Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Networks 50 (2006) 1860–1871
Computer Networks
www.elsevier.com/locate/comnet

A soft-handoff transport protocol for media flows in heterogeneous mobile networks

Antonios Argyriou *, Vijay Madisetti

School of Electrical and Computer Engineering, Georgia Institute of Technology, 777 Atlantic Drive, Atlanta, GA 30332, United States

Received 20 July 2004; received in revised form 21 July 2005; accepted 9 August 2005
Available online 12 September 2005

Responsible Editor: W. Wang

Abstract

In this paper we introduce a protocol for end-to-end handoff management in heterogeneous wireless IP-based networks. The protocol is based on the stream control transmission protocol (SCTP), and employs a soft-handoff mechanism that uses end-to-end semantics for signaling handoffs and for transmitting control messages. The objective of this protocol is first to reduce the home registration delay, and second, to eliminate the tunneling cost that exist in the current IP-based handoff management protocols. While the multihoming feature of SCTP has been suggested as way to realize soft-handoffs, our study is the first one the presents the relative merits of this handoff approach through an analytical methodology. After our theoretical analysis, we evaluate the performance the soft-handoff mobile-SCTP protocol, when media flows with stringent QoS requirements are employed. Our objective is to evaluate whether the soft-handoff mechanism employed by the protocol, can efficiently support media flows in terms of jitter and throughput. We present simulation results that show performance improvements for several vertical handoff scenarios in current and emerging mobile networks.

Keywords: Wireless LAN; SCTP; Mobile IP; Handoff management; Real-time flows

1. Introduction

A high quality of service (QoS) is possible for mobile communications over wireless networks, if it were possible to provide higher bandwidth and tighter control over path delay, jitter, and packet loss. Mobile IP, and its various derivatives [1,2], were designed in order to provide a scalable solution that facilitates handoff between IP-based subnets using encapsulation and tunneling. Its basic functionality is based on three features: Advertising a "care-of address", registering the current location of the mobile host, and tunneling. Agent advertisements, sent by specialized routers called foreign agents (FAs), are used to discover available links within a subnet. However, these mechanisms incur significant overheads like increased handoff delay, triangular routing, and IP encapsulation. All these

* Corresponding author. Tel.: +1 404 4838328.
E-mail addresses: anargyr@ece.gatech.edu (A. Argyriou), vkm@ece.gatech.edu (V. Madisetti).

1389-1286/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.comnet.2005.08.002

issues constitute important performance bottlenecks that must be resolved prior to the widespread acceptability of Mobile IP. Even though a number of optimizations have been reported [2,3], the fundamental overhead in tunneling is unavoidable, while new problems appear mainly related to security [4].

While Mobile IP targets mobility at a macro-level, a number of supplementary mobility management schemes have been developed in order to optimize for the case of frequent handoffs due to increased localized mobility of a host (micro-mobility). Hierarchical Mobile IP [2], Hawaii [5], Cellular IP [6], are some of the approaches that fall into this category. Hierarchical Mobile IP (HMIP) [2] is an extension to the base Mobile IP that allows the creation of a localized and scalable architecture. HMIP solves the problem of excessive signalling delay when the current visited network is far away from the home network. HMIP defines a way of localizing registrations to the visited network so that they do not have to traverse through the home network. This local architecture is realized through an entity called the gateway foreign agent (GFA). The mobile host uses regional registrations to the GFA in order to reduce the number of signaling messages to the home network. Moreover, for frequent intra-domain moves, the signaling delay is reduced since the related signaling messages have only localized significance. However, while reducing some signalling costs, these localized Mobile IP-based approaches share a common overhead—tunneling costs. Moreover, the implied dependence on specialized routing agents (FA/GFA) creates points of failure in the network. Mobile IP with route optimization [3], is used in order to route all the packets destined to a mobile host (MH) directly to its current location and not through its home agent (HA). The authors propose protocol extensions to Mobile IP, so that the correspondent host (CH) can cache the current binding of the MH, and then tunnels the packets for the MH directly to its care-of address, bypassing thus the MH's HA. The authors also propose a mechanism that allow packets in flight to a MH that moves, and packets sent based on an out-of-date cached binding, to be able to forwarded directly to the MH's new address binding. Another approach that shares some features with the proposed one in this paper, is homeless mobile IPv6 [7]. Homeless MIPv6 is basically intended for mobile hosts that do not need to or want to use any explicit home agent. The authors propose a number of modifications to mobile IPv6 so that it does not require, nor assume, any explicit home addresses or agents. One major drawback is that is based on IPv6, making it thus impractical for the current Internet. Moreover, it operates at the network layer and so TCP cannot handle the IP address changes. Finally, this protocol also requires the support of IPSec which incurs significant overhead and adds to the registration delay.

The first approach to handle mobility at the transport layer was TCP-Migrate [8], which is based in extensions of TCP. Location management is implemented with DNS [9] updates sent to the home network notifying it about the new IP address of the mobile host. In the meantime the TCP connection is suspended until the DNS update has taken place. If the delay is large (which is highly probable since the updates are sent to the home network), TCP might suffer from timeouts and so performance will be degraded. Another major handicap is the need for IPSec (which is penalized by a high overhead) every time the IP address is changing. Finally, additional performance degradation is attributed to the increased number of packets being lost during handoffs, because of TCP's inability to accept new socket parameters while in an existing session. Some recent works use the stream control transmission protocol (SCTP) [10] for mobility management [11–13]. These approaches simply state the observation that SCTP can handle changes in IP address, and exploit that for demonstrating a simple handoff. However, neither a detailed analysis of the benefits of such a solution nor a comparative study with current mobility management solutions are presented.

We argue that handoff management would be best handled by a protocol that operates over IP while still being able to inter-operate with network layer mechanisms (i.e. mobile IP). The protocol works at the transport layer and is implemented as extensions to the IETF's standard, the stream control transmission protocol (SCTP). The proposed handoff management protocol is general enough in that since it operates over generalized IP-based connectionless networks (Internet). This feature, will be even more important in the near future, since new incompatible wireless access technologies are emerging, while old ones are changing and evolved.

2. The handoff protocol

Primary design goal of the mobile-SCTP protocol is to operate under various wireless access technologies, to provide various levels of mobility services, and to support different types of QoS architectures. Additional requirements include limited modifications to the core infrastructure to a very minimum, and compatibility with Mobile IP and its various standardized derivatives (i.e. HMIP, MIP-RO). The choice a transport layer protocol appears as solution capable of meeting these challenging requirements.

The distinguishing feature of the protocol is that it completely works around TCP's inherent inability to handle socket parameter changes for ongoing sessions. This is achieved by allowing a transport layer session to maintain more than one active addresses for both the source and the destination. These addresses may be added or deleted dynamically for ongoing sessions. In this way a transport layer session is not defined by a specific IP address and port number pair of the source and the destination. This feature of the protocol removes two significant drawbacks of baseline Mobile IP: first, the overhead of IP encapsulation and tunneling and second, the dynamic addition/deletion of addresses eliminates the need for triangular routing, since a host experiencing handoff need only to notify initially only the CH. While the protocol is new, we will use terminology borrowed from Mobile IP for convenience. This facilitates comparison with specific functionality of MIP, HMIP, and MIP-RO in a straightforward manner. In this section we define in detail the operations performed by mobile-SCTP in order to handle handoffs in an end-to-end decentralized fashion. The description of specific operations of the protocol follow.

2.1. The stream control transmission protocol

The recently adopted IETF stream control transmission protocol (SCTP) is a reliable transport protocol that was initially designed for signaling transport. To its advantage, it proved to be a powerful, feature-rich transport layer protocol, that can operate on top of connectionless IP-based packet networks such as the Internet. One of the most important new ideas that SCTP is using is that of explicit support for multi-homed hosts. A single SCTP association (session) is able to use alternatively anyone of the available IP addresses without disrupting an ongoing session. However, this feature is currently used by SCTP only as a backup mechanism that provides fault tolerance in case of link failures. The state of each IP address (path) is maintained by sending heartbeat messages and allowing thus the detect a specific path failure and switch to another IP address. Another novel feature is that the protocol decouples reliable delivery from message ordering by introducing the idea of streams. A stream is an abstraction that allows applications to preserve in order delivery within a stream but unordered delivery across streams. This feature avoids head-of-line blocking at the receiver in case multiple independent data streams exist in the same SCTP session. Despite all these differences, congestion control was defined similar to TCP, primarily for achieving TCP friendliness [10].

In the next subsections, we describe in detail how handoffs can be handled by the mobile-SCTP protocol.

2.2. Joining the foreign domain

When a mobile host enters a new foreign network or subnet, it must obtain a valid IP address from the visited network (e.g. via DHCP [14]) which is called care-of address (CoA). Normally the responsibility of a DHCP server is to provide an IP address to the MH. If the DHCP server provides a universally routable IP address, then the only responsibility of the foreign network is to keep track of the usage of this address. The only overhead with our protocol, is related to the processing at the DHCP server since it maintains a list of the allocated addresses. However, things are different if the MH obtains a private address, and communicates behind a NAT server [15]. In that case the NAT server effectively hides host movements inside the visited network. Even if the MH changes IP address while moving from one subnet to another, the HA and CH will still be unaware of that.

2.3. Handoff algorithm

In this subsection we describe what steps are followed when there is a need to handoff. The handoff process is presented with the algorithm in Fig. 1. In this algorithm, we assumed that the MH obtains an IP address through the DHCP protocol [14]. However, the same procedure could be followed with IPv6 by using the auto-configuration feature that IPv6 supports [16].

As we said earlier, even with the mobile-SCTP protocol, there is a need for a HA entity in order to provide the location service. However, with the proposed protocol, the CH is notified directly by

1: Mobile host enters a subnet and sends a DHCPDISCOVER message.
2: The DHCP server sends a DHCPOFFER that contains the assigned IP address.
3: The MH uses this new IP address as its primary address for any new connection.
4: When the MH receives an advertisement from a new access point (AP), sends a new DHCPDISCOVER as before.
5: Upon reception of the new DHCPOFFER reply, it sends a DHCPACK to acknowledge the address assignment.
6: Binding update is send to the CH.
7: The CH uses the new address to reach the MH
8: The MH notifies the HA about the current location periodically.

Fig. 1. The handoff algorithm based on DHCP address configuration messages.

the MH when handoff is about to take place through a direct binding update from the MH to the CH. Actually, the binding update procedure is separated in two parts—a binding update that corresponds to the updates sent to the CH during handoffs, and a periodic binding (that also MIP requires) sent directly to the HA. The second part corresponds to a periodic update that has limited relationship, if any, to do with updates due to host movement. This means that there will be cases where a MH has moved to a new location, and has a new IP address, but binding update need not take place, even though mobile IP would require it. We allow a level of "staleness" in the current location of the mobile host so that signaling traffic is reduced. We will later see how we specify this periodic update and analyze the relative tradeoffs. However, when a lag can be accepted in the notification of the HA, we must make sure that binding updates sent to the CH are delivered as soon as possible. These updates may happen while there is an ongoing data transmission, and more specifically in our case can be done proactively (see Section 2.1).

2.4. Binding updates through dynamic address reconfiguration

Recently, IETF has proposed a simple extension to the base SCTP protocol that allows dynamic reconfiguration (addition/deletion) of IP addresses of an existing association between two SCTP endpoints [17]. This SCTP extension was used so that the MH can request from its peer to reset its primary destination address to a new value. The new address in this case will be the IP address that is received when the MH registers with the new visiting network. By doing this, the protocol is able to always maintain an active source or destination address which is topologically correct.

Fig. 2 presents in detail the sequence of messages exchanged during a binding update process. The MH initially goes through a handshaking procedure with the CH that involves four SCTP session setup messages [10]. After that, a connection is established, the data flow is initiated. At some point in the future, the MH may decide to move and at this moment mobile-SCTP gets into action. Upon entering the area of a new access point the MH will start receiving advertisements. The MH immediately sends a registration request and obtains a registration reply. Subsequently, the MH sends a packet with an ASCONF chunk (address configuration [17]) to the CH requesting from it to add the new address to the current association. Upon reception of an ASCONF-ACK packet, the MH starts using the new IP address as its source for all its packets. Clearly, the latency of the registration procedure and the handshake with the CH, represent the total handoff delay. However, a MH must be able to move very fast or the packet arrival rate must be really high so that increased number of packet losses take place. We will analyze this problem in detail in an upcoming subsection, where we will analytically derive the handoff induced packet loss rate.

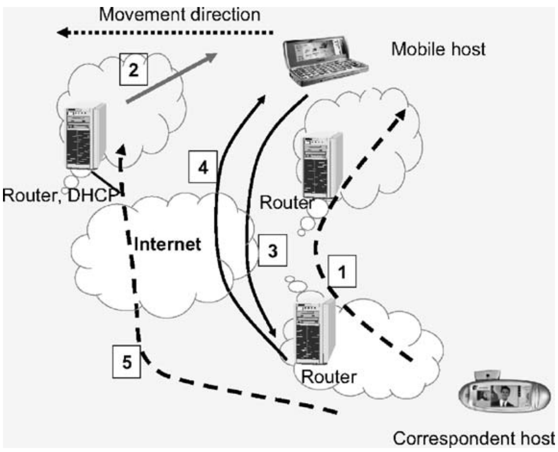

Fig. 2. Message flow of the end-to-end handoff management protocol.

In the communication scenario that we described till now in this paper, we are concerned with flows that are initiated from the MH. However, there are also cases when CH requires connection with the MH. Clearly, the only practical way that this can happen is with the help of the home network, which should be able to provide the current location of the MH. After the CH receives the current location of the MH it can initiate direct communication. This is implemented through a DNS function at the HA which provides the current IP for the CH. Another option would be the exchange of typical mobile IP messages between CH and HA in order to obtain initially the current address binding of the MH [1]. Homeless MIPv6 is a proposal that overcomes the need of a HA, but it is tightly couple with IPv6 routing. Since we are considering IPv4 systems in this paper, this solution cannot be applied. Note that inter-operation with Mobile IP can be smooth since we do not intervene with the inner working of the protocol.

3. Analytical protocol modeling

In this section we define an analytical model that describes the costs related to all operational phases (registration, tunneling, packet delivery etc.) of the mobile-SCTP and the MIP/HMIP protocols. We adopt notation from [18,19] so that comparison with mobile IP becomes easier. At the same time we develop analytical model for two promising mobile IP optimization mechanisms: Hierarchical MIP and MIP with route optimization. We believe that comparison of mobile-SCTP with these two solutions represents a fair comparison with state of the art mobility mechanisms. These protocols are currently considered the most prominent solutions for practical deployment of IP-based mobility services.

3.1. Model parameters and assumptions

All the parameters that are needed by the model can be seen in Table 1. As we go into more details of the protocol model into the next subsections, the meaning of each one of these parameters will become more clear. In these parameters, we can see that the model ignores the effect of queueing delays in the core network. Therefore the model will provide good estimates for the several costs related with the operational faces of the handoff protocol, when the core network is lightly loaded. This assumption is valid in today's Internet, since the bottlenecks appear usually at the access networks, while the core networks are usually underutilized [20].

Table 1
Parameters of the handoff protocol model

Parameter	Value
Packet service time at the FA (s_f)	10
Packet service time at the GFA (s_g)	15
Packet service time at the HA (s_h)	25
Distance cost (δ_D)	0.05
Weight factor (w_1)	0.3
Weight factor (w_2)	0.7
MH per subnet (n)	15
Wireless cost multiplier (ρ)	10
Poisson packet arrival rate (λ_a)	Variable
Mobility ratio (MR) $1/T_f$	Variable

We evaluated the performance of the proposed protocol, based on the assumption that the mobile host will move randomly between the subnets [21]. Given that a network consists of N subnets, and that each subnet consists of r regional networks in case of HMIP, the expectation of a host going out of a regional network at movement M, requiring thus notification of the home network, is given by

$$E[M]_{\text{hmip}} = \frac{2N - r - 1}{N - k}\text{MR}. \tag{1}$$

In case $r = 1$ then each subnet consists of one regional network. As the number r is increased, the probability to move out of a regional network will be increased, increasing thus $E[M]_{\text{hmip}}$. In the case of mobile-SCTP, there are no regional networks. This means that every time the MH moves out of a subnet it must inform the home network. Therefore

$$E[M]_{\text{msctp}} = 2\text{MR}. \tag{2}$$

3.2. Packet delivery

Packet delivery cost is crucial overhead in mobile IP's performance, as packet tunneling is necessary even when the MH moves infrequently [21]. As mentioned earlier, the use of various optimizations like HMIP [2] are not sufficient to eliminate this problem. By moving mobility management "one layer up", the need for tunneling is vanished since the protocol can now accommodate IP address changes.

We now calculate the overhead from packet delivery for MIP, HMIP, MIP-RO, and our end-to-end protocol. As in [18,19] we also define the following variables: s_h and s_g are the packet processing cost at the HA and GFA respectively. L_{ch}, L_{hg}, and L_{gf} are the transmission cost of delivering a packet between the CH to HA, HA to GFA, and GFA to FA respectively. Mobile IP with regional registration has a packet delivery overhead [19]

$$C_{\mathrm{PD}}^{\mathrm{hmip}} = s_h + s_g + L_{hg} + L_{gf} + L_{ch}. \tag{3}$$

Transmission related costs are calculated as follows: $L_{gf} = l_{gf}\delta_D$, where δ_D is a proportionality constant and l_{gf} is the GFA/FA distance. In the case of mobile-SCTP $s_h = 0$ since packets do not go through the HA and so the packet delivery cost with it is simply

$$C_{\mathrm{PD}}^{\mathrm{msctp}} = s_g + L_{cg} + L_{gf}, \tag{4}$$

where L_{cg} is the transmission delay mobile-SCTP, since no tunneling is involved or any other form of intermediate processing of data packets. If λ_{a} is the data packet arrival rate at the HA, then the packet processing delay is analogous to λ_{a} with $s_h = \lambda_{\mathrm{a}}$. In addition, the lookup overhead of the IP routing table has to be calculated which is analogous to its length e, the number of MHs in the subnet n, and course the packet arrival rate λ_{a}. Therefore this cost is $k\lambda_{\mathrm{a}}\log(k)$. We also assume that the distance between CH–HA and HA–GFA is the same, making thus $L_{hg} = L_{ch}$. Eq. (3), will therefore become

$$C_{\mathrm{PD}}^{\mathrm{hmip}} = \lambda_{\mathrm{a}} + k\lambda_{\mathrm{a}}(w_1 nk + w_2 \log(k)) + (l_{gf} + l_{hg})\delta_D. \tag{5}$$

MIP-RO does not suffer from triangular routing but it still has to suffer the tunneling overhead ($c_h = 0$). So

$$C_{\mathrm{PD}}^{\mathrm{mipro}} = s_g + L_{cg} + L_{gf}. \tag{6}$$

Also s_g^{mipro} is the same with s_g^{hmip} because in addition to the routing overhead at the GFA, there is also the tunneling cost present. Only c_h is avoided. Thus,

$$C_{\mathrm{PD}}^{\mathrm{mipro}} = k\lambda_{\mathrm{a}}(w_1 nk + w_2 \log(k)) + (l_{gf} + l_{hg})\delta_D. \tag{7}$$

In addition, in MIP-RO the packets have to be routed to the mobile host, and so the logarithmic $k\lambda_a w_2\log(k)$ IP routing overhead does exist and that is why $s_g \neq 0$.

Finally, $s_g^{\mathrm{msctp}} \neq s_g^{\mathrm{mip}}$, since the GFA in our case does not perform packet decapsulation leading the tunneling cost to zero. But $k\lambda_{\mathrm{a}} w_1 nk$ is zero since there is no decapsulation. So the total packet delivery cost for our protocol becomes

$$C_{\mathrm{PD}}^{\mathrm{msctp}} = k\lambda_{\mathrm{a}} w_2 \log(k) + (L_{gf} + l_{ch,g})\delta_D. \tag{8}$$

3.3. Binding updates

As mentioned earlier, binding or location updates are a requirement in Mobile IP even when the mobile host does not change its care-of address [1]. However, when there is a movement, then either the HA must be informed, or in the case of HMIP, the GFA. We first provide the cost of a binding update for these two cases. As in [18,19] we also define the following variables related to binding costs: C_{hg}, C_{gf}, C_{fm}, are the transmission costs of binding updates between the HA–GFA, GFA–FA, and FA–MH, respectively. Finally ρ is the wireless medium cost and δ is a distance cost unit. So the HA registration cost for HMIP is given by

$$C_{uh} = 2s_f + 2s_g + s_h + 2C_{hg} + 2C_{gf} + 2C_{fm}. \tag{9}$$

Under the HMIP case, the regional registration cost is

$$C_{ur}^{\mathrm{hmip}} = 2s_f + s_g + 2(l_{gf} + \rho)\delta \tag{10}$$

and so the total location update cost per unit time is [19]

$$C_{\mathrm{BU}}^{\mathrm{hmip}} = \frac{E[M]_{\mathrm{hmip}} C_{ur} + C_{uh}}{E[M]_{\mathrm{hmip}} * T_f}. \tag{11}$$

For the MIP-RO case C_{ur} is the same since we assume that MIP-RO operates on top of HMIP. So C_{ur} is again

$$C_{ur}^{\mathrm{mipro}} = 2s_f + s_g + 2(l_{gf} + \rho)\delta, \tag{12}$$

$$C_{\mathrm{BU}}^{\mathrm{mipro}} = \frac{E[M]_{\mathrm{mipro}} C_{ur} + C_{uh}}{E[M]_{\mathrm{mipro}} * T_f}. \tag{13}$$

In mobile-SCTP, the regional registration cost (C_{ur}) is non-existent since the protocol does not send any kind of updates to any local FA or GFA. But it sends directly updates to the CH which can be quite far away and so it has to suffer the cost $C_{u,ch}$ which is higher than C_{ur}. That is why we introduce the aforementioned U_f variable which captures the update frequency to the HA so that we balance out the increased signaling cost due to the direct notification of the CH. The rationale behind U_f being less than 1 is that since we notify directly the CH, the HA does not have to be notified right away for the current connection as we stated earlier (thus the term ‘‘lazy’’ update). Future connections will need the new location of the MH and that is why the higher lag is acceptable. So, when $U_f = 1$ the usual binding frequency is established. The general equation for binding updates (per unit time) in our protocol is

$$C_{\mathrm{BU}}^{\mathrm{msctp}} = \frac{E[M]_{\mathrm{msctp}} C_{u,ch} + C_{uh} U_f}{E[M]_{\mathrm{msctp}} * T_f}, \tag{14}$$

with the transmission cost of binding update between the FA and MH being

$$C_{u,ch} = s_{ch} + 2C_{ch,g} + 2C_{gf} + 2C_{fm}. \tag{15}$$

In this case the packet service time at the correspondent host s_{ch}, is zero since the transport layer endpoint costs are not included, and all the intermediate packet processing costs $2s_h$, $2s_f$, $2s_g$ are equal to zero. The HA registration cost C_{uh}, is the same as in vanilla MIP.

3.4. Packet drop probability

As mentioned in the introduction, the mobile-SCTP protocol utilizes a soft-handoff mechanism. This means that we make the assumption that there are cells with overlapping coverage that remove the possibility of complete detachment from the network. This overlapping means that there are no blackouts due to restricted area coverage, and blackouts can only happen because of the nature of the wireless channel. So in our case, packets may be lost in two cases—due to wireless channel errors, and during handoff when the MH is still receiving data in the old address. However, if connectivity is lost with the previous access point, data packets that were in flight from the CH for this old destination will be lost. Note that the problem is not as severe as in hierarchical mobile IP where a fast or proactive handoff mechanisms have to be used in order to minimize this problem [22]. In our case, the old AP continues to be active and still receives packets while the MH is performing the necessary steps for registering with the a new network.

We showed in the previous section that the average binding update cost per unit time is $C_{\mathrm{BU}}^{\mathrm{msctp}}$. If we consider the packet arrival rate as a Poisson process with average value λ_{a}, then the average number of packet drops per unit time due to binding updates (registration delay) is

$$Drop_{\mathrm{avg}}^{\mathrm{msctp}} = \lambda_{\mathrm{a}} C_{\mathrm{BU}}^{\mathrm{msctp}} \tag{16}$$

and so the packet drop probability due to the latency of binding updates will be

$$P_{\mathrm{BU}}^{\mathrm{msctp}} = C_{\mathrm{BU}}^{\mathrm{msctp}}. \tag{17}$$

However, the packet drop probability also depends on the duration of the disruption time due to handoff in addition to the previously calculated packet loss due to binding update. So the probability of handoff induced packet losses is the probability of the one way latency L (between CH and MH), being smaller than the handoff duration X

$$P_h = P\{L < X\}. \tag{18}$$

We define as f_L the distribution of the end-to-end latency. Concerning the disruption time X, it is a parameter that depends on the mobility management protocol used. In general it can have a fixed value for a specific protocol [19]. Therefore, the only random variable that has to be calculated in Eq. (18), is that of the network latency L_N. We model the one way network latency distribution f_{L_N}, as a shifted Gamma distribution with probability density function [23,24]

$$f_{L_N}(t) = \begin{cases} \dfrac{\lambda e^{-\lambda t} (\lambda t)^{v-1}}{(v-1)!} & \text{if } t \geqslant 0, \\ 0 & \text{it } t < 0, \end{cases}$$

and a cumulative distribution function (c.d.f.) of

$$F_{L_N}(t) = \begin{cases} 1 - e^{-\lambda t \frac{(\lambda t)^v}{v!}} & \text{if } t \geqslant 0, \\ 0 & \text{if } t < 0. \end{cases}$$

This distribution has mean v/λ, and variance v/λ^2. Therefore, by using the previous equations, we can obtain the packet loss probability due to handoff P_h. The total packet loss probability for the mobile-SCTP, MIP, and HMIP protocols, can be derived by combining the previous equations.

4. Simulation results for elastic traffic

In Fig. 3 we show the topology used throughout our experiments in this section. The values for delay and bandwidth of the links shown in this figure, were set to various values according to the specific experiment. The simulation scenario that we used is the following: initially at time X the MH initiates an FTP data flow from the correspondent host (CH). According to the scenario, at time X the

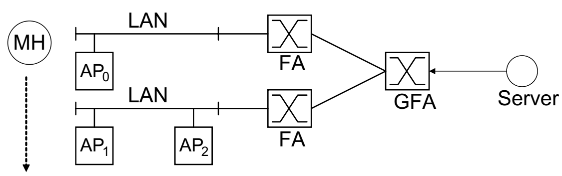

Fig. 3. WLAN topology used for simulations.

MH starts moving away from the first AP, at a speed of 10 m/s, and is heading towards the other AP (see Fig. 3). MH follows the procedure mentioned earlier in order to notify the CH. At time Y, contact with the old AP is lost and the MH can only the new IP. Note that there is a degree of overlapping of the two neighbor cells which corresponds to a few seconds when a host moves with speed of 10 m/s.

We made another modification to the Mobile IP implementation in ns-2 simulator [25], so that it supports smooth handoffs. Packets are buffered in the old AP during handoff, and they are forwarded to the new AP upon completion of the MH registration. The smooth handoff principle was proposed in mobile IP route optimization draft (MIP-RO) [3]. In this way our comparisons are against current state of the art MIP approaches. Generally, our simulations were favoring HMIP. This is because this process has to be done securely [4]: a registration key is established between a foreign agent and a mobile host during the registration process. As a result the whole handoff process is slowing down because this additional delay. However, we did not included this delay in the simulations because we did not have any real life measurement for setting it. On the other hand SCTP, does not need this kind of mechanism since it inherently supports a cookie-based mechanism for securing association between two parties [10].

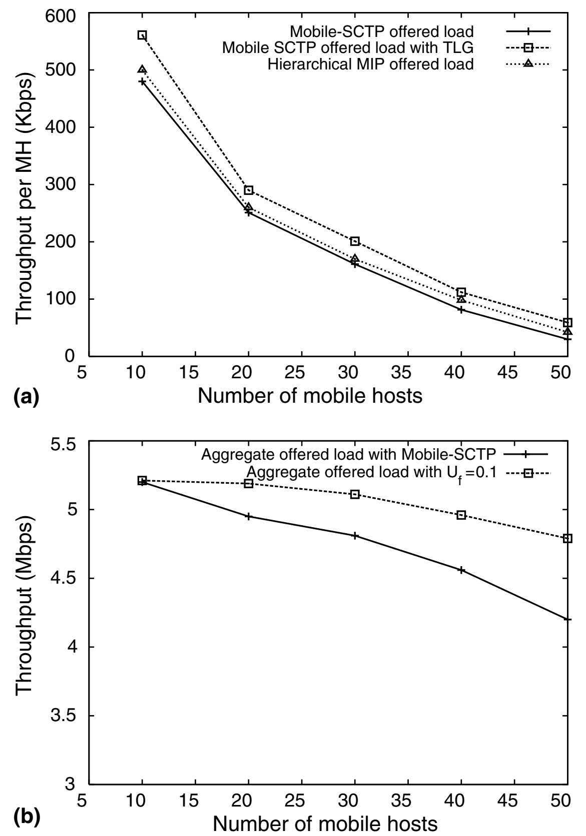

Fig. 4. Throughput for different handoff management protocols. (a) Throughput normalized per mobile host and (b) aggregate throughput.

4.1. Evaluating protocol scalability

For evaluating the ability of the protocol to scale as the number of mobile hosts per subnet is increased, we performed a series of simulation experiments. The capacity of the wired links is l0 Mbps and the delay to 10 ms. Concerning the wireless 802.1b link we set its effective capacity to 60%. In addition, we use the random waypoint movement model for setting the movement pattern of the MHs, that exists in the ns-2 simulator. The experiments were repeated 100 times for each scenario, that involved the use of specific number of mobile hosts per subnet. Fig. 4 presents throughput results for this set of simulations.

As the number of mobile hosts is increased in the subnets, contention for the wireless medium is increased, leading thus to an increase of the packet loss rate. This phenomenon will lead not only to the reduction of the actual goodput, or offered load, for a MH but also to loss of binding updates and even bigger throughput reduction. That is the reason we observe an increased in the number of binding update messages, as the number of mobile hosts is increase in Fig. 4(a). As a general observation though, we can say that the load due to the binding updates to the correspondent host, is less than 1% of the offered load or actual data rate.

In addition we evaluated how the effect of lazy binding updates can affect scalability, and the overall performance of the system. Fig. 4(b) presents results for this part of our simulation. It is interesting to note that while the number of mobile hosts is increased, the baseline mobile-SCTP protocol suffers from degradation in the overall throughput. Additional packets were lost in this case, since the MH with mobile-SCTP, was waiting for the binding update to reach the CH. However, when we set $U_f = 0.1$, which means that the MH notifies the HA after 10 new registrations, we observe a throughput improvement. The reason is now that the MH basically only communicates with the CH and occasionally notifies the HA. This result stresses the significance of the binding updates to the HA, which should be kept to a minimum, unless there is a need to keep the precise location of the MH at all times.

4.2. Results for downlink data flows

Data flows on the downlink, pose the most difficult problems since they require some sort of buffering in case a MH moves out of the range of its old AP. Fig. 5 shows how the sequence number progressed when vanilla SCTP was employed under Mobile IP with a downlink flow present. We can see indeed from Fig. 5(b) that our system has no losses, while in Fig. 5(a) even if we used buffering at the old AP and then data were forwarded to the new AP after the registration phase was finished. Even in this case, with buffering, we see that registration delay is a problem. Even a small delay can cause spurious re-transmissions and even lead to RTO timeouts. This situation could also happen if the FA is located far from the MH, adding thus more to the handoff delay.

5. Simulation results for media flows

A relatively different network setup was used for evaluating the performance of mobile-SCTP protocol in heterogeneous wireless networks. We used the again the ns-2 simulator to evaluate the performance of the handoff algorithm. According to the simulation scenario, handoff is performed at the 30th second from two heterogeneous links. These links can be WLAN, GPRS, or UMTS. Link buffer is setup to be 7 packets for GPRS and 20 packets for UMTS, the propagation delays 300 and 150 ms, respectively, while the bitrates are 30 Kbps for GPRS and 384 Kbps for UMTS [26]. For WLAN

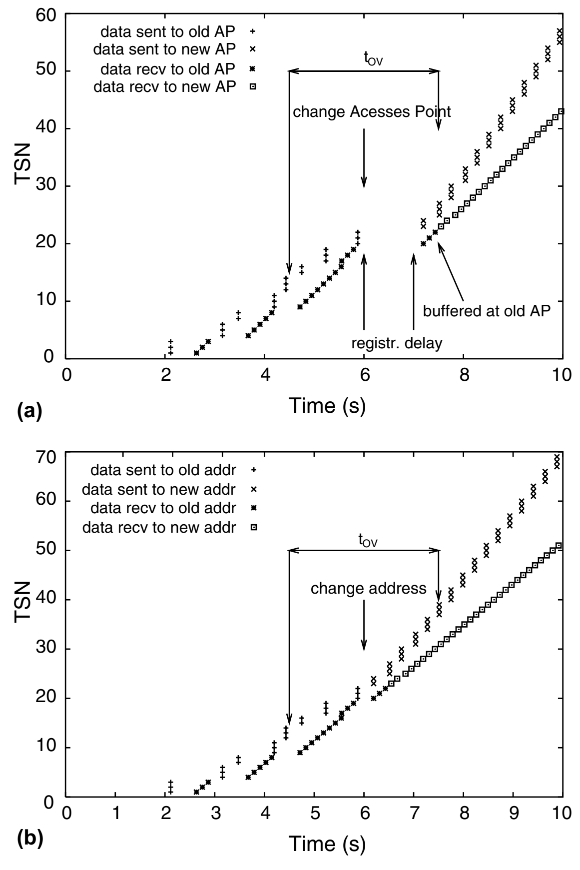

Fig. 5. Sequence number progression for downlink data flows in low bandwidth symmetric links. Host speed 20 m/s. (a) vanilla SCTP over MIP-RO and (b) mobile-SCTP.

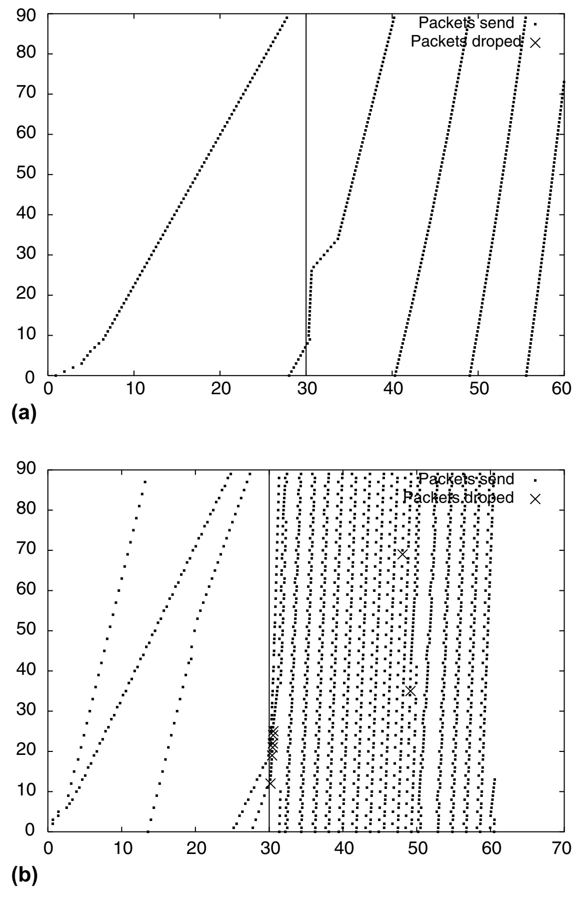

Fig. 6. Vertical handoff from GPRS to UMTS link. (a) Packet sequence numbers for TFRC and (b) packet sequence numbers for mobile-SCTP.

802.11b, the same parameters were set to 20 packets, 10 ms, and 6 Mbps, respectively.

Fig. 6 depicts the case of a vertical handoff from a GPRS to a UMTS link with the TFRC rate control protocol [27], and mobile-SCTP. Note that the TFRC protocol is especially engineered for the support of media flows that have low delay and jitter requirements. However, the performance of a protocol like this, suffers from significant performance problems in case of handoffs, and the only solution for fast adaptation to the new link, is an explicit L2 notification [26]. However, the mobile-SCTP protocol does not require any explicit handoff notification as TFRC. The result is that the AIMD congestion control algorithm used by SCTP, adapts faster to the new link conditions as a handoff from a GPRS link to a UMTS link takes place.

Now in Fig. 7, we present similar figures but for handoff from a fast UMTS link to a slower GPRS. The situation in this case is different since TFRC cannot adapt fast to the new link characteristics resulting into big number of losses (Fig. 7(a)). However, a window-based AIMD congestion control algorithm, during handoff results into a reduced number of losses since its algorithm adapts fast to the new link. A bunch of losses cause mobile-SCTP to reduce its rate faster than TFRC. When buffering of in-flight packets is used at the old AP, during handoff, then TFRC can perform better since is can eliminate several packet losses (Fig. 8(a)). However, the adaptation to the new link conditions remains slow, due to the inherent behavior of the algorithm. When the mobile-SCTP protocol is used, we can see in Fig. 8(b) that packet losses are also eliminated and throughput is no throttled back. In this case we have the best combination where both packet losses are eliminated and fast rate adaptation is achieved.

We also evaluated the performance of mobile-SCTP during handoffs in terms of the experienced

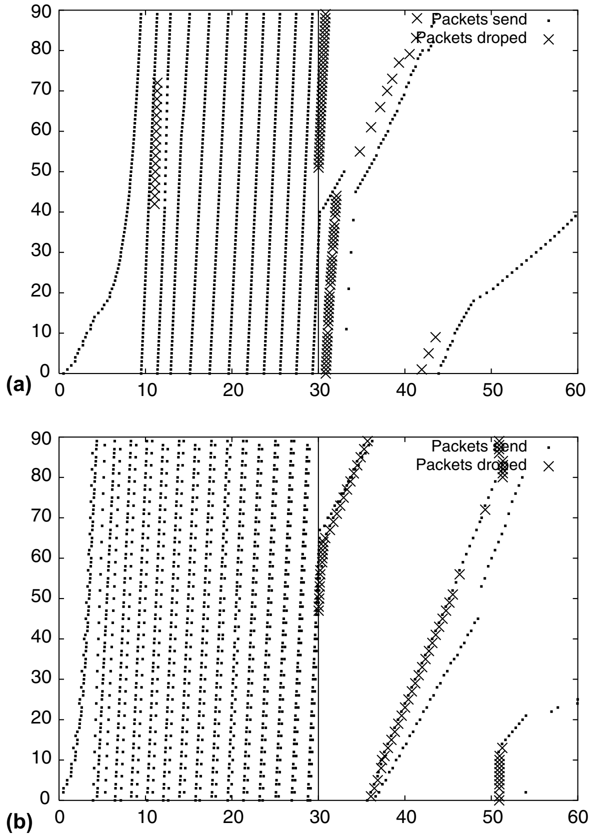

Fig. 7. Vertical handoff from UMTS to GPRS link. (a) Packet sequence numbers for TFRC and (b) packet sequence numbers for mobile-SCTP.

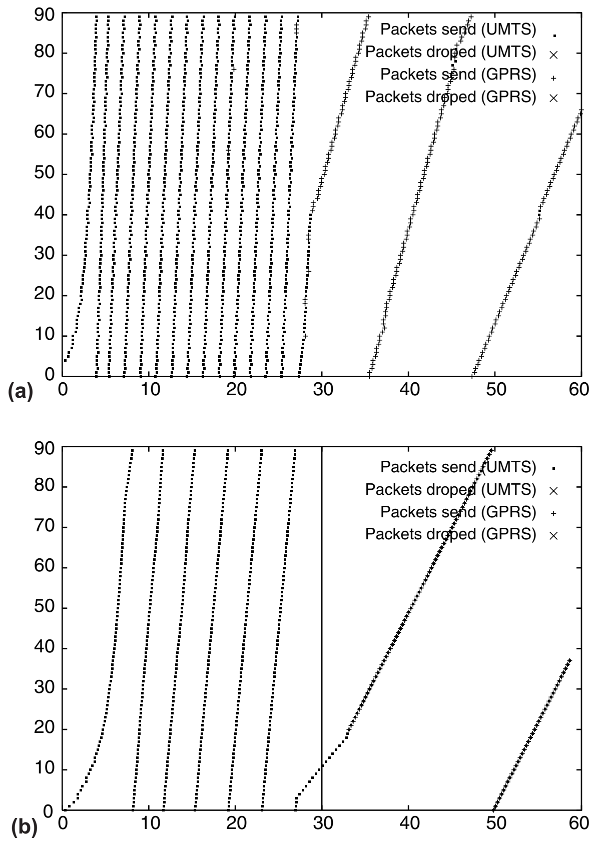

Fig. 8. Vertical handoff from UMTS to GPRS link with buffering. (a) Packet sequence numbers for mobile-SCTP and (b) packet sequence numbers for TFRC.

jitter, since the TFRC protocol has as its primary objective stable jitter so that media flows can be supported. Fig. 9 presents results for a UMTS to WLAN handoff experiment. Even though TFRC suffers generally from less jitter, the adaptation to the experienced jitter in the new link is slow (Fig. 9(b)). For SCTP however during the handoff, jitter instantaneously increases and drops fast when connection with the new link is achieved (Fig. 9(a)). In Fig. 10(a) we present results for instantaneous jitter for UMTS to GPRS handoff with TFRC. In this figure we see that TFRC offers relatively stable jitter. For SCTP however, when it performs a handoff from a fast WLAN to a slower UMTS, it suffers from jitter for quite some time after the time of handoff. This is the result of the reliable nature of the protocol which uses a timer for controlling retransmission of lost packets during handoff, similar to TCP.

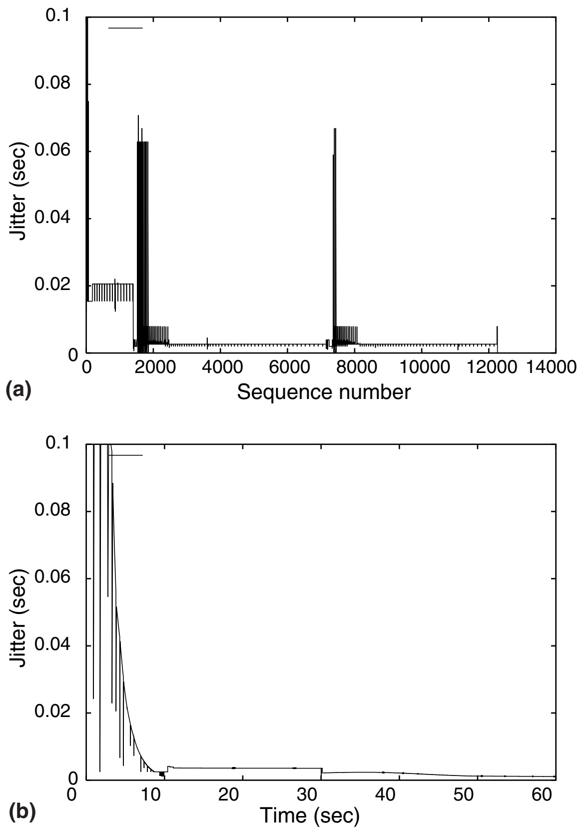

Fig. 9. Jitter for handoff experiment with CBR flows from a UMTS to WLAN 802.11b link. (a) Handoff from UMTS to WLAN with SCTP and (b) handoff from UMTS to WLAN with TFRC.

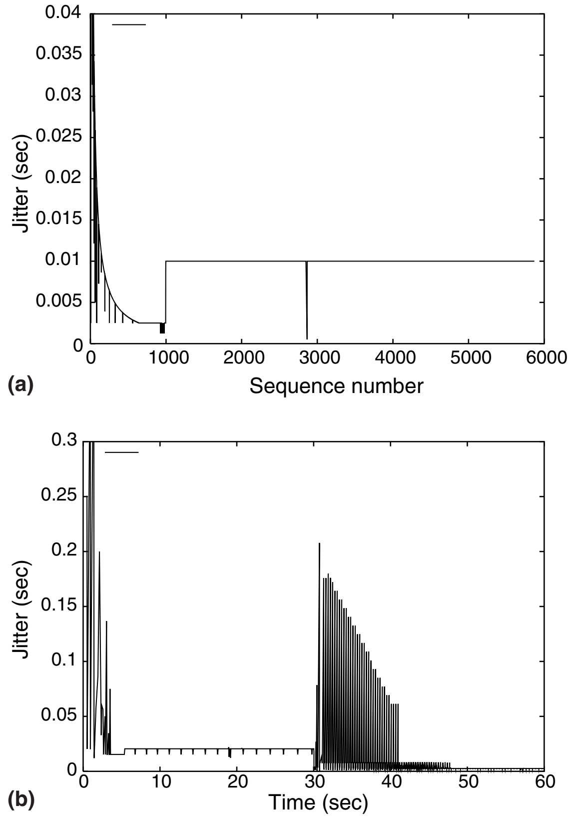

Fig. 10. Jitter for handoff experiment with CBR flows. (a) Handoff from UMTS to GPRS with TFRC and (b) handoff from WLAN to UMTS with SCTP.

6. Conclusions

In paper we presented a lightweight protocol for end-to-end handoff management in heterogeneous wireless IP-based networks. The protocol is based on existing enhancements to the stream control transmission protocol (SCTP), that employs a soft-handoff mechanism for handoff signaling. Our simplified first-order models, show the significant overhead that the mobile IP tunneling is introducing to the overall system performance. Minimization of binding updates to the home network are key component for the improved protocol performance when compared with HMIP, since reduced signalling load increases scalability. We also presented simulation results that for vertical handoff scenarios between WLAN, UMTS and GPRS networks. Our results suggest that the AIMD congestion control algorithm, adapts faster to the new link when handoff from a slow to a fast link takes place (GPRS/UMTS or UMTS/WLAN). This situation leads to better bandwidth utilization which is increased even more when the soft-handoff mechanism is employed due to the reduced number of packet losses.

References

[1] C. Perkins, IP Mobility Support for IPv4, RFC 3220, January 2002.
[2] E. Gustafsson, A. Jonsson, C.E. Perkins, Mobile IPv4 Regional Registration, draft-ietf-mobileip-reg-tunnel-07.txt.
[3] C. Perkins, D.B. Johnson, Route Optimization in Mobile IP, draft-ietf-mobileip-optim-11.txt, September 2001.
[4] W. Haitao, Z. Shaoren, The security issues and countermeasures in Mobile IP, in: ICII, vol. 5, Beijing, 2001, pp. 122–127.
[5] R. Ramjee, T.F.L. Porta, S. Thuel, K. Varadhan, S.Y. Wang, HAWAII: A domain-based approach for supporting mobility in wide-area wireless networks, in: ICNP, 1999, pp. 283–292. Available from: <citeseer.nj.nec.com/ramjee99hawaii.html>.
[6] A. Campbell, J. Gomez, S. Kim, A. Valko, C. Wan, Z. Turanyi, Design, implementation, and evaluation of Cellular IP, 2000. Available from: <citeseer.nj.nee.com/article/campbell00design.html>.
[7] P. Nikander, J. Lundberg, C. Candolin, T. Aura, Homeless Mobile IPv6, 2001. Available from: <http://www.tml.hut.fi/pnr/publications/draft-nikander-mobileip-homelessv6-01.txt>.
[8] A.C. Snoeren, H. Balakrishnan, An end-to-end approach to host mobility, in: ACM International Conference on Mobile Computing and Networking (MOBICOM), Boston, MA, 2000, pp. 155–164.
[9] J. Postel, Domain name system structure and delegation, RFC 1591, 1994.
[10] R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, V. Paxson, Stream Control Transmission Protocol, RFC 2960, October 2000.
[11] W. Xing, H. Karl, A. Wolisz, M-SCTP: Design and prototypical implementation of an end-to-end mobility concept, 2002. Available from: <citeseer.nj.nec.com/558661.html>.
[12] M. Riegel, M. Tuexen, Mobile SCTP, draft-riegel-tuexen-mobile-sctp-02.txt, 2002.
[13] S.J. Koh, M.J. Chang, M. Lee, mSCTP for soft handover in transport layer, IEEE Communication Letters 8 (3) (2004) 189–191.
[14] R. Droms, Dynamic Host Configuration Protocol, RFC 2131, March 1997.
[15] K. Egevang, P. Francis, The IP Network Address Translator (NAT), RFC 1631, 1994.
[16] C.P.D. Jonhson, J. Arkko, Mobility Support in IPv6, RFC 3775, June 2004.
[17] R. Stewart, Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration, draft-ietf-tsvwg-addip-sctp-07.txt, March 2003.
[18] W. Chen, J. Ho, Performance analysis of adaptive location management for mobile IP, Technical Report 97-CSE-13, Southern Methodist University, 1997.
[19] J. Xie, I.F. Akyildiz, A novel distributed location management scheme for minimizing signaling costs in mobile IP, IEEE Transactions on Mobile Computing 1 (3) (2002) 163–175.
[20] Digital Subscriber Line (DSL) and the Evolution of Asynchronous Transfer Mode (ATM) Networks, 2004. Available from: <http://www.iec.org/online/tutorials/dsl_atm/topic01.html>.
[21] I. Stojmenovic, Handbook of Wireless Networks and Mobile Computing, John Wiley and Sons, 2002.
[22] P. Reinbold, O. Bonaventure, A survey of IP micromobility protocols, 2002. Available from: <citeseer.nj.nec.com/reinbold02survey.html>.
[23] J. Chakareski, B. Girod, Rate-distortion optimized packet scheduling and routing for media streaming with path diversity, in: IEEE Data Compression Conference (DCC), 2003.
[24] Y.-C. Lee, Error resilient video streaming over lossy networks, Ph.D. thesis, Georgia Institute of Technology, 2003.
[25] Network simulator. URL `http://www.isi.edu/nsnam/ns/`.
[26] A. Gurtov, Effect of vertical handovers on performance of TCP-friendly rate control, in: ACM Computer Communications Review, 2004.
[27] M. Handley, S. Floyd, J. Pahdye, J. Widmer, TCP Friendly Rate Control (TFRC): Protocol Specification, RFC 3448, January 2003.

Antonios Argyriou received the degree in electrical and computer engineering from the Democritous University of Thrace, Greece, in 2001, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, in 2003 and 2005, respectively. His primary research interests include wireless networks, mobile computing and multimedia communications. He is a member of the IEEE and ACM.

Vijay Madisetti is a professor of electrical and computer engineering at the Georgia Institute of Technology. He splits his time among teaching, research and entrepreneurship. His interests are design, prototyping, and packaging of electronic systems, virtual prototyping, embedded software systems, and computer networks. He obtained his Ph.D. in electrical engineering and computer science from the University of California at Berkeley. He is a member of the IEEE and the Computer Society.