Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT•

Computer Communications 27 (2004) 697–707

computer communications

www.elsevier.com/locate/comcom

Analysis of SIP-based mobility management in 4G wireless networks

Nilanjan Banerjee*, Wei Wu, Kalyan Basu, Sajal K. Das

Center for Research in Wireless Mobility and Networking (CReWMaN), Department of Computer Science and Engineering,
The University of Texas at Arlington, Arlington, TX 76019-0015, USA

Received 23 October 2003; accepted 24 October 2003

Abstract

Providing seamless mobility support is one of the most challenging problems towards the system integration of fourth generation (4G) wireless networks. Because of the transparency to the lower layer characteristics, application-layer mobility management protocol like the Session Initiation Protocol (SIP) has been considered as the right candidate for handling mobility in the heterogeneous 4G wireless networks. SIP is capable of providing support for not only terminal mobility but also for session mobility, personal mobility and service mobility. However, the performance of SIP, operating at the highest layer of the protocol stack, is only as good as the performance of the underlying transport layers in such a heterogeneous environment. In this paper we analyze the handoff performance of SIP in a IP-based 4G network with Universal Mobile Telecommunication System (UMTS) and Wireless LAN (WLAN) access networks. Analytical results show that the handoff to a UMTS access network introduces a minimum delay of 1.4048 s for 128 kbps channel, while for handoff to a WLAN access network the minimum delay is 0.2 ms. In the former case the minimum delay is unacceptable for streaming multimedia traffic and requires the deployment of soft-handoff techniques in order to reduce the handoff delay to a desirable maximum limit of 100 ms.

Keywords: Application layer mobility management; Streaming multimedia; Handoff

1. Introduction

Fuelled by the advancement of wireless technologies and the emergence of multimedia data services, cellular wireless networks have evolved to their third generation (3G) in just two decades. However, comprehensive 3G wireless networks are yet to be available due to the costly deployment and upgrade of already deployed system equipment. It may also be possible that 3G technology will never be fully deployed. Other predictions foresee a 'generation jump' directly to 4G wireless networks [18,24]. The major task towards 4G architecture is system integration [19,22], where a unified wireless access system is to be established through the integration of the services offered by current access technologies such as General Packet Radio Service (GPRS), CDMA2000 or Wireless LAN (WLAN) as well as future wireless access technologies such as Universal Mobile Telecommunication System (UMTS). The trend towards packet switched technologies and increasingly general use and acceptance of the Internet Protocol (IP) indicate that different wireless access networks are to be connected to an IP-based core network, namely the Internet. Conceptually, a 4G wireless network architecture can be viewed as many overlapping wireless Internet access domains as shown in Fig. 1. In this heterogeneous environment, a mobile host (MH) is equipped with multiple (often called multi-mode) wireless interfaces to connect to any or all wireless access networks anytime anywhere. Therefore, providing *seamless mobility support* is one of the most challenging problems for the system integration in 4G wireless networks.

Several mobility protocols have been proposed for wireless Internet [8,10,11,15,17,21,25]. Although these protocols have the common goal of location transparency, they differ a lot from each other due to choices made during design and implementation phases. These protocols can be broadly classified based on the layer of their operation, such as those operating in the network layer [15], transport layer [21] and application layer [11]. The dependency of these mobility protocols on the access networks reduces progressively as we move up on the protocol stack [5].

* Corresponding author.
E-mail addresses: banerjee@cse.uta.edu (N. Banerjee), wuwei@cse.uta.edu (W. Wu), basu@cse.uta.edu (K. Basu), das@cse.uta.edu (S.K. Das).

0140-3664/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2003.10.009

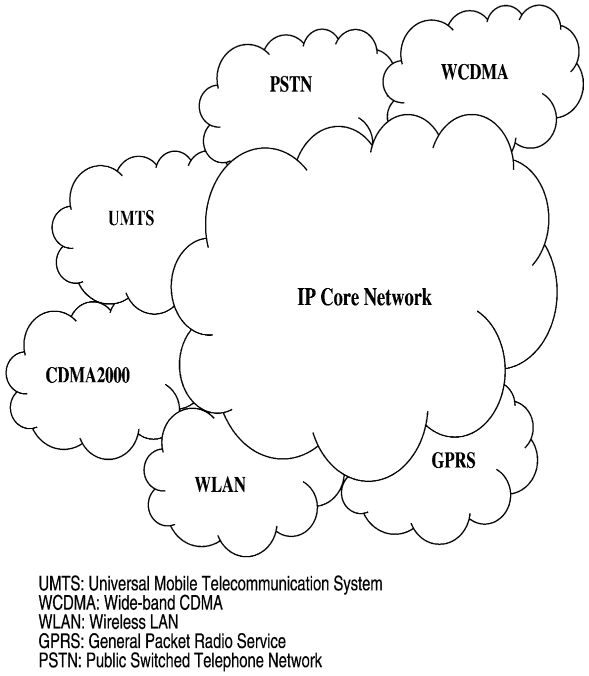
PSTN
WCDMA
UMTS
IP Core Network
CDMA2000
WLAN
GPRS
UMTS: Universal Mobile Telecommunication System
WCDMA: Wide-band CDMA
WLAN: Wireless LAN
GPRS: General Packet Radio Service
PSTN: Public Switched Telephone Network

Fig. 1. Conceptual view of a 4G wireless network architecture.

Among them, Mobile IP [15] and Session Initiation Protocol (SIP) [11] have been standardized by Internet Engineering Task Force (IETF) [23] as the mobility solutions for the network layer and application layer, respectively. Although Mobile IP seems to be the architecturally right protocol for providing IP Mobility in the wireless Internet, it requires significant changes in the underlying networking infrastructure. Application layer protocols, however, are transparent to the lower layer characteristics. They maintain the true end-to-end semantics of a connection and are expected to be the right candidate for handling mobility in a heterogeneous environment. Indeed, SIP has been accepted by the third Generation Partnership Project (3GPP) as a signaling protocol for setting up real-time multimedia sessions. SIP is capable of supporting not only terminal mobility but also session mobility, personal mobility and service mobility. Therefore, SIP seems to be an attractive candidate as an application layer mobility management protocol for heterogeneous 4G wireless networks. However, SIP uses TCP or UDP to carry its signaling messages and hence is limited by the performance of TCP or UDP over wireless links. In addition, SIP entails application layer processing of the messages, which may introduce considerable delay. These are the prime factors behind the handoff delay while using SIP as the mobility management protocol.

European Telecommunications Standards Institute (ETSI) [2] has defined in a quantitative way four different classes of performance—*best, high, medium, and best effort*—for voice traffic and streaming media over IP networks [4]. The first two classes specify the type of IP telephony services that have the potential to provide a user experience better than the Public Switched Telephone Network (PSTN). Medium class has the potential to provide a user experience similar to common wireless mobile telephony services. Best effort class includes the type of services that will provide a usable communications service but may not provide performance guarantees. The specification for the end-to-end media packet delay for the best and high classes of services is less than 100 ms, while for medium and best effort classes the delay is less than 150 and 400 ms, respectively. In fact, a handoff delay of more than 200–250 ms makes voice conversations annoying. Clearly, the handoff delay, being a component of the total end-to-end delay, should also abide by these delay limits. Thus, it is evident that for quality of service (QoS) sensitive streaming multimedia traffic belonging to either *best* or *medium* class, the handoff delay should be less than 100 ms.

In this paper we investigate the performance of SIP as a mobility management protocol in a heterogeneous access networking environment predicted for 4G wireless networks. In particular, we perform a case study of SIP-based handoff delay analysis using SIP to handle terminal mobility in a IP-based network. Two different types of access technologies, viz. UMTS and IEEE 802.11b based WLAN, have been considered for the IP-based network. Analytical results show that for WLAN networks the handoff delay is suitable for streaming media but for UMTS network the minimum handoff delay does not meet the specifications. More precisely, handoff to a UMTS network from either another UMTS network or a WLAN, introduces a minimum delay of 1.4048 s for 128 kbps channel, while a handoff to a WLAN access network from another WLAN or a UMTS network, the minimum delay is 0.2 ms. Clearly, in the former case the minimum delay is unacceptable for streaming multimedia traffic and requires the deployment of soft-handoff techniques to reduce the handoff delay and keep it within a desirable maximum limit of 100 ms.

The rest of the paper is organized as follows. In Section 2, we describe the system architecture for the case study. In Section 3, we model and analyze the handoff performance of SIP for the sample architecture and present numeric results to evaluate the performance of SIP-based terminal mobility management. Section 4 concludes the paper.

2. System architecture

Telecommunication networks are gradually shifting from circuit switched to packet switched networks. At the same time the applications are converging to multimedia based applications. For our case study, we have considered an architecture conceptually similar to IP-based 4G networks in terms of heterogeneity in access network technologies. A logical view of the architecture considered is presented in Fig. 2. The architecture is primarily focused on wireless

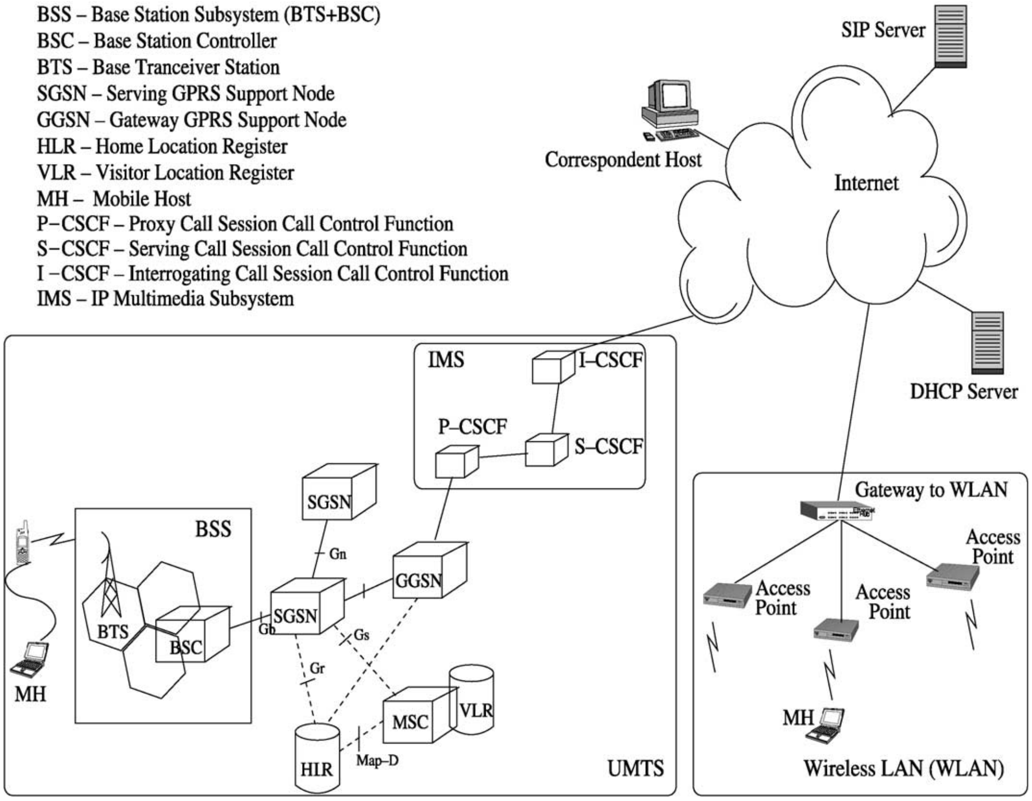
BSS – Base Station Subsystem (BTS+BSC)
BSC – Base Station Controller
BTS – Base Tranceiver Station
SGSN – Serving GPRS Support Node
GGSN – Gateway GPRS Support Node
HLR – Home Location Register
VLR – Visitor Location Register
MH – Mobile Host
P–CSCF – Proxy Call Session Call Control Function
S–CSCF – Serving Call Session Call Control Function
I –CSCF – Interrogating Call Session Call Control Function
IMS – IP Multimedia Subsystem
SIP Server
Correspondent Host
Internet
DHCP Server
IMS
I–CSCF
P–CSCF
S–CSCF
SGSN
Gn
BSS
GGSN
SGSN
Gb
Gs
Gr
BTS
BSC
MH
MSC
VLR
HLR
Map–D
UMTS
Gateway to WLAN
Access Point
Access Point
Access Point
MH
Wireless LAN (WLAN)

Fig. 2. 4G Architecture considered for case study.

mobile multimedia networking and is constructed around an IP core network (the Internet) with two different types of the access networks viz. UMTS and WLAN. The UMTS Release 5 multimedia architecture [1] has been proposed by 3GPP to provide multimedia based services in an all-IP environment. However, complete migration to UMTS networks may not be possible in recent future and a heterogeneous environment could evolve with several of the existing access technology like IEEE 802.11 based WLAN, operating with emerging core networks. This observation forms the basis of our selection criteria for the architecture to be studied in this paper.

UMTS Release 5 defines GPRS/EDGE[1] radio access network (GERAN) as its access technology. We have assumed only GPRS access network due to its wide acceptance. GPRS networks are built on existing GSM (Global System for Mobile Communications) [3] networks by adding a new class of network nodes called the GPRS support nodes (GSN). A *serving GPRS support node* (SGSN) is responsible for mobility and link management, and delivering packet to the MH under its service area. A *gateway GPRS support node* (GGSN) acts as an interface between the GPRS network and the external packet data networks (the Internet in this case). Home Location Register (HLR) and Visited Location Register (VLR) are two databases to keep user location information for mobility management. These databases are derived from legacy GSM architecture. A location register in the SGSN keeps track of the current VLR for a user.

A salient feature of UMTS Release 5 standardization is the new subsystem, known as the IP Multimedia Subsystem (IMS) that works in conjunction with the Packet Switched Core Network (PS-CN) for supporting legacy telephony service as well as new multimedia services. The IMS enables an IP-based network to support both IP telephony services as well as the multimedia services. SIP is the signaling protocol used between the MH or User Equipments (UE) and the IMS as well as with its internal components. As far as the SIP signaling is concerned, the main component of the IMS involved is the Call Session Control Function (CSCF), which is basically a SIP server. The CSCF performs a number of functions such as multimedia session control and address translation function (i.e. evolution of digit translation function). In addition, the CSCF must perform switching function for services, voice coder negotiation for audio communication, and handling the subscriber profile (analogous to the Visitor Location Register). The CSCF play three roles, viz. the Proxy CSCF (P-CSCF) role, the Interrogating CSCF (I-CSCF) role and the Serving CSCF (S-CSCF) role. P-CSCF is the mobile's first point of contact with the IMS network; I-CSCF is responsible for selecting the appropriate S-CSCF based on load or capability; S-CSCF is responsible for mobile's session management.

[1] Enhanced data rates for GSM evolution.

The other access network technology considered is IEEE 802.11 based WLAN. A WLAN access network consists of several access points (AP) providing the radio access to the MH. The APs are connected to the backbone IP network with an ethernet switch. A DHCP (Dynamic Host Configuration Protocol) [9] server is used to assign an IP address to a visiting MH.

We assume that an MH moving between UMTS network and WLAN has separate network interfaces to connect to these networks. The MH after moving to a UMTS network or a WLAN switches to the respective interface in order to attach to the corresponding access network infrastructures. The switch over instant is identified by the reception of the GPRS pilot signal in a UMTS network and the characteristics beacon in a WLAN.

2.1. In-session or mid-call handoff with SIP

SIP is a simple scalable, text-based protocol that offers a number of benefits, including extensibility and the provision for call/session control. The main entities in SIP are user agents, proxy servers and redirect servers. A user is generally identified using an email like address, such as *user@userdomain*, where *user* is the user name and *userdomain* is the domain or numerical address. There exist various methods defined in SIP, viz. INVITE, ACK, BYE, OPTIONS, CANCEL, and REGISTER. Apart from the signaling function SIP inherently supports personal mobility and can be extended to support service and terminal mobility [20,26].

Terminal mobility requires SIP to establish connection either during the start of a new session, when the terminal or the MH has already moved to a different location, or during the middle of a session. The former situation is referred to as *pre-call mobility* while the latter is known as *mid-call mobility*. For pre-call mobility the MH re-registers its new IP address with its 'home' by sending a REGISTER message, while for mid-call mobility, the terminal needs to intimate the correspondent host (CH) or the host communicating with the MH, by sending an INVITE message about the terminal's new IP address and updated session description. In principle, this is similar to Mobile route optimization [16]. The CH starts sending data to the new location as soon as it gets the re-INVITE message. Hence, the handoff delay is essentially the one-way delay for sending an INVITE message from the MH to the CH. Here the home refers to the redirect or SIP server in the home network of the MH. The MH needs to register with the redirect server in the home network for future calls. High level messaging of SIP-based mid-call mobility management is depicted in Fig. 3. However, in mid-call mobility management, before sending the SIP re-INVITE message there are some procedures that need to be completed to get the MH attached to the wireless access network infrastructure. For example, an MH attaches to the GPRS radio access of a UMTS network using the GPRS Attach and Packet Data Protocol (PDP) Context Activation procedure, while for the WLAN it uses DHCP to attach to the WLAN.

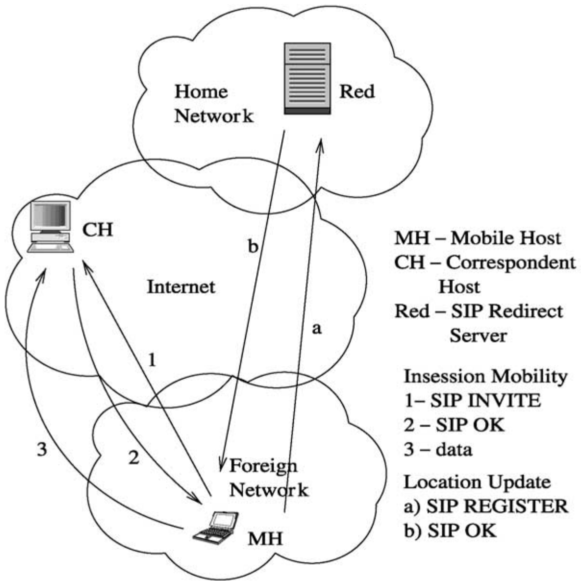

Fig. 3. SIP-based mid-call terminal mobility management.

Now, mobility in such a heterogeneous networking environment can give rise to the following four cases: (i) MH moves from a UMTS network to another UMTS network, (ii) MH moves from a UMTS network to a WLAN network, (iii) MH moves from a WLAN network to a UMTS network, and (iv) MH moves from a WLAN network to another WLAN network. Since our concern here is to analyze the delay incurred in the handoff procedure, the above four cases can be mapped to only two cases of interest:

- MH moving to a UMTS network
- MH moving to a WLAN.

This is because the handoff delay is caused mainly by the message exchange that occurs while an MH attaches to a new access network (either UMTS or WLAN in our case) followed by the location update. These two cases are discussed in more details as follows.

(1) *MH moving to a UMTS network from another UMTS network or a WLAN*: When an MH moves to a UMTS network, it performs two key functions to initiate a handoff.

• Data Connection Setup that involves the execution of two procedures known as GPRS Attach and the PDP Context Activation. This establishes the data path required to carry the SIP related messages to the P-CSCF through the GGSN, which acts as the gateway for the P-CSCF. The messages involved in the GPRS Attach and the PDP Context Activation procedures are shown in Figs. 4 and 5. The steps are described as follows. As a part of the GPRS Attach procedure, the MH sends an Attach message (1) to the SGSN (responsible for mobility management, logical

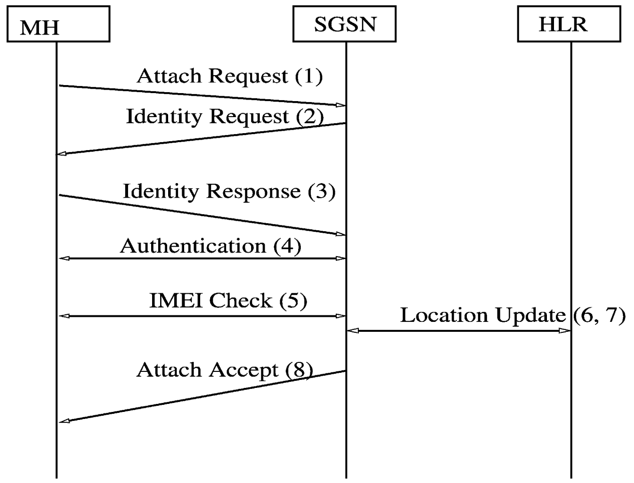

Fig. 4. GPRS Attach procedure.

link management, and authentication and charging functions in a UMTS network) with the MHs International Subscriber Identifier (IMSI). The SGSN uses the IMSI to authenticate (messages 2, 3, 4 and 5) the MH with its HLR. Successful authentication is followed by the SGSN sending a location update to the HLR (messages 6 and 7). The SGSN finally completes the Attach procedure by sending an Attach Complete message (8) to the MH. Thus a logical association is established between the MH and the SGSN.

Once an MH is attached to an SGSN, it must activate a PDP address (or IP address) to begin packet data communication. Activation of PDP address creates an association between the MHs current SGSN and the GGSN (acting as the interface between the GPRS/UMTS backbone network and the external packet data networks) that anchors the PDP address. A record of such an association is known as the PDP context. The PDP context transfer is initiated by the MH by sending PDP Context Activation message (9) to the SGSN. The SGSN after receiving this Activation message discovers the appropriate GGSN (messages 10 and 11). It selects GGSN capable of performing functions required for the SIP related activities. The SGSN and the GGSN create special paths for the transfer of SIP messages to the P-CSCF, which is identified by the GGSN. The corresponding IP address of the P-CSCF is sent along with the activation accept message (messages 12–16).

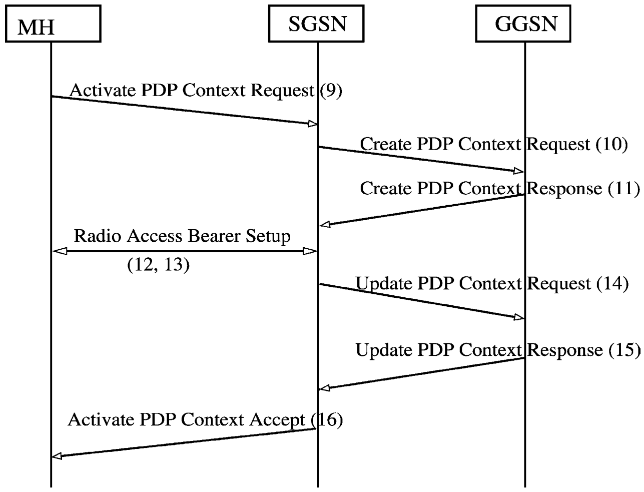
MH
SGSN
GGSN
Activate PDP Context Request (9)
Create PDP Context Request (10)
Create PDP Context Response (11)
Radio Access Bearer Setup
(12, 13)
Update PDP Context Request (14)
Update PDP Context Response (15)
Activate PDP Context Accept (16)

Fig. 5. PDP Context Activation procedure.

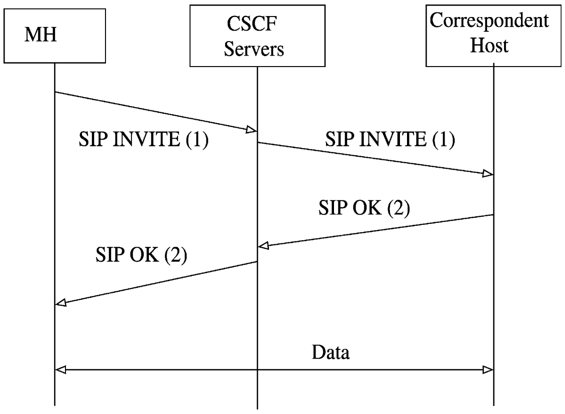

Fig. 6. Messages involved in SIP-based mid-call terminal mobility management.

- The SIP message exchange for re-establishing the connection is shown in Fig. 6. The MH re-invites the CH to its new temporary address by sending SIP INVITE message (1) through the P-CSCF, S-CSCF and the I-CSCF servers. The INVITE message uses the same call identifier as in the original call setup and contains the new IP address at the new location. Once the CH gets the updated information about the MH, it sends an acknowledge message (2) while starting to send data.

(2) *MH moving to a WLAN from another WLAN or a UMTS network*: When an MH moves to a WLAN it goes through the following major steps to update its location with the CH.

- The MH goes through DHCP registration procedure to secure a new IP address for its new location. The message exchanged in the registration procedure is shown in Fig. 7. When the MH identifies the presence of WLAN after receiving the characteristics beacons, it broadcasts DHCP DISCOVER message (1) to discover the DHCP server willing to lend it with registration service. The appropriate DHCP server sends out DHCP OFFER message (2) to offer service to the requesting MH. The MH on receiving this OFFER message sends a DHCP REQUEST message (3) to the DHCP server to confirm the offer made. The DHCP server then sends the MH an DHCP ACK message (4) with information such as the new IP address to be assigned to the MH.
- The SIP message exchange to re-establish the connection is similar to that for UMTS networks, where the MH, after acquiring the new IP address, re-invites the CH to its new address by sending SIP INVITE message (messages 5, 6, and 7).

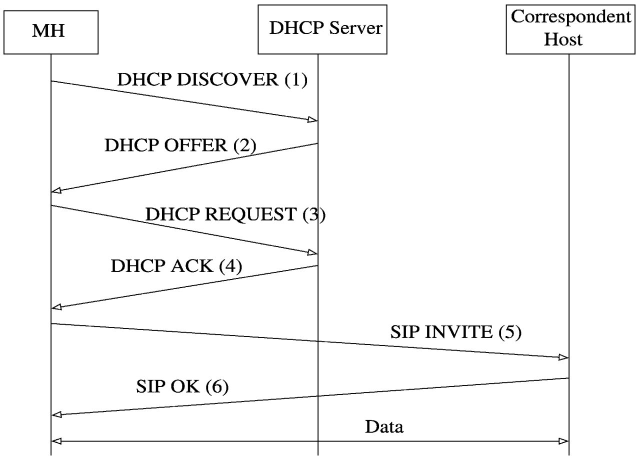

Fig. 7. DHCP registration procedure.

3. Handoff delay analysis

In this section we derive the handoff delay introduced due to the wireless link in the UMTS and WLAN access networks and the queuing delay in the different servers in the signaling path. We show that the handoff delay due to the wireless access is congenial to streaming media for WLAN and not for UMTS networks. Let us denote the handoff delay during the mid-call terminal mobility by D_{Handoff}, which can be divided into two parts: (i) the delay occurred during the attachment procedure and (ii) the delay due to location update using the SIP INVITE message. During each of these procedures, messages are transported over the wireless access link, which introduces major delays in comparison with the queuing and transmission delay introduced by the high speed backbone networks.

To compute the delay for transmitting messages over the wireless links in the access networks, we have used the delay models for frame and packet transmission over a wireless link under various link error conditions, proposed in Ref. [7]. The outdoor operation of GPRS radio access networks makes it more vulnerable to noise, thus increasing the bit-error rate (BER) for the wireless channel. To improve the BER performance for transmission of packets over wireless links, a semi-reliable link-layer retransmission mechanism like the Radio Link Protocol (RLP) is used on top of the MAC (Medium Access Control) layer. However, due to much higher bandwidth and the indoor operation of WLAN, no such retransmission scheme is used. So we need to consider two types of wireless delay models for our case study.

(a) *Transmission delay with RLP (for GPRS radio access)*: The analysis considers the following parameters.

- p, probability of an RLP frame being in error in the air link;
- k, number of frames in a packet transmitted over the air;
- D, end-to-end frame propagation delay over the air link (typical values of the order of 100 ms).
- τ, interframe time of RLP (typical values of the order of 20 ms for GPRS).

The effective packet loss P_f seen at the transport layer, with RLP operating underneath, is given as:

$$P_f = 1 - p + \sum_{j=1}^{n}\sum_{i=1}^{j} P(C_{ij}) = 1 - p(p(2-p))^{n(n+1)/2} \qquad (1)$$

where n is the maximum number of RLP retransmission trials and C_{ij} (representing the first frame received correctly at the destination) is the ith retransmission frame at the jth retransmission trial. For $n = 3$ (typical value), the packet loss rate with RLP for packets with k frames is given as:

$$q = 1 - (1 - p(p(2-p))^6)^k$$

Considering the RLP retransmissions, the transport delay in transmitting a packet over the RLP is given by:

$$D' = D + (k-1)\tau + \frac{k(P_f - (1-p))}{P_f^2} \times \left(\sum_{j=1}^{n}\sum_{i=1}^{j} P(C_{ij})\left(2jD + \left(\frac{j(j+1)}{2} + i\right)\tau\right)\right) \qquad (2)$$

Interested readers can find further details on the derivations of these expressions in Ref. [7].

Next we determine the value of k corresponding to different types of messages involved in the GPRS Attach and PDP Context Activation procedures. The maximum size of the messages exchanged in the GPRS Attach procedure is 43 bytes [3]. Now for a 9.6 kbps channel there are $9.6 \times 10^3 \times 20 \times 10^{-3} \times \frac{1}{8} = 24$ bytes in each frame. Therefore, the number of frames (k) to be transferred for a single message in the Attach procedure is $43/24 \approx 2$. Similarly, for a 19.2 kbps or higher bandwidth (e.g. 128 kbps) channel, the number of frames per message is $k = 1$. On the other hand, the maximum size for the PDP Context Activation messages is 537 bytes [3]. So the number of frames for PDP messages are $k = 537/24 \approx 23$, 12 and 2 for 9.6, 19.2 and 128 kbps channels, respectively.

Using the expression for delay model in Eq. (2), the delays corresponding to the GPRS Attach and the PDP Context Activation procedures can be determined as: $D_{\text{Attach}} = 8D'$ and $D_{\text{PDP}} = 4D'$. This is because, as shown in Figs. 4 and 5, GPRS Attach and PDP Context Activation procedures requires 8 and 4 message exchanges, respectively.

(b) *Transport Delay Without RLP (for WLAN)*: In this case there is no RLP retransmission. Instead, due to packet loss, retransmission may be done by upper layer protocols like TCP or DHCP until there is a successful transmission. Let N_m be the number of such retransmissions. The DHCP packet loss rate in this case is $q = 1 - (1-p)^k$, where p is the probability that a frame is in error and k is the number of frames in a packet transmitted. The average delay for

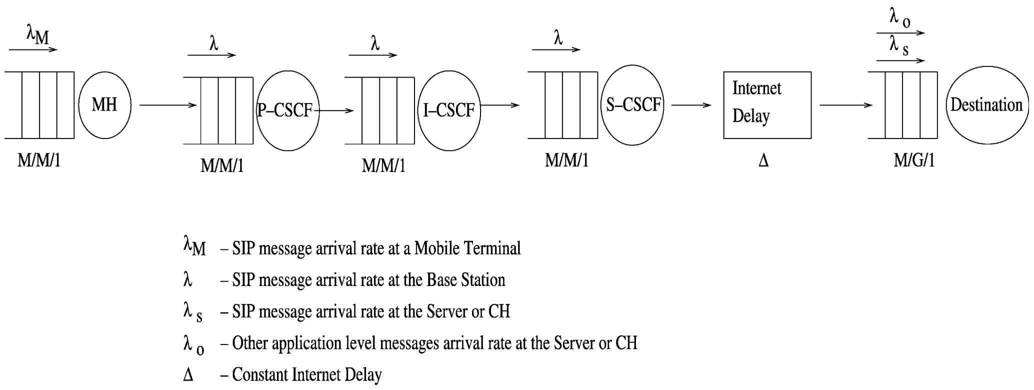
λM
MH
M/M/1
λ
P–CSCF
M/M/1
λ
I–CSCF
M/M/1
λ
S–CSCF
M/M/1
Internet Delay
Δ
λ o
λ s
Destination
M/G/1
λM – SIP message arrival rate at a Mobile Terminal
λ – SIP message arrival rate at the Base Station
λ s – SIP message arrival rate at the Server or CH
λ o – Other application level messages arrival rate at the Server or CH
Δ – Constant Internet Delay

Fig. 8. Queuing model for analyzing delay in SIP-based session setup for UMTS.

successfully transmitting a TCP or DHCP packet with no more than N_m retransmissions is given as,

$$D'' = (k-1)\tau + \frac{D}{(1-q^{N_m})(1-2q)} + \frac{1-q}{1-q^{N_m}} \times D\left[\frac{q^{N_m}}{1-q} - \frac{2^{N_m+1}q^{N_m}}{1-2q}\right] \tag{3}$$

Note that the DHCP messages have a maximum length of 548 bytes. Also the IEEE 802.11 standard specifies that the WLAN frame duration is 3.5 ms. So, using similar calculations as for the case with RLP, we get $k = 1$ for both 2 and 11 Mbps WLAN. Also the end-to-end transmission delay, D, of the wireless channel = 0.27 and 0.049 ms for 2 and 11 Mbps channel, respectively. The interframe time, $\tau = 0.001$ s, is independent of the bit rate. Now, the delay due to DHCP registration is given by $D_{\text{DHCP}} = 4D''$, since it is shown in Fig. 7 that DHCP registration requires 4 message exchanges.

3.1. Delay for handoff to UMTS network from another UMTS network or a WLAN

Since SIP is an application layer protocol, the processing of SIP messages in the intermediate and destination servers may take considerable time due to the queuing of messages that need to be accounted for. Rough estimates of the queuing delays can be obtained using the classical queuing theory based waiting time formulas.

The major delays occur in the MH, the P-CSCF, I-CSCF, S-CSCF and the destination server due to the queuing of the SIP messages. This is shown in Fig. 8.

To compute the queuing delay we have assumed an M/M/1 queuing model for the MH as well as the CSCF servers, and a priority based M/G/1 model for the destination server. The rationale behind these assumptions is that while the MH and the CSCF servers perform dedicated jobs, the destination server may be busy with a variety of jobs other than serving the SIP messages and thus may have general service time distribution. Table 1 lists the parameters used in the analysis and their meanings. Although, it has been shown that Internet delay varies between 100 ms and 1 s [14], emerging high-speed technologies like Generalized Multiprotocol Label Switching (GMPLS) [6] provide efficient traffic engineering to reduce this Internet delay to a nominal fraction of the minimum allowed end-to-end delay (in the order of few ms). Hence, the major concern is with the delay introduced by the wireless links in the access networks.

Table 1
List of system parameters

Parameters	Symbols
λ_M	SIP message arrival rate at the UE/MH
λ	SIP message arrival rate at the CSCF servers
μ	Processing rate for each SIP message in the UE/MH
ρ_s	Destination and the CSCF server load
λ_s	SIP message arrival rate at the destination
μ_s	Processing rate for each SIP message at the destination
ρ_o	Load at the destination for messages other than SIP
λ_o	Arrival rate at the destination for messages other than SIP
μ_o	Processing rate at the destination for messages other than SIP
Δ_I	Internet delay in transmitting of SIP messages
D_{MH}	Queuing rate at the MH
D_{RLP}	Delay in transmitting a packet over an RLP link in UMTS network
$D_{\text{P-CSCF}}$	Queuing delay at the P-CSCF server
$D_{\text{I-CSCF}}$	Queuing delay at the I-CSCF server
$D_{\text{S-CSCF}}$	Queuing delay at the S-CSCF server
D_{Dest}	Queuing delay at the destination (CH)
D_{SIP}	Queuing delay at the P-CSCF server
D_{GW}	Queuing delay at the gateway to the WLAN
D'	Transport delay with RLP
D''	Transport delay without RLP

Let us now determine the queuing delay of a SIP message at the MH, the intermediate CSCF servers, the destination and the transport delay over the wireless access. We assume that multiple MHs are served by the CSCF servers, although there are some load balancing functions in the CSCF servers. Hence $\lambda_M \leq \lambda$, or λ_M is a fraction of λ. The SIP message transmission delay, $D_{\text{SIP-UMTS}}$, for GPRS radio access of a UMTS network can be computed as

$$D_{\text{SIP-UMTS}} = D_{\text{MH}} + D_{\text{RLP}} + D_{\text{P-CSCF}} + D_{\text{I-CSCF}} + D_{\text{S-CSCF}} + \Delta_I + D_{\text{Dest}} \tag{4}$$

Using the results from queuing theory [12] and the parameters presented in Table 1, the delay components are estimated as follows.

$$D_{\text{MH}} = \frac{1}{\mu - \lambda_M} \tag{5}$$

$$D_{\text{P-CSCF}} = D_{\text{I-CSCF}} = D_{\text{S-CSCF}} = \frac{\rho_s}{\lambda(1-\rho_s)} \tag{6}$$

$$D_{\text{Dest}} = \frac{\frac{1}{\mu_s}(1-\rho_o-\rho_s) + R}{(1-\rho_o) + (1-\rho_o-\rho_s)} \tag{7}$$

where $R = \lambda_o \bar{X}_1^2 + \lambda_s \bar{X}_s^2/2$; $\bar{X}_1^2$ and $\bar{X}_s^2$ are the second moments of μ_o and μ_s, respectively. The expression for D_{Dest} is obtained by using the result of a non-preemptive priority-based M/G/1 queue [12]. Since our objective is to estimate the SIP message processing delay, we have considered only those messages having higher priority than SIP messages and ignored other lower priority messages. The derivation of the delay, D_{RLP}, requires us to adopt a transport layer based delay model over wireless links. Now SIP messages work with both TCP and UDP. Since we are dealing with wireless links and TCP is a reliable protocol, the SIP messages are assumed to be sent over TCP. So a delay model for TCP transmission over wireless links is required.

According to the model used and the results reported in Ref. [7], the delay to transmit a TCP segment consisting of k frames over a radio link with RLP operating over it, is given by

$$\begin{aligned} D_{\text{RLP}} = {} & D(k-1)\tau + \frac{k(P_f-(1-p))}{P_f^2} \\ & \times \left(\sum_{j=1}^{n} \sum_{i=1}^{j} P(C_{ij}) \left(2jD + \left(\frac{j(j+1)}{2} + i \right) \tau \right) \right) \\ & + \frac{2Dq(1-q)}{1-q^{N_m}} \\ & \times \left[1 + \frac{4q(1-(2q)^{N_m-2})}{1-2q} - \frac{q(1-q^{N_m-2}}{1-q} \right] \end{aligned} \tag{8}$$

where $n = 3$ is the maximum number of RLP retransmission trials, N_m is the number of TCP retransmissions, τ, p, P_f, and C_{ij} are the same parameters as defined earlier, and $q = 1-(1-p)^k$ is the packet loss rate.

To derive the value of k (number of air link frames), we have assumed that a TCP segment is carried in one packet. We assume that the air link frame duration is 20 ms. As derived earlier, a 9.6 kbps radio channel can afford 24 bytes in each frame. Also, we assume that the size of one SIP message is 500 bytes. Therefore, the number of air link frames in a SIP message is $k = 500/24 \approx 21$. For 19.2 and 128 kbps channels the number of frames are $k = 11$ and $k = 2$, respectively.

3.2. Delay for handoff to a WLAN network from another WLAN or a UMTS network

For the WLAN network, the queuing delays are shown in Fig. 9. Different parameters used are also listed in Table 1. The corresponding transmission delay for a SIP message, $D_{\text{SIP-WLAN}}$, can be calculated in the same manner as $D_{\text{SIP-UMTS}}$ and is given as follows.

$$D_{\text{SIP-WLAN}} = D_{\text{MH}} + D'' + D_{\text{GW}} + \Delta_I + D_{\text{Dest}}$$

Here D_{GW} is essentially the same as the queuing delay at any of the CSCF servers and is given as follows.

$$D_{\text{GW}} = \frac{\rho_s}{\lambda(1-\rho_s)}$$

For both 2 and 11 Mbps WLAN networks, the number of frames corresponding to a SIP message (500 bytes) is $k = 1$. Although, typically in a WLAN, the SIP-based control messages and the data use the same channel, we have assumed that the control messages would have higher preemptive priority than the data frames and would not wait for pending data frame transmission.

Once we have the estimate for all the components, we can determine the total handoff delay for UMTS networks

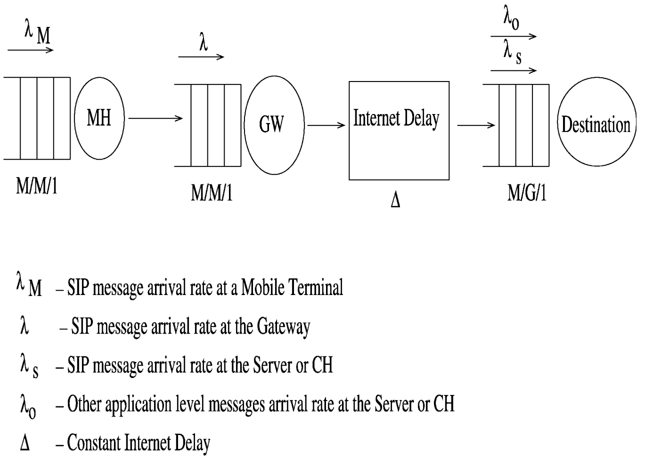
λM
λ
λ0
λs
MH
GW
Internet Delay
Destination
M/M/1
M/M/1
Δ
M/G/1
λM – SIP message arrival rate at a Mobile Terminal
λ – SIP message arrival rate at the Gateway
λs – SIP message arrival rate at the Server or CH
λ0 – Other application level messages arrival rate at the Server or CH
Δ – Constant Internet Delay

Fig. 9. Queuing model for analyzing delay in SIP-based session setup for WLAN.

Table 2
System parameter values

Parameters	Values
μ	4×10^{-4} s
ρ_s	$\lambda/\mu(\lambda < \mu)$
ρ_o	0.7
Δ_I	200 ms
N_m	10

as $D_{\text{Handoff}} = D_{\text{Attach}} + D_{\text{PDP}} + D_{\text{SIP-UMTS}}$ and for WLAN as $D_{\text{Handoff}} = D_{\text{DHCP}} + D_{\text{SIP-WLAN}}$.

3.3. Numerical results

In this section we present the results for the handoff delay computation in a SIP-based multimedia session using the delay models described in Section 3.2. The values used for different system parameters are given in Table 2.

We have assumed the SIP message arrival rate (λ) and the processing rate at the CSCF servers and the destination server, are the same (i.e. $\mu_s = \mu$). Also we have assumed $\lambda_M = 0.1\lambda$. The derivation of D_{Dest} involves the second moment of the processing rate at the destination, which can be derived once the mean and variance are given. For our analysis, we have assumed the standard deviation of the processing rates at the destination is 5% of the mean. Now $\bar{X}_1^2 = E[X_1^2]$ and $\bar{X}_s^2 = E[X_s^2]$. Also $E[X_1^2] = \sigma_1^2 + (E[X_1])^2$ and $E[X_s^2] = \sigma_s^2 + (E[X_s])^2$, where σ_1^2 and σ_s^2 are the respective variances. Substituting μ_o and $mu;_s$ for $E[X_1]$ and $E[X_s]$ and the values for the variances, we get $R = 0.501[\rho_o^2 + \rho_s^2]$.

As mentioned before, due to the varying nature of the Internet delay and the computing power of the intermediate servers, it is difficult to characterize the end-to-end handoff delay. With proper traffic engineering (e.g. GMPLS), the Internet delay can be made to suit the application requirements. Hence we focus on the component of the handoff delay introduced due to the wireless access networks to get an estimate of the minimum handoff delay. Subsequently we have also estimated the end-to-end handoff delay assuming a constant value for the Internet delay and some representative values for the computing capabilities of the servers as shown in Table 2. Fig. 10 shows the increase of the handoff delay component due to the wireless access only, with the increase of channel FER for channel bandwidth of 9.6, 19.2 and 128 kbps, when the MH moves to a UMTS network. Table 3 shows the corresponding end-to-end handoff delay including the queuing delay at different servers and the transmission delay over the Internet.

Fig. 11 shows the handoff delay component due to wireless access with the increase of SIP-based session request rate. The request rate of the SIP-based session in an MH is assumed to be $\lambda_M = 50$ requests/s when the channel FER is varied. On the other hand, the channel FER is kept constant at 0.05 when the arrival rate (λ_M) for SIP-based session is varied.

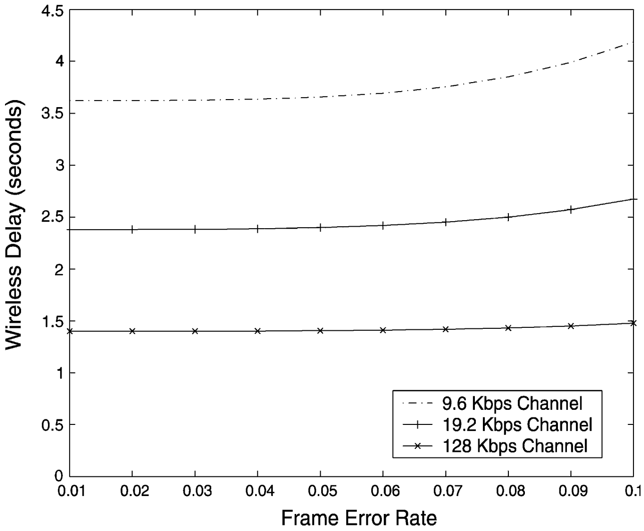

Fig. 10. Handoff delay vs. the channel FER—MH moving to UMTS network from another UMTS network or a WLAN.

The corresponding variation of handoff delay with the channel FER and SIP session request rate, for the case

Table 3
Handoff delay components (MH moving to UMTS network from another UMTS network or a WLAN)

Channel bandwidth	Channel FER	Processing delay	Wireless delay
9.6 kbps	0.010000	0.214408	3.620081
	0.020000	0.214408	3.621036
	0.030000	0.214408	3.624850
	0.040000	0.214408	3.634785
	0.050000	0.214408	3.655453
	0.060000	0.214408	3.692894
	0.070000	0.214408	3.754647
	0.080000	0.214408	3.849831
	0.090000	0.214408	3.989220
	0.100000	0.214408	4.185324
19.2 kbps	0.010000	0.214408	2.380042
	0.020000	0.214408	2.380538
	0.030000	0.214408	2.382520
	0.040000	0.214408	2.387681
	0.050000	0.214408	2.398418
	0.060000	0.214408	2.417867
	0.070000	0.214408	2.449945
	0.080000	0.214408	2.499389
	0.090000	0.214408	2.571796
	0.100000	0.214408	2.673663
128 kbps	0.010000	0.214408	1.400010
	0.020000	0.214408	1.400137
	0.030000	0.214408	1.400650
	0.040000	0.214408	1.401998
	0.050000	0.214408	1.404818
	0.060000	0.214408	1.409945
	0.070000	0.214408	1.418423
	0.080000	0.214408	1.431517
	0.090000	0.214408	1.450721
	0.100000	0.214408	1.477774

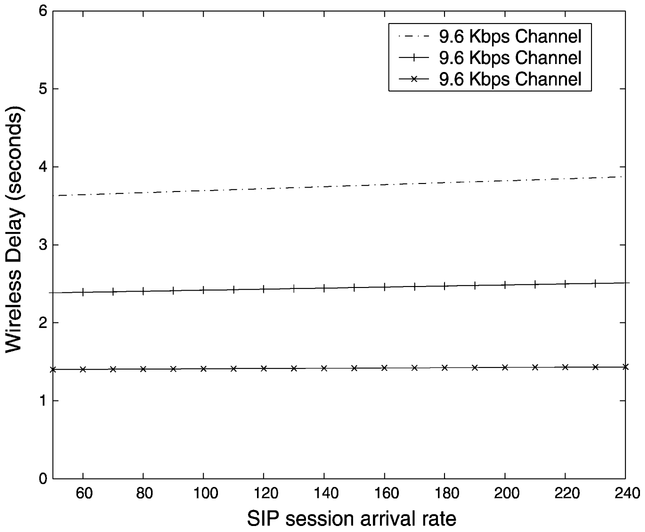

Fig. 11. Handoff delay vs. session request rate (λ_M)—MH moving to UMTS network from another UMTS network or a WLAN.

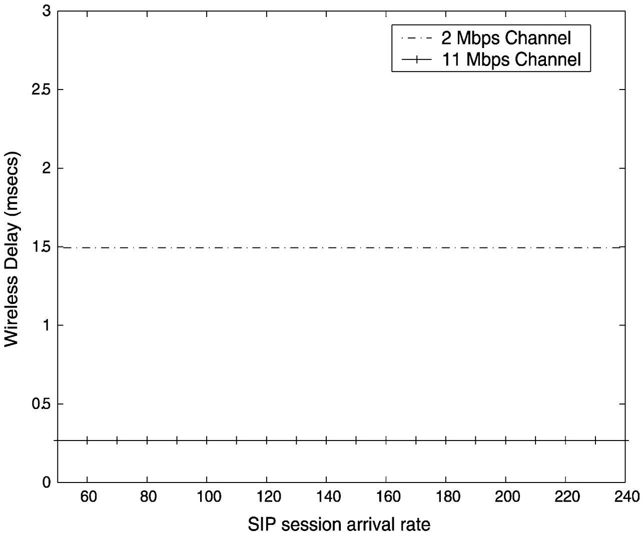

Fig. 13. Handoff delay vs. session request rate (λ_M)—MH moving to WLAN from another WLAN or a UMTS network.

when the MH moves to WLAN are given in Figs. 12 and 13. Table 4 shows the end-to-end handoff delay for an MH moving to a WLAN access network.

Observe that the component of handoff delay due to wireless access for a 128 kbps GPRS radio access of a UMTS network is 1.404818 s, where the channel FER is 0.05 and the SIP-based multimedia session arrival rate is 50/s. Whereas for the 11 Mbps WLAN, the handoff delay is only 0.267 ms. As mentioned earlier, to ensure QoS for streaming multimedia the maximum handoff delay should be ideally less than 100 ms and not more than 200 ms. Clearly, this requirement cannot be satisfied for a UMTS network even with a channel data rate of 128 kbps. However, soft-handoff techniques, such as those using 'make before break connection', may be used to counter the delay in the wireless link and meet the handoff delay requirement for the multimedia services using SIP as a mobility management protocol. Something similar to this has been done to reduce this handoff delay in Ref. [13], using shadow registration concept. However with higher speed data access (of the order of Mbps) as promised in emerging 4G networks [24], this auxiliary mechanisms would become redundant and SIP would be able to meet the performance requirement. For WLAN access networks, this is not as much of a problem because the wireless component of the delay is only around 0.2 ms.

As shown in Table 4, excluding the constant Internet delay, the end-to-end handoff delay is only around 1.9 ms. This leaves

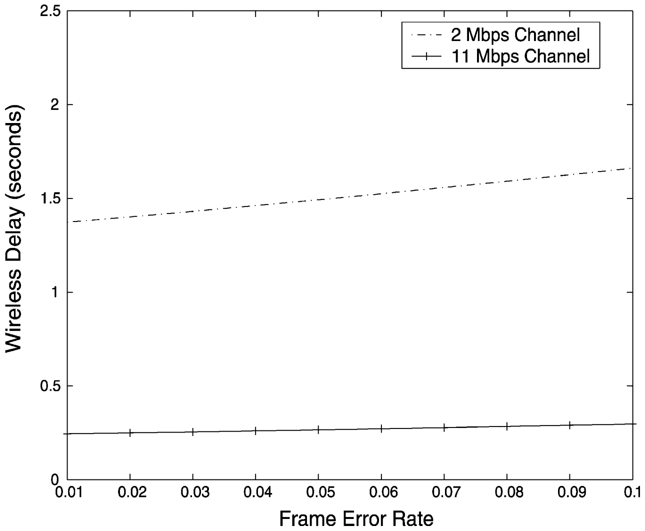

Fig. 12. Handoff delay vs. the channel FER—MH moving to WLAN from another WLAN or a UMTS network.

Table 4
Handoff delay components (MH moving to WLAN from another WLAN or a UMTS network)

Channel bandwidth	Channel FER	Processing delay	Wireless delay
2 Mbps	0.010000	0.201908	0.001373
	0.020000	0.201908	0.001402
	0.030000	0.201908	0.001431
	0.040000	0.201908	0.001462
	0.050000	0.201908	0.001493
	0.060000	0.201908	0.001525
	0.070000	0.201908	0.001558
	0.080000	0.201908	0.001592
	0.090000	0.201908	0.001626
	0.100000	0.201908	0.001661
11 Mbps	0.010000	0.201908	0.000246
	0.020000	0.201908	0.000251
	0.030000	0.201908	0.000256
	0.040000	0.201908	0.000262
	0.050000	0.201908	0.000267
	0.060000	0.201908	0.000273
	0.070000	0.201908	0.000279
	0.080000	0.201908	0.000285
	0.090000	0.201908	0.000291
	0.100000	0.201908	0.000297

a leverage of about 98 ms for the Internet delay. As mentioned before, with appropriate traffic engineering deployed in the Internet the end-to-end handoff delay can be restricted well within the stipulated maximum limit of 100 ms. The end-to-end handoff delay presented in Tables 3 and 4 are computed using the constant Internet delay of 200 ms.

4. Conclusions

Several mobility protocols have been proposed for Wireless Internet targeting different layers of the network protocol stack to achieve different goals [5]. Although each of them has the same goal of providing location transparency, the dependency of the mobility protocols on the underlying layers reduces as they operate higher in the protocol stack. In the next generation networks, a variety of wireless network technologies are supposed to co-exist, and hence no single network specific mobility protocol is expected to work for all of them. The design of a uniform mobility protocol that will work across all the different networks, requires tremendous efforts. The only solution seems to implement the mobility management functionality in the application layer, where there is least amount of dependency on the lower layers. SIP, accepted widely as a signaling protocol but capable of providing mobility support at the application layer, satisfies this criterion. We have performed a case study for a heterogeneous network with IP backbone having a combination of UMTS and WLAN access networks, to evaluate the performance of SIP-based mobility management. Results show that the minimum handoff delay introduced by the GPRS radio access of UMTS network is 1.4048 s for 128 kbps channel bandwidth, while the corresponding delay is around 0.2 ms for a 11 Mbps WLAN access network. Thus, the handoff delay while moving to a GPRS radio network, unlike the WLAN access network, is unacceptable for streaming multimedia. The major bottleneck found in our case study is the GPRS wireless access. Thus in order to comply with the maximum limit of the handoff delay, soft-handoff and advance resource reservation techniques need to be deployed.

References

[1] The Third Generation Partnership Project (3GPP), http://www.3gpp.org.

[2] The European Telecommunications Standards Institute, http://www.etsi.org.

[3] 3GPP TS 23.060, General Packet Radio Service, Description, Stage 2, December 2001.

[4] Telecommunications and Internet Protocol Harmonization over Networks, QoS Class Specification, TS 101329-2, http://www.3gpp.org.

[5] N. Banerjee, W. Wu, S.K. Das, S. Dawkins, J. Pathak, Mobility support in wireless Internet, IEEE Wireless Communications 10 (5) (2003) 51–61. October, 2003.

[6] L. Berger, Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions, RFC 3473 Internet Engineering Task Force, January, 2003.

[7] S.K. Das, E. Lee, K. Basu, S.K. Sen, Performance optimization of VoIP calls over wireless links using H.323 protocol, IEEE Transactions on Computers 52 (6) (2003) 742–752.

[8] S. Das, A. McCauley, A. Dutta, A. Misra, K. Chakraborty, S.K. Das, IDMP: an intra-domain mobility management protocol for next-generation wireless networks, IEEE Wireless Communications 9 (3) (2002) 38–45. June.

[9] R. Droms, Dynamic Host Configuration Protocol, RFC 2131 Internet Engineering Task Force, March, 1997.

[10] A. Grilo, P. Estrela, M. Nunes, Terminal independent mobility for IP (TIMIP), IEEE Communication Magazine (2001) 34–41. December.

[11] M. Handley, H. Schulzrinne, E. Schooler, J. Rosenberg, SIP: Session Initiation Protocol, RFC 2543, Internet Engineering Task Force, March, 1999.

[12] L. Kleinrock, QUEUING SYSTEMS vol. I: Theory, Wiley, New York, 1975.

[13] T.T. Kwon, M. Gerla, S.K. Das, S. Das, Mobility management for VoIP service: mobile IP vs. SIP, IEEE Wireless Communications 9 (5) (2002) 66–75. October.

[14] V. Paxson, End to End Internet Packet Dynamics, Proceedings of SIGCOMM, 1997, pp. 139–152.

[15] C.E. Perkins, I.P. Mobility, IP Mobility Support for IPv4, RFC 3220, 2002, January.

[16] C. Perkins, D. Johnson, Route Optimization in Mobile IP, draft-ietf-mobileip-optim-11.txt, September 2001.

[17] T.La. Porta, R. Ramjee, L. Lee, L. Salgerelli, S. Thuel, IP-based access network infrastructure for next-generation wireless data networks, IEEE Personal Communications 7 (4) (2000) 34–41. August.

[18] Y. Raivio, 4G—Hype or Reality, Proceedings of the IEEE 3G Mobile Communication Technologies (2001) 346–350. March.

[19] D. Raychaudhuri, 4G Network Architectures: WLAN Hot-Spots, Infostations and Beyond, International 4G Forum, London, UK, May, 2002.

[20] H. Schulzrinne, E. Wedlund, Application-layer mobility using SIP, ACM SIGMOBILE Mobile Computing and Communication Review 4(3), 47–57, 2000

[21] A.C. Snoeren, H. Balakrishnan, An end-to-end approach to host mobility, Proceedings MobiCom (2000) 155–166. August.

[22] The Focus project on 4G Mobile Network Architectures and Protocols, WINLAB, Rutgers University, http://www.winlab.rutgers.edu/pub/docs/focus/MobNet2.html.

[23] The Internet Engineering Task Force, http://www.itef.org.

[24] U. Varshney, R. Jain, Issues in emerging 4G wireless networks, IEEE Computer Magazine 34 (6) (2001) 94–96. June.

[25] C.-Y. Wan, A.T. Campbell, A.G. Valko, Design, implementation and evaluation of cellular IP, IEEE Personal Communications 7 (4) (2000) 42–49. August.

[26] E. Wedlund, H. Schulzrinne, Mobility Support using SIP, Second ACM/IEEE International Workshop on Wireless and Mobile Multimedia (WoWMoM'99), August, 1999, pp. 76–82.