INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 10(5):686-701, October, 2015.

Adaptive Probe-based Congestion-aware Handover Procedure Using SIP Protocol

R. Libnik, A. Svigelj

Rok Libnik
Telekom Slovenije, d.d.,
Cigaletova 15, 1000 Ljubljana, Slovenia
rok.libnik@telekom.si

Ales Svigelj*
1. Department of Communication Systems,
Jozef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
ales.svigelj@ijs.si
2. Jozef Stefan International Postgraduate School
Jamova cesta 39, 1000 Ljubljana, Slovenia
*Corresponding author: ales.svigelj@ijs.si

Abstract: Wireless technologies have evolved very rapidly in recent years. In the future, operators will need to enable users to use communication services independently of access technologies, so they will have to support seamless handovers in heterogeneous networks. In this paper we present a novel adaptive congestion aware Session Initiation Protocol (SIP) based procedure for handover in heterogeneous networks. In the proposed algorithm the handover decision is based in addition to signal strength, also on target network congestion status, which is tested during the conversation. As SIP protocol was used, the proposed procedure is independent of access technologies. For performance evaluation of proposed procedure we developed a purpose built simulation model. The results show that the use of the proposed adaptive procedure significantly improves the QoE of VoIP users, compared to reference scenario, in which only signal strength was used as the trigger for handover decision.

Keywords: Session Initiation Protocol (SIP), seamless handover, heterogeneous networks, performance evaluation, congestion awareness.

1 Introduction

Fixed, nomadic and mobile telecommunications networks, which provide voice and data services, are nowadays converging toward a seamless heterogeneous telecommunication network. Due to many different wireless access technologies, comprising licenced (e.g. GSM/UMTS, HSPA, WiMAX, and LTE) and unlicensed (e.g. WLAN) access, there is a need for a uniform access to converged services. Such services should be independent of the access technologies, providing seamless connectivity and sufficient quality of experience (QoE). In the future networks the wireless access networks will play the key role, as their inherent characteristics of the limited radio bandwidth and channel properties, are of the paramount importance for the provision of appropriate transmission rates and quality of service (QoS).

In addition, WLAN is increasing its popularity, in particular in home/business environment (limited coverage areas). Thus, to enable mobile users to communicate using a variety of different access technologies, terminals have to support several network interfaces. Terminal manufacturers are already producing multi mode terminals and the number of interfaces is bound to increase with technology limits. Operators, on the other hand are starting to offer fixed mobile converged services. With increasing demand from the users to communicate independent of access technologies, operators will need to offer seamless handover between heterogeneous access networks (i.e. vertical handover). In homogeneous networks handover techniques (i.e. horizontal handover) are well studied in the literature and already integrated in mobile networks. The horizontal handover is usually triggered by received signal strength (RSS) only. In the future, this will not be sufficient and other parameters should be taken into account (network congestion status in our case) in handover decision. Thus, new mechanisms need to be developed.

The main contribution of this paper is an advanced adaptive SIP based procedure for congestion aware handover in heterogeneous networks, together with its performance evaluation in simulation environment.

The remainder of the paper is organized as follows. In the following subsections the related research work on mobility management techniques and support for handovers when using the SIP protocol is described. In Section 2 the novel adaptive congestion aware SIP based procedure for handover in heterogeneous networks is presented, while its performance evaluation is presented in Section 3. The paper ends with conclusions in Section 4.

1.1 Mobility management

Mobility management techniques are defined as techniques that support user movement within and between different networks. The handover process can be in general divided into three phases: (i) Handover information gathering phase, (ii) handover decision phase and (iii) handover execution phase [1].

In the first phase (i.e. handover information gathering phase) a mobile node collects not only network information, but also information about the other components of the system such as network properties, mobile devices, access points, and user preferences [1]. The information/parameters typically collected/measured are the following [1] [2] [3] [4] [5]:

- Availability of neighbouring network
- Received Signal Strength (RSS), Signal Noise Ratio (SNR), Carrier to Interference Ratio (CIR), Signal to Interference Ratio (SIR), Bit Error Ratio (BER),
- Delay, jitter
- Throughput,
- Economic price of the usage of the network.
- The Mobile device's state by gathering information about battery status, resources, speed, and service class.
- User preferences information such as budget and services required, preferred network operator.
- Context information.

As seen, some parameters (SNR, delay, jitter, bandwidth, and power consumption) are network/hardware related and cannot be influenced by the user, while others (price, preferred network operator) represent parameters that can be selected/set by the user. Gathered parameters can be grouped also according to the origin of the parameters. They can be provided by the network (i.e. service independent) or can be provided by application/service (i.e. service dependent).

In the second phase (i.e. handover decision phase) the decision for handover is made, based on criteria function, taking into account different information/parameters, which were gathered during the first phase. The second phase is one of the most critical processes during the handover, as in this phase, the decision about time (if and when) and to which network (selecting the best network fulfilling requirements) the handover is made is taken. In a homogeneous network environment the decision about time usually depends on RSS values, while the selection of the network is not an issue since the same networking technology (horizontal handover) is used. In heterogeneous networks the selection of the appropriate network is quite complex, as many parameters/information obtained from the different information sources (i.e. network, mobile devices, and user preferences) must be evaluated in order to make the best decisions. In this paper we are focusing on vertical handover solutions based on the combination of different parameters. Selection of appropriate parameters and handling of appropriate trigger algorithm in handover decision phase is of a paramount importance, as wrong handover decision can lead into unsatisfied users. This is in particular important when performing handover using real time applications such as Voice over IP (VoIP), where the assurance of appropriate level of Quality of Experience (QoE) in the target access network, represents the main challenge. Operator's backbones usually do not present a bottleneck as they are well maintained and controlled in order to provide an adequate level of QoS.

In the last (third) phase (i.e. handover execution), traffic flow is handed over to the target network. This means that all the traffic is sent using the new connection, while the connection with the old network is terminated. This phase should also guarantee a smooth session transition process.

In this paper we focused in particular on the second phase (i.e. handover decision), proposing new procedure that improves the handover decision and consequently the user experience when performing handover using VoIP applications in heterogeneous networks.

Handover in heterogeneous network can be performed using different protocols and this has been the subject of several studies [1] [5] [6]. At the network layer, Mobile IP (MIP), defined in [8], has been most frequently selected [8] [9] [10] [11] as the protocol for handover. With the modifications presented in [12] it can provide greater support for real time services on a Mobile IPv4 network, by minimizing the period of time when an mobile node (MN) is unable to send or receive IPv4 packets due to delay in the Mobile IPv4 Registration process. In [13] authors proposed an enhancement of Mobile IP (MIP) called MIP with Home Agent Handover (HH-MIP) to enjoy most of the advantages of Route Optimization MIP (ROMIP) but with only a small increase of signalling overhead. The most widely used protocols at the transport layer are TCP and UDP. Both have some limitations for mobility support. However, a new solution called mobile SCTP (mSCTP) has been developed to enable IP addresses to be added, deleted and changed during active SCTP association [14] [15]. For mobility management at the application layer, SIP is usually selected as the most favoured protocol [9] [10] [11] [12] [16] [17] [18]. SIP runs on top of several different transport protocols and is today's most widely used protocol for IP telephony penetrated in both terrestrial and satellite networks [19]. The advantage of using the SIP protocol for handover execution is, that SIP is an application layer protocol, and thus its use does not have a great impact on the network changes needed. The transport independence of SIP means that it does not require great network involvement in handover execution. However, the application usually needs to be improved / customized to support handover and only SIP based application can perform handover. Application level solutions based on tunnelling [20], using SIP only for signalling can overcome this problem. The biggest advantage of using SIP is its wide adoption in real operator environments, since almost all operators that are offering VoIP services are using SIP for signalling. In addition, SIP is used in many operator environments and has been selected as the primary signalling protocol in IMS (IP Multimedia Subsystem) networks. In this paper we focus on solutions that can be deployed easily in a real operator environment. Thus, we decided to focus on the use of SIP for mobility management, which is shortly described in the next subsection.

1.2 SIP mobility

SIP protocol [21] [22] is an application layer signalling protocol for establishing, modifying, and terminating Internet multimedia sessions. These sessions include Internet telephone calls, multimedia distribution, and multimedia conferences. SIP invitations used to create sessions carry session descriptions that allow participants to agree on a set of compatible media types.

When using SIP protocol for IP telephony in operator's environment, the regulatory and security issues are forcing operators to provide additional functionalities that can affect the architecture of the IP telephony solution [23]. Usually, the Session Border Controller (SBC) is added to their network [24], which can lead to a conflict with SIP architectural principles. SIP based SBCs typically handle both signalling and media, resulting in call flow to be changed in a way that all RTP streams are routed via SBC (see also Figure 4).

With minor modifications, SIP can support four types of mobility. Terminal mobility enables devices to move between subnets and be accessible to other hosts and to continue any ongoing session when they move. Session mobility enables users to maintain a session while moving from one terminal to another. Personal mobility allows users to use the same set of services, even when changing devices or network attachment points. Service mobility enables users to be identified by the same logical address, even if the user is at different terminals.

In this paper we are focusing on terminal mobility, for which two types of mobility management have been defined: pre-call mobility and mid call mobility. SIP protocol supports several applications; however, in this paper we have selected IP telephony as a typical representative of real time application.

In the pre-call mobility scenario MN gets a new IP address prior to the call, thus this operation does not affect the quality of IP telephony service and will therefore not be discussed further.

In the mid call mobility scenario first a call is established between the corresponding node (CN) and the MN that is in the home network. When MN moves to the target network, it gets new IP address and sends SIP re INVITE message to CN and informs it about the location change. The new RTP session is then established. The limitation of this approach is that the SIP server is not informed about the location change. Some solutions have been presented in the literature in which MN informs the SIP server about the location change after sending the SIP re INVITE message [13]. However, in a real operator environment, information about location change needs to be sent to the SIP server prior to starting a new SIP session between MN and CN. This should be done, for example, to support proper charging, since prices can differ between networks.

To overcome the limitation of mid call mobility in [25] we have proposed SIP enhanced mid call scenario (SEMCS)

In SEMCS scenario, a call is established between the CN and the MN that is in the home network After moving to target network, the MN sends the SIP re-INVITE message to the SIP server to inform it about the location change. The SIP server then forwards the SIP re INVITE message to the CN . After the acknowledgement the new RTP session is established using (new) IP address of the target network.

In [25] we proposed message exchange via SIP server and we focused on handover execution based on SNR ratio only. That procedure was upgraded in [26], where we are presenting a novel procedure for handover decision based on congestion detection (CAHP-C). In this paper, we are extending the CAHP-C procedure described in [26] with an adaptive procedure for congestion aware handover.

2 CAHP-A adaptive procedure for congestion aware handover

As stated in the introduction in today's mobile/wireless homogenous networks the trigger for handover is usually based on SNR only. This is sufficient as the whole network uses the same access technology and is under control of one operator offering service. In the future heterogeneous networks, the target access network could be covered by different operator or it is not even operator's network. The later is usually the case with open WLAN network in a hotel or in a congress centre, where there are several access points (AP) connected to single xDSL connection. On one hand the users can have good signal coverage and very fast (e.g. 52 Mbit/s) connection between MN and AP, while on the other hand, when traffic from all APs is gathered on single xDSL line bandwidth becomes a bottleneck. Thus, the SNR level is not sufficient parameter for handover decision and should not be performed only when SNR level exceeds the threshold, but should be done based also on other parameters (e.g. congestion status) [26]. This is especially important when providing real time applications (IP telephony in our case), where ability to provide sufficient QoE for the user is the most important and where performing handover to congested network could lead to degraded service level.

Our solution is described in [26], where we defined basic congestion aware handover procedure (CAHP-C), which enables efficient handover performance, taking into account also the congestion status of the target network. It should be noted, that receiving signal power measurement remains the prerequisite for handover (i.e. handover cannot take place to a network lacking or with limited signal coverage).

In our study two groups of networks with different characteristics were defined. Networks in the first group are reliable and expensive (e.g. UMTS, HSPA and LTE), while networks in the second one are cheaper or even free of charge and unreliable (e.g. WLAN). In order to present our solution more clearly we selected one representative from each group. The HSPA was selected from the first group of networks and WLAN from the second group. Those two will be used in the rest of this paper. Please note, that our approach can be used between any two networks (e.g. also between different WLAN networks). We assumed that congestion (i.e. QoE degradation) can happen in the WLAN network only. We are not focusing on "who" or "what" is causing the additional delay (e.g. MAC, TCP, routing, low bandwidth on the access link), but only on the fact that if there is a delay, which is not acceptable for IP telephony, we do not use the corresponding access network.

As the proposed CAHP-A procedure is an extension to CAHP-C procedure we provide short description of CAHP-C emphasizing only the main characteristics, as partially depicted in Figure 1, while in depth description of CAHP-C can be found in [26].

2.1 CAHP-C

The CAHP-C procedure is used only if SNR exceeds the predefined threshold [26]. When SNR is below threshold, the proposed procedure is not used. In order to detect possible congestion in the WLAN network before the handover is executed, we proposed Pre-probe algorithm. The congestion is detected with delay testing. In order to monitor congestion (i.e. round trip delay measurement) we defined new SIP message named SIP *pre_ PROBE* , which is sent before the SIP re INVITE message. As characteristics of the WLAN access network can also change during the call, we defined another algorithm named Mid-probe algorithm. Another SIP message called *SIP mid_ PROBE* was defined, which is used to check the congestion status of the WLAN access network when in use. After receiving the responses the MN calculates the average delay D_{pre} (Pre-probe algorithm) or D_{mid} (Mid-probe algorithm) from MN to SBC. We also defined two parameters T_{pre} and T_{mid}. When user is trying to handover to WLAN network, the parameter T_{pre} defines the period, when SIP *pre_ PROBE* messages are sent again if measured delay is above predefined threshold T_d (i.e. 200ms in our case) [26]. When user is already using WLAN network the parameter T_{mid} defines the period when *SIP mid_ PROBE* messages are sent again if the measured delay is below T_d [26]. Those two parameters are the most important in CAHP-C procedure and need to be set carefully as they affect the level of signalling overhead and the speed of detecting possible congestion. The main limitation of CAHP-C procedure was that

T_{pre} and T_{mid} parameters are constant, which means that they do not change if the network characteristics are improved. Thus, we are proposing CAHP-A procedure in which T_{pre} and T_{mid} parameters change adaptively according to current network characteristics (see also Figure 1). The CAHP-A procedure is presented in the next chapter.

2.2 CAHP-A

The SIP *pre_PROBE* and *SIP mid_PROBE* messages present additional traffic in the network and can be seen as signalling overhead. As the size of the messages is the same as in RTP traffic, the use of newly defined messages does not have significant effect on backbone traffic. However, such increase of signalling affects the SBC, which needs to handle additional messages. In this paper we are extending the CAHP-C procedure with adaptive calculation of T_{pre} and T_{mid}, as depicted in Figure 1 (bolded blocks).

In order to keep the signalling overhead as low as possible we propose that parameters T_{pre} and T_{mid} change adaptively according to characteristics of network (i.e. measured delay) as the networks differ by the ability to provide sufficient QoE. Some of them get congested likely (e.g. in congress centre) while the other used by a single user (e.g. at home) are not. With the Pre probe algorithm WLAN network is tested prior handover.

In the case that calculated delay D_{pre} is above threshold T_d the parameter T_{pre} change according to difference between D_{pre} and T_d. After handover, the WLAN network is tested with Mid probe algorithm. In case that calculated delay Dmid is below threshold Î¤d the parameter Tmid should change according to difference between D_{mid} and T_d. The calculations happen after the delay measurements (see bolded block in Figure 1).

In order to achieve such dependency we defined equations (1) and (2) which are used for calculation of T_{pre} and T_{mid} respectively:

$$T_{pre} = \begin{cases} N/A; & D_{pre} \leq D_{max} \\ T_{max} \cdot \left(\frac{D_{pre}-D_{min}}{D_{max}-D_{min}} - 1\right); & D_{max} < D_{pre} < 2 \cdot D_{max} - D_{min} \\ T_{max}; & D_{pre} \geq 2 \cdot D_{max} - D_{min} \end{cases} \tag{1}$$

$$T_{mid} = \begin{cases} T_{max}; & D_{mid} \leq D_{min} \\ T_{max} \cdot \left(1 - \frac{D_{mid}-D_{min}}{D_{max}-D_{min}}\right)^{\alpha}; & D_{min} < D_{mid} < D_{max} \\ N/A; & D_{mid} \geq D_{max} \end{cases} \tag{2}$$

where T_{max} represent maximum possible time set for T_{pre} or T_{mid}, D_{min} minimum (i.e. delay of non congested network) and D_{max} maximum delay that does not have effect on QoE (T_d in proposed algorithm). As seen the values for T_{pre} and T_{mid} are function of α and measured D_{pre} or T_{mid}. By selecting different values for α we can define different curves that define T_{pre} and T_{mid}. It is worth noting, that different α can be set for calculating T_{pre} and T_{mid}. Some possibilities are presented in Figure 2 and Figure 3, where parameter α was set to 1/16, 1/8, 1/4, 1/2, 1, 2, 4, 8, and 16.

When T_{pre} and T_{mid} are low more SIP probe messages are sent as they are more frequent and the possible degradation of service is detected faster. In such a manner we managed to maintain QoE of the user. When fewer messages are sent (T_{pre} and T_{mid} are high) the degradation of service may not be detected fast enough and degradation of QoE can be expected. However, more frequent sending of SIP *pre_PROBE* and *SIP mid_PROBE* messages T_{pre} and T_{mid} are high) increases signalling overhead. From Figure 2 and Figure 3 it can be seen that level of overhead changes also with parameter α. Thus, the α can be seen as the parameter that defines signalling overhead (i.e. the bigger the α, the bigger the signalling overhead). When α is approaching 0

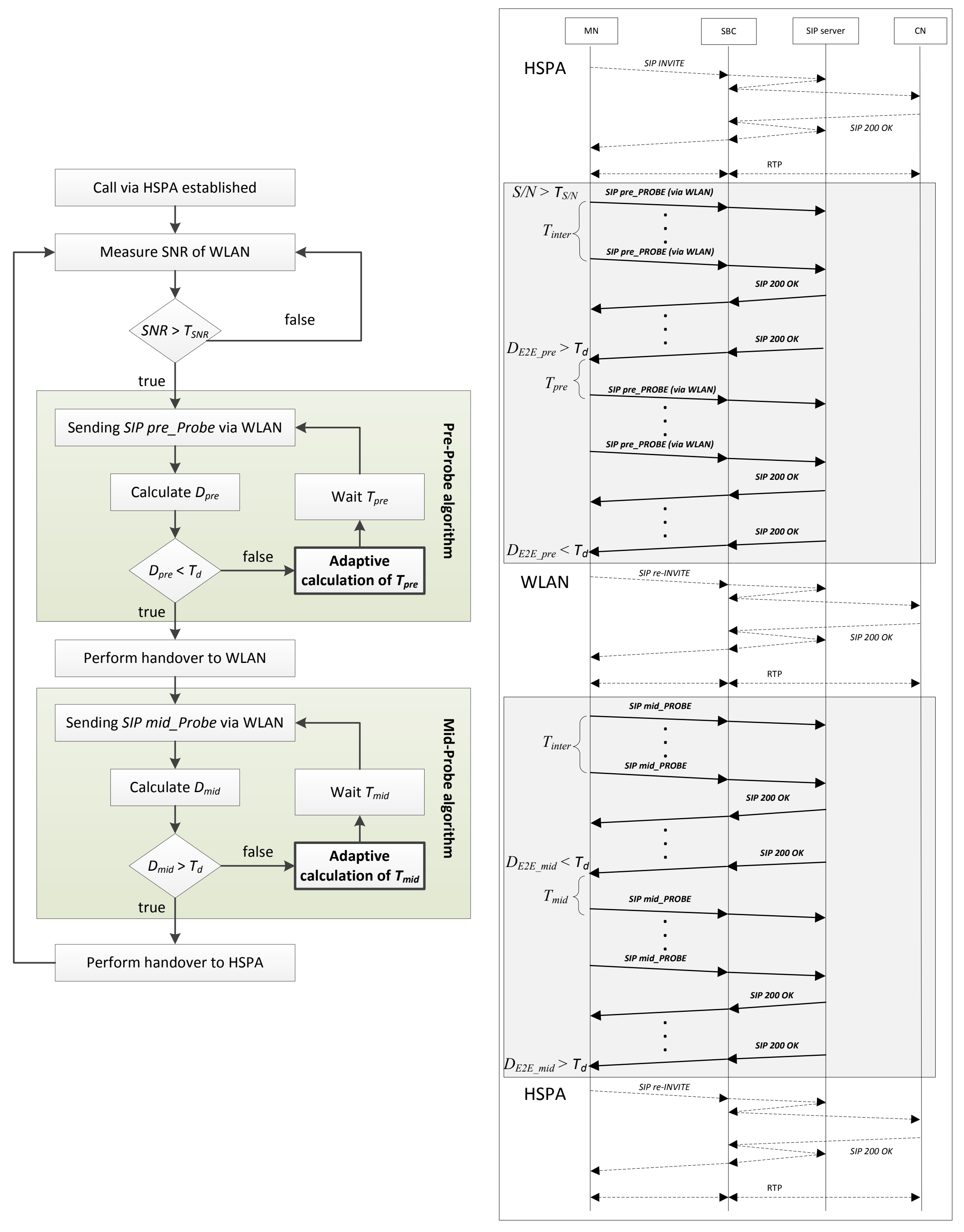

Figure 1: Flow diagram of CAHP-A procedure and message exchange.

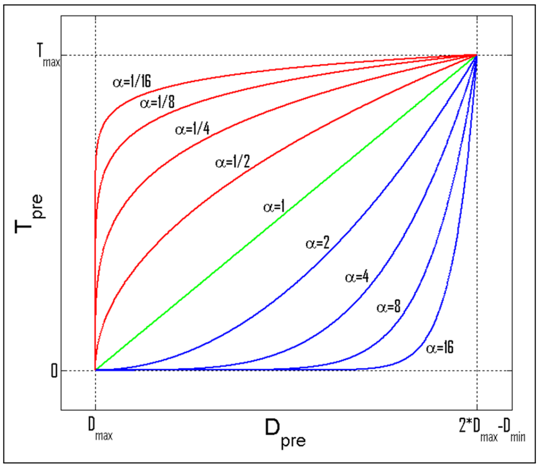

Figure 2: Dependency between SIP D_{pre} and SIP T_{pre} for different values of α

the values for T_{pre} and T_{mid} are approaching T_{max} and when α is approaching infinity the values the values for T_{pre} and T_{mid} are approaching 0 s.

In the proposed handover CAHP-A procedure two assumptions were made: (i) the procedure starts when MN is connected to both current and target networks and (ii) MN is capable of sending SIP messages via WLAN interface while, for RTP the HSPA network is still used. Both functionalities were necessary to perform seamless handover from HSPA to WLAN network. In the case that MN would not be capable of connecting to two different networks at the same time, first HSPA connection should be terminated and only then the establishment of WLAN connection would start. By second functionality the MN sends newly defined and standard signalling SIP messages via WLAN when still connected to HSPA. If MN would not have such capabilities, again HSPA connection should be terminated first and SIP messages could be sent only when WLAN connection would be established. Without those functionalities the connection would be lost several times which would affect QoE of the user.

The SIP *pre_PROBE* and SIP *mid_PROBE* messages are sent to SBC, through which RTP packets are also sent. Thus, the use of the proposed probes in the target networks can provide the real status of the ability of the network to provide an adequate level of QoE for IP telephony. As SIP protocol was used for sending the newly proposed messages, this approach is completely independent of the lower layers (i.e. transport, network and link). Furthermore, it can be used independently of the protocol used in lower layers and easily integrated in the operator's environment. It means that the role of lower layers protocols stays the same (i.e. providing IP connectivity) and that they do not need to be modified. The application uses theirs functionalities (e.g. measurement of signal, establish physical connection, IP connection).

Our proposal is also easily scalable, as operators add additional SBCs to their network, when the number of users and traffic flows increases and existing SBCs can not serve all the signalling and RTP traffic load.

3 Performance evaluation of the CAHP-A procedure

For performance evaluation of the proposed handover procedure presented in section 2 we developed a simulation model of a telecommunication system, which is discussed in [26]. The simulation model comprises two networks, WLAN and HSPA. Two hosts, MN and CN, that are using IP telephony as an application, were also defined. The G.711 codec was used for VoIP.

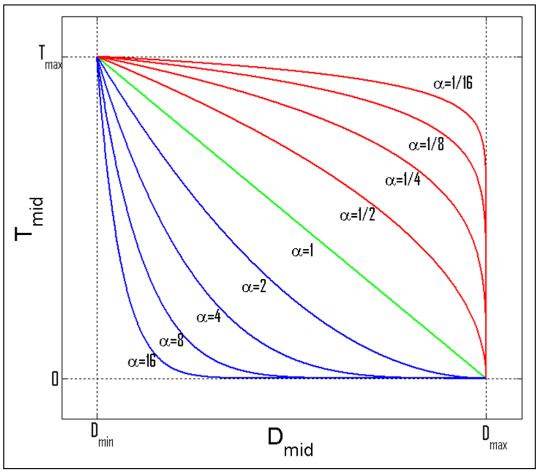

Figure 3: Dependency between D_{mid} and T_{mid} for different values of α

Silence suppression was not used, resulting in a constant RTP traffic stream of 100 packets/s. The network architecture of the simulation model, resembling real operator environment, is presented in Figure 4.

The following assumptions were made in the simulation scenario:

- HSPA network is always available;
- WLAN network has limited coverage, but its access link between Router 1 and IP network representing fixed operator, could become congested, as it aggregates traffic from all WLAN access points (AP);
- The usage of WLAN network is prioritized by the user, which means that MN will always try to perform a handover when this network is available;
- MN is a dual mode handset capable of sending RTP packets and SIP INVITE messages at the same time via different interfaces.

The first two assumptions were made just for simulation scenario in order to validate proposed procedure in environment of two networks, where one is reliable and the second is unreliable. We focused only on access link traffic load, as this is usually critical part of connection from MN to SBC. As described in chapter 2, in proposed procedure the congestion will be detected by measuring delays from MN to SBC.

The simulation model of the communication system was developed using the discrete event, object-oriented modelling simulation tool OPNET Modeler [26]. It has an open source code of commonly used protocols, which is very convenient for performance evaluation of user developed/enhanced mobility management mechanisms [28]. It enables network modelling and simulation for designing new protocols and technologies, together with performance evaluation of existing and newly developed optimized protocols and applications.

OPNET supports SIP telephony but it does not support handover on the application layer, thus some pre-defined process models that incorporate SIP procedures were customized [25]. Beside modification of process's functionalities, we also define the newly proposed SIP *pre_ PROBE* and *SIP mid_ PROBE messages*, which are used for congestion testing of the WLAN access network.

Figure 4 shows the simulation network configuration. The MN is a dual mode terminal, capable of connecting to WLAN and to HSPA network. Both networks are connected via IP networks to the SBC. The CN is an IP phone connected to SBC.

To increase traffic in the WLAN network access link, other clients were added (represented by laptops in Figure 4), which generated additional UDP traffic in the LAN network.

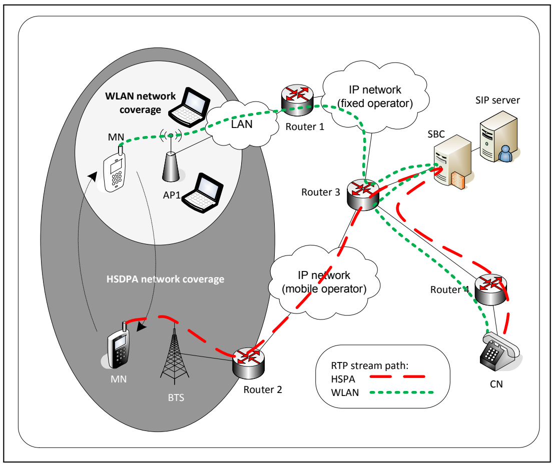

Figure 4: Network architecture of the simulation model.

3.1 Simulation scenario

The proposed CAHP-A procedure was evaluated in simulation model. In order to get statistically representative results, one long call that lasted 28,800 s (i.e. 8 hours) was simulated. However, the results can be easily applied to the scenario with several users that are making calls with the total sum of conversation time 28,800 s.
In defined network architecture WLAN network access link was randomly congested with additional UDP traffic, which was generated by additional VoIP clients. The duration of each "congestion" was distributed exponentially with mean value of 20 s. The time between two congestions was also distributed exponentially with mean value of 10 s. Such distribution enabled us to test the proposed procedure several times in different traffic conditions. The user movements were defined by changing SNR ratio. The SNR values were measured in real environment and imported in the simulation model.

Table 1: Simulation scenarios

Scenario	T_{mid}(s)	T_{pre}(s)	α
R_1	N/A	N/A	N/A
R_2	0	0	N/A
S_1	adaptive	adaptive	1/16
S_2	adaptive	adaptive	1/8
S_3	adaptive	adaptive	1/4
S_4	adaptive	adaptive	1/2
S_5	adaptive	adaptive	1
S_6	adaptive	adaptive	2
S_7	adaptive	adaptive	4
S_8	adaptive	adaptive	8
S_9	adaptive	adaptive	16

In order to evaluate the proposed procedure, we prepared eleven scenarios. Two first scenarios were defined to get reference results for two opposite situations. In the first reference scenario (R_1) the proposed procedure was not used and handover was made based on SNR only. In the second scenario (R_2) parameters T_{pre} and T_{mid} were set to 0 s, which gave us maximal sending frequency of newly defined SIP messages SIP *pre_PROBE* and SIP *mid_PROBE*. We added additional 9 scenarios (S_1 - S_9) with CAHP-A procedure in which parameters α was set to 1/16, 1/8, 1/4, 1/2, 1, 2, 4, 8 and 16. All simulation scenarios are summarised in Table 1.

3.2 Simulation results

In the simulation several results were collected. From the user's point of view the end to end delay should not be significantly affected by performing handovers. We measured end to end delay of IP telephony for each packet. This enabled us to get the cumulative simulation time with end to end delay above 200 ms, which will be presented in results. The cumulative time when HSPA interface was used will be also presented as this affects the cost of communication, which is also very important to the user, as he wants to minimise the usage of expensive network (i.e. HSPA), while still having the appropriate QoE. The signalling increase caused by additional newly proposed messages during handovers was also tracked. If users perform large number of handovers the probe messages will increase signalling traffic and traffic load of the SBC, which is important from the operator's point of view. Thus, the number of overhead messages will be also presented in the results.
The end to end delay distribution for simulation scenarios is presented in Figure 5. For the sake of clarity only delays above 200 ms are presented, as those delays significantly affect QoE of the user. Delays above 400 ms occur in R_1 only, in which cumulative time, with end to end delay above 200 ms, presents 48.7% of all communication time. That means that in such conversation the QoE of the user is highly degraded. It can be seen that in all other scenarios the cumulative conversation time above 200 ms is much smaller (see percentages above histograms in Figure 5), resulting in QoE to be highly improved compared to R_1. The best results are measured in scenarios from S_7 to S_9. It can be seen that results in S_7 to S_9 are even better than in R_2, where T_{pre} and T_{mid} were set to 0 s as in R_2 the signalling overhead itself was causing additional congestion and deteriorate the results.

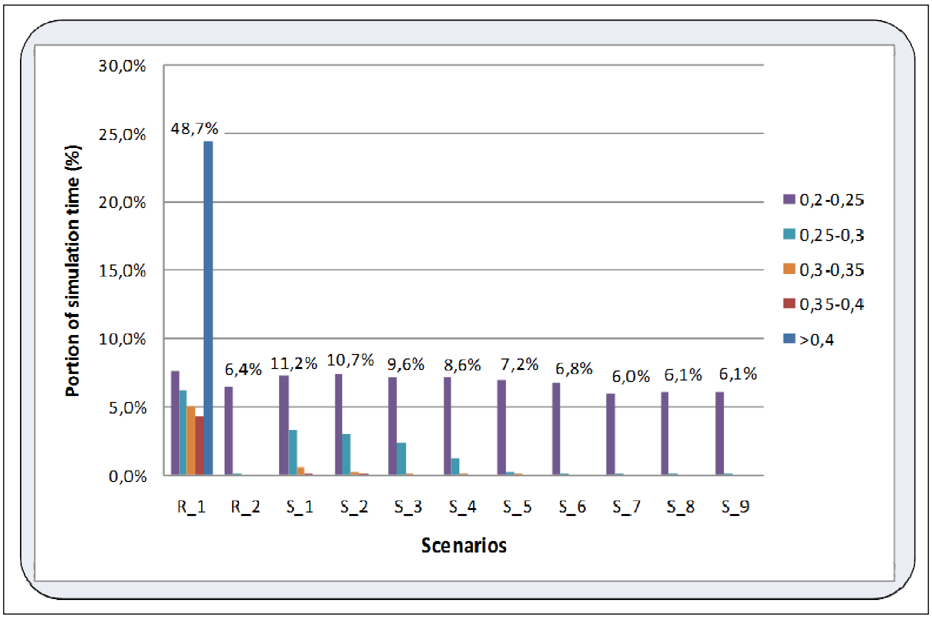

Figure 5: End to end delay distribution

The biggest share of cumulative simulation time above 200 ms is measured in R_1, thus we took this scenario as a reference for normalizing other scenarios as presented in (1):

$$NSTaT_i = \frac{STaT_i}{STaT_{R1}} \cdot 100\% \tag{3}$$

where $NSTaT\hat{I}\text{¤}_i$ normalized simulation time above threshold (200 ms) of scenario i, $STa\hat{I}\text{¤}_i$ simulation time above threshold for scenario i, $STa\hat{I}\text{¤}T_{R1}$ simulation time with delay above 200 ms for scenario R_1. Normalized values are presented in Table 2.
It can be seen that the $NSTa\hat{I}\text{¤}_i$ is decreasing with increase of parameter α in scenarios S_1-S_9. The best results are achieved in S_7, where α was set to 4.
Cumulative conversation time when HSPA interface is used is also presented in Table 2. In R_1 the HSPA was used only for 16.4% of simulation time, as handovers were performed based on SNR only. In all other scenarios (R_2 and S_1 - S_9) the measured HSPA usage is between 52.0% and 59.6%, which give us only 7.6 percentage points of difference.

Table 2: Normalized cumulative simulation time above 200 ms

Scenario	$NSTa\hat{I}\text{¤}T_i$	Cum. conv. time on HSPA
R_1	100.0%	16.4%
R_2	13.1%	59.6%
S_1	23.1%	52.0%
S_2	21.9%	52.6%
S_3	19.7%	55.8%
S_4	17.6%	55.0%
S_5	14.7%	57.4%
S_6	13.9%	53.7%
S_7	12.3%	58.1%
S_8	12.6%	59.3%
S_9	12.5%	57.7%

Normalized signalling overhead caused by newly proposed SIP*pre_ PROBE* and SIP *mid_ PROBE* messages is presented in Figure 6. Maximum number of signalling overhead messages is, as expected, measured in R_2 ($T_{pre} = T_{mid} = 0$ s) in which about 273 thousand overhead messages were sent. This scenario was taken as a reference for normalizing other scenarios as presented in (2):

$$NSO_i = \frac{nSM_i}{nSM_{R2}} \cdot 100\% \tag{4}$$

where NSO_i presents normalized signalling overhead for scenario i, nSM_i number of signalling messages of scenario *i*, nSM_{R2} number of signaling messages of scenario R_2.
From Figure 6 it can be seen that the signalling overhead is highly decreased in scenarios form S_1 to S_9 (where CAHP-A was used), with lowest overhead in S_1 (only 4.4 % of R_2 signalling traffic). The results show that parameter α defines signalling overhead (i.e. the bigger the α, the bigger the signalling overhead) proofing our theoretical analysis in chapter 2. As in the reference scenario R_1 CAHP-A procedure was not used, no messages were sent, thus we get result of 0.0

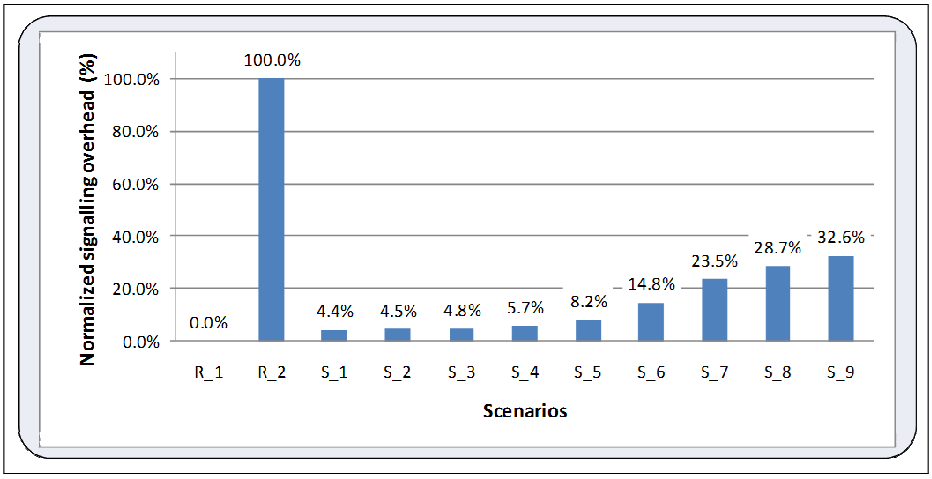

Figure 6: Normalized signalling overhead

3.3 Discussion

Among presented results we are looking for optimal values for parameter α (defining T_{pre} and T_{mid}). As the results presenting simulation time when HSPA was used (Table 2) are not very dependent on T_{pre} and T_{mid} (7.6 percentage points of difference among scenarios), these results are excluded from further analysis. From the results of end to end delay distribution (Figure 5) and normalized signalling overhead (Figure 7) it can be seen, that high signalling overhead results in low end to end delay (i.e. good QoE) and vice versa. Thus, compromise will be needed between user's and operator's objectives.

Users want QoE to be as good as possible. To assure this, the end to end delay should be sufficiently low during VoIP session also when using unreliable network (i.e. WLAN network in or case). In all scenarios we measured end to end delay between 250 and 300 ms, thus end to end delay of 300 ms presents upper limit for sufficient QoE. From results (Figure 5) it can be seen that scenarios R_2 and S_6 - S_9 meet that condition as no packet appeared with delay above 300 ms.

From the operators point of view the signalling overhead needs to be low in order to save SBC resources. The limit for signalling overhead is harder to set, thus we defined two conditions. Our requirement is that signalling overhead is lowered for at least (a) 70% and (b) 80% compared to R_2. Results presented in Figure 6 show that scenarios S_1 to S_8 all meet condition (a) and scenarios S_1 to S_6 meet condition (b).

As seen, scenario S_6 appears in all groups. Based on defined conditions the optimal setting of parameter α is 2. By using those setting, cumulative time with measured end-to end delay above 200 was decreased for 86% (from 48.7% to 6.8%) compared to reference scenario R_1, while signalling overhead was lowered for 85% compared to reference scenario R_2.

The comparison of the results for optimal setting ($\alpha = 2$) with the results obtained for different constant values of T_{pre} and T_{mid} described in [26] shows, that when using CHAP-A, the end-to-end delay is significantly decreased, in particular for the delays higher than 250 ms, while the signalling overhead stays almost the same.

4 Conclusions

In this paper we presented a novel adaptive SIP based procedure for congestion aware handover in heterogeneous networks. With newly proposed Pre-probe and Mid-probe algorithms the handover decision is (in addition to SNR) based also on target network congestion status. In order to analyze and evaluate the proposed procedure several scenarios were prepared. The results show that using the proposed adaptive CAHP-A procedure, the QoE of VoIP users was significantly improved, compared to reference scenario, in which only signal strength was used as the decision for handover. This is achieved by eliminating handovers to unreliable highly congested target network, which could cause degradation of service. With the CAHP-A procedure the unreliable network is tested also after handover. In the case of the detection of congestion the handover back to reliable network is triggered, which prevents the QoE degradation. In the proposed procedure SIP protocol was used for sending the proposed probe messages. As it runs on the application layer, our solution is completely independent of the underlying access technologies and thus applicable easily to next generation wireless systems. Furthermore, it is also independent of the lower layer protocols used. Another advantage of using SIP protocol for messages is that SIP usage is increasing in operators environments. In this paper we have focused only on measurement of end-to-end delay based on transmission of the proposed SIP*pre_ PROBE* and SIP*mid_ PROBE* messages. In order to further improve the handover decision algorithm, we will in our further work evaluate more parameters (e.g. user profiles, other network parameters, context awareness) to make even more efficient handover decision. In addition, the historic information about particular network can also be used in order to make the handover decision even more efficient, in particular where there are more than two WLAN networks available.

Bibliography

[1] J. M. Barja , C. T. Calafate, J.-C. Cano, P. Manzoni (2011); An overview of vertical handover techniques: Algorithms, protocols and tools, *Computer Communications*, 34(8):985-997

[2] Q.-T. Nguyen-Vuong et al. (2008); A user-centric and context-aware solution to interface management and access network selection in heterogeneous wireless environments, *Computer Networks*, doi:10.1016/j.comnet.2008.09.002

[3] E. Aruna, R.S. Moni (2012); Optimization of Vertical Handoff Decision Algorithm for Wireless Networks, *International Journal of Computers Communications & Control*, 7(2):218-230, DOI: http://dx.doi.org/10.15837/ijccc.2012.2

[4] F. Patriarca, S. Salsano, F. Fedi; Efficient Measurements of IP Level Performance to Drive Interface Selection in Heterogeneous Wireless Networks;

[5] S. Ghahfarokhi, N. Movahedinia (2012); Context gathering and management for centralized context-aware handover in heterogeneous mobile networks. *Turk J Elec Eng & Comp Sci*; 20(6), doi:10.3906/elk-1101-1042

[6] C. Lozano-Garzon, N. Ortiz-Gonzalez, Y. Donoso (2013); A Proactive VHD Algorithm in Heterogeneous Wireless Networks for Critical Services, *International Journal of Computers Communications & Control*, 8(3): 425-431, DOI: http://dx.doi.org/10.15837/ijccc.2013.3

[7] C. Perkins, IP Mobility Support for IPv4, RFC 3344, August 2002.

[8] I. F. Akyildiz, J. Xie, S. Mohanty (2004); A survey of mobility management in next-generation all-IP-based wireless systems, *IEEE Wireless Communications*, 16-28.

[9] H. Fathi, R. Prasad, S. Chakraborty (2005); Mobility management for VoIP in 3G systems: evaluation of low-latency handoff schemes, *IEEE Wireless Communications*, 96-104.

[10] P.M.L. Chan, R.E. Sheriff, Y.F. Hu, P. Conforto, C. Tocci Mobility management incorporating fuzzy logic for heterogeneous a IP environment. *IEEE Communications Magazine* 01/2002; DOI:10.1109/35.968811

[11] S. Mohanty, I. F. Akyildiz (2007); Performance Analysis of Handoff Techniques Based on Mobile IP, TCP-Migrate, and SIP, *IEEE Tranactions on mobile computing*, 6(7):731-747.

[12] K. El Malki (2005); Low Latency Handoffs in Mobile IPv4, INTERNET-DRAFT, October 2005.

[13] J.-Y. Chen , C.-C. Yang, L.-S. Yu (2010); HH-MIP: An Enhancement of Mobile IP by Home Agent Handover. *EURASIP Journal on Wireless Communications and Networking*. 2010. doi:10.1155/2010/653838

[14] L. Ma, F. Yu, V. C. M. Leung, Tejinder Randhawa (2004); A new method to support UMTS/WLAN vertical handover using SCTP, *IEEE Wireless Communications*, 44-51.

[15] S. J. Koh, M. Jeong Lee, Maximilian Riegel, Mary Li Ma, Michael Tuexen (2004); Mobile SCTP for Transport Layer Mobility, *IETF* Internet Draft.

[16] J. Zhang, H. Chan, V. Leung (2007); A SIP-based seamless-handoff (S-SIP) scheme for heterogeneous mobile networks. *IEEE WCNC*.

[17] G. P. Silvana, H. Schulzrinne, SIP and 802.21 for Service Mobility and Pro-active Authentication, *Communication Networks and Services Research Conference*, 2008, doi 10.1109/C-NSR.2008.61

[18] N. Banerjee, W. Wu, K. Basu, S. K. Das (2004); Analysis of SIP-based mobility management in 4G wireless networks, *Computer Communications*, 27, 697-707.

[19] Ali M, Liang L, Sun Z, Cruickshank H. (2011); Optimization of SIP Session Setup for VoIP over DVB-RCS Satellite Networks, *Inderscience International Journal of Satellite Communications Policy and management (IJSCPM)*, 1 (1): 55-76. doi: 10.1504/I-JSCPM.2011.039741

[20] M. Bonola, S. Salsano, A. Polidoro. UPMT: Universal Per-application Mobility management using Tunnels. *IEEE GLOBECOM* 2009, 30 Nov - 4 Dec 2009, Honolulu, Hawaii.

[21] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, SIP: Session Initiation Protocol, *IETF RFC3261*, June 2002.

[22] I. Basicevic, M. Popovic, (2008); Use of SIP Protocol in Development of Telecommunications Services, *The Journal of The Institute of Telecommunications Professionals*, 01/2008, 2.

[23] T. Aljaz, B. Imperl, A. Svigelj (2008); Border gateway function performance requirements for the lawful intercept of voice at IMS architecture. *AEU, Int. j. electron. commun.* (Print), 62(8):610-621, doi: 10.1016/j.aeue.2007.08.006.

[24] J. Hautakorpi, G. Camarillo, R. Penfield, A. Hawrylyshen, M. Bhatia, Requirements from SIP (Session Initiation Protocol) Session Border Control Deployments, Internet-Draft, October 23, 2008, URL: http://www.ietf.org/internet-drafts/draft-ietf-sipping-sbc-funcs-07.txt

[25] R. Libnik, A. Svigelj, G. Kandus (2008); Performance evaluation of SIP based handover in heterogeneous access networks, *WSEAS transactions on communications*, 7(5): 448-458.

[26] R. Libnik, A. Svigelj, G. Kandu (2010); A novel SIP based procedure for congestion aware handover in heterogeneous networks. *Computer Communications*, 33(18): 2176-2184, doi: 10.1016/j.comcom.2010.09.007.

[27] OPNET Technologies (2014); http://www.opnet.com.

[28] R. Libnik, G. Kandus, A. Svigelj (2011); Simulation model for performance evaluation of advancde SIP based mobility management techniques, *International Journal of Communications*, 5(1):26-35.