the Reactive Approach, while negative values indicate that the Proactive Approach exhibits a better performance.

In general higher T_{AN} values benefit the Proactive Approach. This is due to the fact that the “Reactive Calls” requires a higher number of interactions with the called endpoint. Additionally, the effect of the T_{3PCC} is also remarkable. For a fixed value of T_{AN}, higher values of T_{3PCC} entail a better performance of the Reactive Approach. The “Direct Calls” require less 3PCC processing events.

As a result, it can be concluded that the performance comparison in terms of the call setup times is not only a function of the traffic mix pattern, but also depends on the specific delay values of the different individual contributions. Thus, a specific analysis should be carried out for each MRF deployment scenario.

As a step further, we evaluate the threshold traffic mix ratios for two specific mobile access networks. For the first scenario, we fix the T_{AN} value to 50 ms, which is measured as the median value in modern UMTS networks [6]. Regarding the second scenario, we set up the T_{AN} value to 18 ms, based on the 36 ms value reported in [7] as the median RTT in current 4G LTE networks. The T_{3PCC} is also fixed to a constant value of 15 ms, associated to low session handling loads. Meanwhile, T_{CN}, T_{UE} and T_{MRF} remain as in Table II.

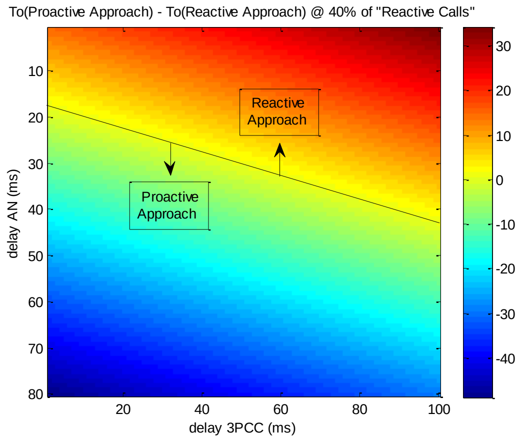

Fig. 6. Comparative analysis of average call setup times

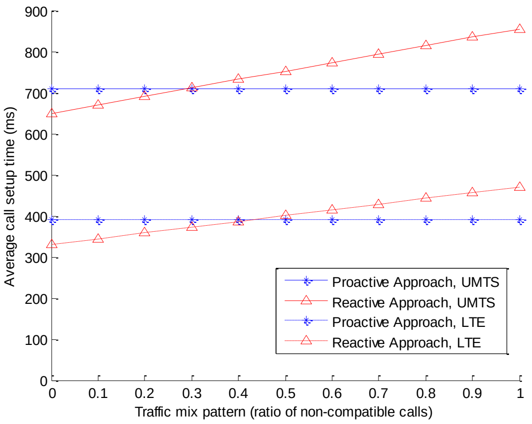

Fig. 7. Required call setup times for selected access network delays

Under these conditions, we analyze the estimated system call setup times for the different possible configurations. As shown in Fig. 7, the overall experienced call setup time is almost the double for the case of UMTS. Considering the Proactive Approach with the assumed delay contributions, every session will suffer a call establishment time of 390 ms with LTE and 710 ms with UMTS. Regarding the Reactive Approach, the average delays experienced are a function of the traffic mix. For LTE, we find a range from 330 ms when all the involved endpoints are compatible to 471 ms when none of the involved endpoints share compatible media characteristics. In the case of UMTS, the call setup time is in the range from 650 ms to 855 ms. Also, it is assumed that any combination of end users using different access network technologies can be easily analyzed following a similar approach.

Considering the specific case study proposed in this section, the thresholds to choose between the two alternative operating modes are around 29.3% of non-compatible calls for the case of UMTS, and 42.5% of non-compatible calls for the case of LTE. These types of values could be used in order to make final decisions about the convenience of one operating mode or the other one, taking also into account the performance results in terms of signaling overhead.

V. Conclusions

This paper has aimed to lay out the basis for a useful tool that may aid system administrators to determine the most suitable approach for the deployment of an IMS-based interoperability framework. Based on the standardized IMS elements and procedures, we analyze the expected performance of two alternative operating modes (i.e. proactive approach and reactive approaches) in different possible scenarios. Different thresholds are determined to select the best performing operating mode, including thresholds in the traffic mix pattern and several individual delay thresholds.

The performance evaluation is carried out taking into account two metrics: the signaling overload and the average call setup times. The former can be directly inferred from a detailed analysis of the required signaling procedures. For the latter, a more thorough analysis is needed since different types of messages, nodes and network segments impute different levels of delay contributions o the overall process.

Section IV of this paper provides a series of results that may be used in similar use cases. First, the ratio of calls between media incompatible UEs determines the average signaling overhead. The traffic mix thresholds for deploying one approach or another one are determined for the considered network architecture. In addition, the different sources of delay are analyzed based on experimental data. Several results are illustrated in order to gauge the combined effects of different variable input parameters. For illustration purposes, a comparative example of two modern radio access networks (i.e. UMTS and LTE) is provided. In general, the traffic mix threshold is higher for UMTS networks, since low delay radio links such as LTE benefit the Reactive Approach.

In addition to the presented performance results, the identification of the individual delay contributions to the different procedures is a remarkable result of the paper. The described methodology can be used in other network scenarios, adapting the individual delay contributions to the specific use case. The analysis of procedures and the type of performance results may be also useful for system administrators. The manager should estimate the specific expected traffic conditions and delay ranges for their systems. In order to make a final decision, they also need to gauge the significance of the signaling overhead and the call setup times for their systems.

To the best of the authors' knowledge, there are no similar studies or tools in the state of the art taking into account the heterogeneous set of input parameters considered in this paper.

References

[1] 3rd Generation Partnership Project (2000-). IP Multimedia Subsystem (IMS); Stage 2. Technical Specification 23.228. 3GPP. http://www.3gpp.org/ftp/Specs/html-info/23228.htm.

[2] Infonetics Research, "Mobile VoIP Services and Subscribers" report , July 2013.

[3] 3rd Generation Partnership Project (2005-). IP Multimedia Subsystem (IMS) emergency sessions. Technical Specification 23.167. 3GPP. http://www.3gpp.org/ftp/Specs/html-info/23167.htm.

[4] 3rd Generation Partnership Project (2001-). IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3. Technical Specification 24.229. 3GPP. http://www.3gpp.org/ftp/Specs/html-info/24229.htm.

[5] 3rd Generation Partnership Project (2009-). IP Multimedia Subsystem (IMS) media plane security. Technical Specification 33.328. 3GPP. http://www.3gpp.org/ftp/Specs/html-info/33328.htm.

[6] P. Romirer-Maierhofer, F. Ricciato, A. D'Alconzo, R. Franzan and W. Karner, "Network-Wide Measurements of TCP RTT in 3G," in Proc. of the First International Workshop on Traffic Monitoring and Analysis (TMA '09), 2009, pp. 17-25.

[7] M. Laner, P. Svoboda, P. Romirer-Maierhofer, N. Nikaein, F. Ricciato and M. Rupp, "A Comparison Between One-way Delays in Operating HSPA and LTE Networks," in Proc. of 10th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2012, pp. 286-292