Fast-handoff Schemes for Application Layer Mobility Management

Ashutosh Dutta, Sunil Madhani, Wai Chen
Telcordia Technologies Inc., 444 Hoes Lane, Piscataway, NJ 08854

Onur Altintas, Toyota InfoTechnology Center, 444 Hoes Lane, Piscataway, NJ 08854

Henning Schulzrinne
Computer Science Department, Columbia University, New York, NY 10027

Abstract—
In order to ensure proper quality of service for real-time communication in a mobile wireless Internet environment it is essential to minimize the transient packet loss when the mobile is moving between different cells (subnets) within a domain. Network layer mobility management schemes have been proposed to provide optimized fast-handoff for multimedia streams during a client's frequent movement within a domain. This paper introduces application layer techniques to achieve fast-handoff for real-time RTP/UDP based multimedia traffic in a SIP signaling environment. These techniques are based on standard SIP components such as user agent and proxy which usually participate to set up and tear down the multimedia sessions between the mobiles. Unlike network layer techniques, application layer techniques do not have to depend upon any additional components such as home agent and foreign agent. It thus provides a network access independent solution suitable for application service providers.

I. Introduction

Handover delay during a mobile's frequent movement in a wireless environment consists of latency factors at different layers: layer 2 movement detection, IP address discovery and configuration, and signaling to redirect the media to mobile's destination. Latency as a result of layer 2 detection is mostly determined by the access method being used such as 802.11b or CDMA. While layer 2 detection may help trigger the IP address discovery process, latency contributed by the IP address configuration can be attributed to the types of protocol being used such as DHCP [1], PPP [2], etc. References [3], [4] and [5] provide approaches to reduce the latency associated with the first two factors. Once the client is configured with the new IP address in its new domain, signaling from the client to redirect the media will contribute to the third component of latency. This paper focuses on the techniques to reduce the transient data loss due to media re-direction in a SIP based mobility environment. Figure 1 shows different latency factors during a node's movement represented as t_1, t_2 and t_3.

In order to provide seamless mobility support to the clients in a mobile wireless Internet environment, several variations of Mobile IP [6] such as [7], [8], [9], [10] have been proposed. Mobile wireless Internet telephony uses SIP [11] signaling to establish and tear down multimedia sessions. These multimedia sessions are mostly based on RTP/UDP [12] and have different delay, error characteristics than standard TCP/IP-based

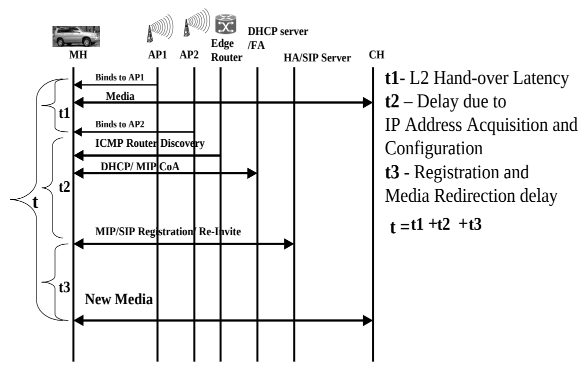

Fig. 1. Handoff Latency Factors

applications. Application layer mobility management scheme [13], [14] uses SIP to provide an alternative mobility solution for real-time interactive and streaming traffic. This mobility scheme does not depend upon components such as home agent and can be deployed by any third party application service provider without much dependence upon the underlying network elements.

The rest of the paper is organized as follows. Section II describes the related work in the area of fast-handoff for Intra-domain mobility. Section III provides several alternate techniques for SIP based fast-handoff. Implementation details of one these techniques are described in section IV. We finally conclude the paper in section V.

II. Related Work

Once the mobile moves to a new domain, most of the movement is limited to the subnets within the new domain until the mobile moves out again to another domain. However even in this case the communicating mobiles can be far apart or the mobile may be far away from the home agent. Thus it is necessary to limit the signaling traversal within the domain itself during the mobile's intra-domain mobility. There are several layer 3 based intra-domain mobility management solutions such as [15], [5], [16], [17], [18] to help reduce the transient data loss. However, some of these techniques require changes in the end systems and depend upon additional components in the home or foreign domains. Since SIP provides an application layer mobility management solution for real-time traffic [13], it is helpful to augment this approach with fast-handoff techniques. Vakil et al [19] provides a virtual soft-handoff approach for CDMA-based wireless IP networks using SIP signaling. But it does not provide a generalized solution suitable for other types of access networks. Thus, it is desirable to design an architecture that can provide a generalized SIP based fast-handoff solution for real-time traffic independent of type of access network.

III. SIP Fast-handoff techniques

SIP user agents and SIP servers are the core parts of a SIP based architecture. Every SIP user in the wireless Internet has a home proxy server, that provides the authentication for the roaming user by interacting with AAA servers and other SIP servers in the visited domain. It is natural to assume that each visited domain has a SIP proxy where a visiting mobile can register during its movement [20]. In some cases however it may be desirable to install multiple SIP servers to ensure redundancy and proper load balancing by means of DNS (Domain Name System) based "SRV" mechanism.

We propose several methods based on SIP signaling to help alleviate transient data loss due to signaling as a result of continuous hand-offs within a domain. In a typical scenario when the mobile host moves from one subnet to another it changes the layer 2 attachment, obtains a new IP address and sends a new Re-INVITE to the CH (Correspondent Host). During a consecutive move if the CH happens to be far away from MH (Mobile Host), then it takes a long time before the data from CH gets re-directed to the new address, thus resulting in transient data loss. Some of the methods that can be applied to help take care of the transient data loss can be achieved by intercepting and forwarding the transient traffic or by multicasting these packets pro-actively for a short period of time.

Visited domains can consist of several subnets. Every move to a new subnet causes the MH to send a re-INVITE to the CH containing its new care-of address. If the re-INVITE request gets delayed due to path length or congestion, media packets will continue to be directed to the old address. We assume that the visited network has an outbound proxy that every client gets configured with via DHCP. We enhance this proxy with the ability to temporarily register visitors [20]. The visitor obtains a temporary, random identity from the visited network and uses it as its address-of-record to register with the registrar in the visited network. The MH also informs the home registrar of this temporary address. It then only updates that registration with its current local IP address. This process speeds up registrations, but does not address the "delayed binding update" issue. In-transit packets can be redirected to a unicast or multicast address based on the movement pattern of the mobiles and usage scenario. Below, we describe several ways to achieve fast handoff using SIP: using a SIP registrar and RTP translator or NAT (Network Address Translator); using the outbound proxy; using a multicast agent; and B2BUA (Back to Back UA) as a mobility agent. Figure 2 illustrates the motivation for a SIP based fast-handoff scenario within the Internet.

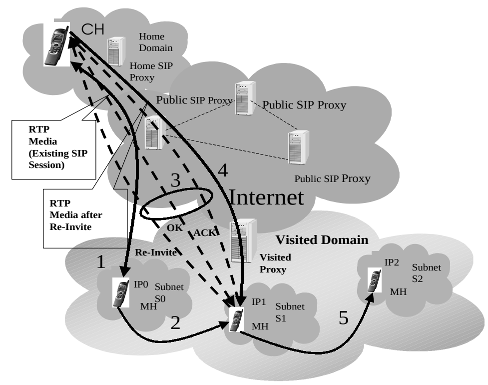

Fig. 2. SIP-based fast-handoff applicability

Following subsections describe some possible ways of achieving fast-handoff using SIP based mechanism.

A. SIP registrar and RTP translator or NAT

For every move the mobile makes within a domain while obtaining a new IP address, it sends a Re-INVITE to the correspondent host so that the new traffic gets forwarded to the new destination of the mobile. Because of the distance between CH and MH and congestion associated with the routing in the network, SIP Re-INVITE may get delayed. Thus, during this time, transient traffic is still sent to the old destination, and thus not being received by the mobile. In order to take care of this transient data an application layer approach that combines both SIP and RTP translator has been designed. This approach provides a complete application layer technique.

Each subnet within a domain is equipped with an RTP translator [12] that provides application-layer forwarding of RTP packets for a given address and UDP port to a given network destination. (RTP applications generally do not care about the source IP address of RTP packets, using just the synchronization source identifier (SSRC) to identify the source.) Figure 3 shows a sequence of operations when a mobile host moves from one network to another. The SIP server here can act like a registrar or proxy. The visited-network registrar described earlier receives the registration update from the MH that has just moved, and immediately sends a request to the RTP translator in the network that the MH just left. The request causes the RTP translator to bind to the old IP address used by the MH and forward any incoming packets to the new address of the MH. After a set interval or after no media packets have been received by the RTP translator, the RTP translator relinquishes this old address and removes the forwarding

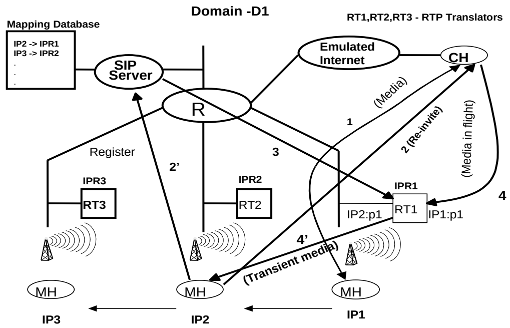

Fig. 3. RTP translator based fast-handoff

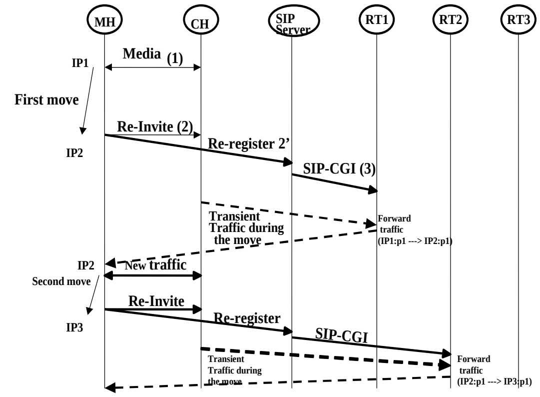

Fig. 4. Fast-handoff Flow

table entry, assuming that the re-INVITE has reached the CH. In practice this time period should be no longer than a few seconds at most.

In figure 3, RT1, RT2 and RT3 are RTP translators in the respective subnets. RTP translators forward the traffic associated with one IP address/port number to another IP address/port number. RTP translator in each of these subnets intercepts the traffic meant for the mobile host and sends it to the new address of the mobile host after capturing it. As the mobile host moves from one location to another and obtains a new address, (say, IP2), it sends a register message to the SIP registrar. The server in turn looks up in its database and sends a message to the appropriate RTP translator to forward the transient traffic to the mobile's new address. This message can be via SIP-CGI [21] or a simple UDP signaling-based triggering can be used. If the distance between CH and MH is large enough, then it may take some time for re-INVITE to reach CH before the new media gets re-directed to the proper address. It is noteworthy to mention that, these RTP translators can also forward the transient traffic to a duration limited multicast address until the new data comes from the CH. In the absence of timely ageing, there is a likelihood that some other client may like to use this address. Hence it is absolutely essential that there should be a mechanism to age this address out of the virtual interface of the RTP-translator as soon as the traffic redirection stops from the previous subnet. De-activation can be triggered as soon as the correspondent host stops sending the packets to mobile's previous network or the mobile goes back to its original subnet.

There are several ways the ageing can be implemented to remove the virtual interface in the previous subnet. As a most common approach the virtual interface can timeout after particular time period say a preset expiry time. A second approach can be to monitor the traffic destined to the mobile's old address, if it does not see for a while then it removes the virtual interface (deletes the ARP entry) from the proxy server thus releasing that IP address to be used by another mobile. Another way of de-activating the virtual interface will be to signal the SIP server that packets have started arriving directly from the correspondent host. This can be done by observing the source address of the packets, since the mobility proxy with iptables actually changes the source address to be that of the correspondent host. Figure 4 shows a flow diagram for the fast-handoff operation using RTPtrans approach.

B. *SIP outbound proxy*

Using SIP outbound proxy is another method of supporting fast-handoff. SIP requests typically traverse a SIP proxy in the visited network, the *outbound proxy*. It can use the data in the MH-to-CH re-INVITE to configure the RTP translator or NAT. The advantage of this approach is that the outbound proxy usually has access to the Session Description Protocol (SDP) information containing the MH media address and port, thus simplifying the configuration of the translator or NAT (Network Address Translator). On the other hand, this outbound proxy has to remember the INVITE information for an unbounded amount of time and become call stateful, since it needs the old information when a new re-INVITE is issued by the MH.

C. *B2BUA approach*

Another way of providing fast-handoff is by using a back-to-back SIP user agent (B2BUA). A B2BUA consists of two SIP user agents where one user agent receives a SIP request, possibly transforms the SDP parameters and then has the other part of the B2BUA re-issue the request. A B2BUA in each domain needs to be addressed by the MH in the visited domain. The B2BUA issues a new request to the CH containing its own address as the media destination and then forwards the packets, via RTP translator or NAT, to the MH. This approach however requires some cooperation from the MH. As noted, the INVITE request needs to be addressed explicitly to the B2BUA, as otherwise end-to-end encryption of the body may prevent the B2BUA from inspecting it. B2BUA in addition to setting up a call between CH and MH, also sets up a

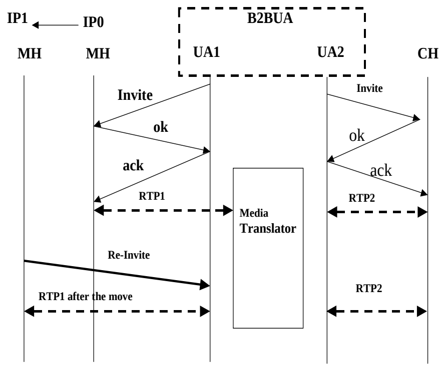

Fig. 5. B2BUA fast-handoff flow

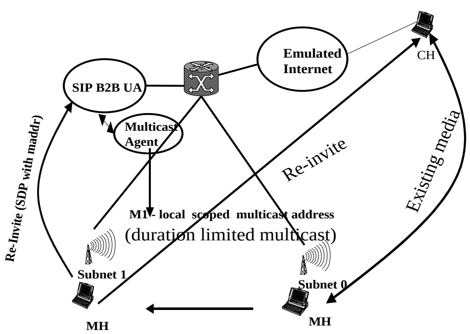

Fig. 6. SIP fasthandoff with multicast agent

call between itself and CH. As the mobile moves to a new subnet, it sends a Re-invite message and also an Invite message to the B2BUA. A time bound transient session is established between B2BUA and the MH where a transient data is delivered until the new media arrives from CH. This can continue for MH's each subsequent move within a domain. Figure 5 shows the flow associated with this approach.

D. SIP Fast-handoff using Multicast Agent and B2BUA

Locally scoped multicast may help to avoid packet losses if the MH can predict that it will be moving to a new subnet shortly. In that case, it informs the visited registrar or B2BUA of a temporary multicast address as its contact or media address. Once the MH has arrived in its new subnet, it updates the registrar or B2BUA with its new unicast address, while continuing to listen to multicast address. Using scoped multicast is only effective if the MH can quickly acquire a multicast address. Multicast agent may co-locate with the first-hop router or can co-exist with the B2BUA or SIP proxy. Figure 6 illustrates this technique.

IV. Implementation Details

We have implemented the SIP based fast-handoff architecture as described in Figure 2 in multimedia test-bed. The basic components are SIP registrar, SIP UA, RTP translator. Modified version of Columbia University SIPC was used in the experimental test-bed. All the network elements CH, MH and SIP proxy are Linux based with kernel version 2.4.7-10. This enables iptables NAT functionality to change the source and destination address of the media packets for different applications such as audio and video. In this particular implementation both RTP translator and SIP proxy co-exist. Etherape measurement tool was modified so that GUI can show all the packet redirection including media and signaling traffic between different entities on the screen. Tcpdump tool was used to capture the RTP packets with time-stamp on CH, SIP proxy and SIP mobile host. It provides a time scale of how much time it took to forward the packet and packet loss associated.

SIP signaling re-INVITE was synthetically delayed using NIST delay simulator to simulate the network congestion or distance between CH and MH. Important feature of this experiment was to capture the transient packets as the mobile moves from one subnet while Re-INVITE takes a long time to reach the correspondent host. We assume that the common network router is not too far from the subnet routers in each adjacent cell, thus the registration message will not take that much time to reach the SIP server compared to the re-INVITE signal. Both VIC (Video Conferencing Tool) and RAT (Robust Audio Tool) were used to measure the performance of audio and video streaming traffic, respectively. Two methods (e.g., rtptrans and NAT iptables) were used to direct the transient traffic from the previous subnet to the new one. It was observed that although "rtptrans" tool actually changes the IP address of the outgoing traffic to be that of the RTPtranslator, but iptables with NAT functionality (Mobility Proxy) does not change the source IP address rather maintains it to be that of CH. VIC and RAT application behaved bit differently with rtptrans because VIC uses connected socket with bind() and does not pick up the packets if the source address is different.

We tried few experiments with re-INVITE traversal values of 100 ms, 200 ms, 500 ms, 1 sec, 2 sec, 3 sec and 5 sec delay to show how RTP translator helps speed up the delivery of RTP packets and enhances smooth handoff mechanism. Both RTPtrans approach and mobility proxy methodology were used to capture the traffic and forward this to the new location of the client.

A. Performance Evaluation

Number of transient RTP packets forwarded contributing to smooth handoff will depend upon the type of optimization technique implemented. Following provides a quantitative analysis using RTP translator methodology.

The time taken for complete subnet movement including IP address acquisition and layer 2 movement is Ts. Time taken for Re-INVITE to reach CH is Ti (mostly decided by

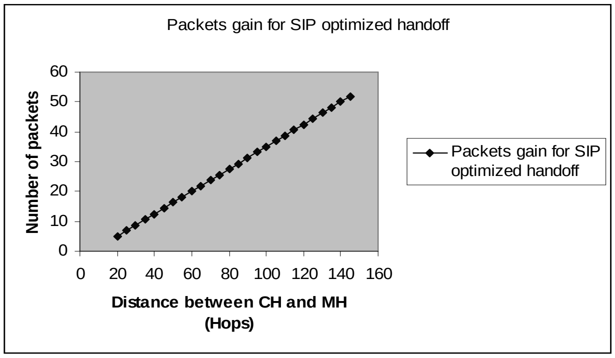

Fig. 7. Packet gain for SIP optimized fast-handoff

the distance factor). Time taken to process Re-INVITE at CH is Tp, time taken to register at the SIP proxy is Tg, time taken for the SIP registrar to forward the packet after capturing is Tf. Packet generation rate at CH is Pr per second.Thus total number of packets lost during a simple Re-Invite with subnet movement is Pi= (Ts+Ti+Tp)*Pr. Total number of packets lost during using SIP registration and RTPtrans is Prt= (Ts+Tg+Tf)*Pr.

According to [14] a sample Re-Invite processing time at the CH is about 100 ms. Complete registration takes about 150 ms. It takes about 200 ms to complete the subnet movement and IP address acquisition including the layer 2 detection. We measured the packet delay due to redirection at the registrar as less than 1ms when iptables-based NAT approach was used, where as RTP translator approach added 4 ms of delay. In a 802.11b environment it lost about 15 packets due to the delay associated with Re-INVITE processing time, registration and packet forwarding.

Figure 7 compares the efficiency of SIP based optimized handoff approach using a combination of mobility proxy and RTP translator with basic SIP based mobility management without fast-handoff technique. As it shows number of packets gained increases as distance in terms of number of hops increases between CH and MH given a specific packet generation rate.

There is a likelihood that duplicate packets are received during the mobile's movement between the subnets. RTP packets have their own sequence numbers associated and thus should not be harmful. Although mechanisms similar to described in [16] can be adopted to take care of duplicate packets from being sent up to the application. Rapid handoff between multiple subnets, ageing, registration with the SIP servers are some of the factors that may contribute to the scalability when deployed in a real network.

V. Conclusions

In this paper we presented several application layer fast-handoff approaches based on SIP based mobility management. These novel approaches help reduce transient data loss due to a mobile's frequent intra-domain handoff by limiting the frequent signaling update within a domain. Results from analysis and implementation for RTPtrans based approach show that by employing these optimization technique we achieve up to 80 percent improvement of Packet-to-Loss factor compared to basic SIP based mobility management meant for real-time traffic.

References

[1] R. E. Droms, "Dynamic host configuration protocol," RFC 2131, Internet Engineering Task Force, Mar. 1997.

[2] W. Simpson, "The point-to-point protocol (PPP)," RFC 1661, Internet Engineering Task Force, July 1994.

[3] J.-O. Vatn and G. C. Maguire, "The effect of using co-located care-of addresses on macro handover latency," in *14th Nordic Tele-traffic Seminar*, (Technical University of Denmark, Lyngby, Denmark), Aug. 1998.

[4] A. McAuley, S. Das, S. Madhani, S. Baba, and Y. Shobatake, "Dynamic registration and configuration protocol (DRCP)," internet draft, Internet Engineering Task Force, July 2000. Work in progress.

[5] K. Malki *et al.*, "Low latency handoffs in mobile IPv4," internet draft, Internet Engineering Task Force, June 2003. Work in progress.

[6] C. Perkins, "IP mobility support for IPv4," RFC 3344, Internet Engineering Task Force, Aug. 2002.

[7] A. Campbell, J. Gomez, S. Kim, A. G. Valk, C.-Y. Wan, and Z. R. Turnyi, "Design, implementation, and evaluation of cellular IP," *IEEE Personal Communications Magazine*, vol. 7, pp. 42–49, Aug. 2000.

[8] R. Ramjee, T. F. LaPorta, L. Salgarelli, S. Thuel, K. Varadhan, and L. Li, "IP-based access network infrastructure for next-generation wireless networks," *IEEE Personal Communications Magazine*, vol. 7, pp. 34–41, Aug. 2000.

[9] S. Das, A. Dutta, A. McAuley, A. Misra, and S. Das, "IDMP: an intra-domain mobility management protocol for next generation, wireless networks," *IEEE Personal Communications Magazine*, June 2002.

[10] A. C. Snoeren and H. Balakrishnan, "An end-to-end approach to host mobility," in *ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, (Boston, Massachusetts, USA), pp. 155–166, Aug. 2000.

[11] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. R. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: session initiation protocol," RFC 3261, Internet Engineering Task Force, June 2002.

[12] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: a transport protocol for real-time applications," RFC 3550, Internet Engineering Task Force, July 2003.

[13] E. Wedlund and H. Schulzrinne, "Mobility support using SIP," in *2nd ACM/IEEE International Conference on Wireless and Mobile Multimedia (WoWMoM)*, (Seattle, Washington), Aug. 1999.

[14] A. Dutta, F. Vakil, J. Chen, M. Tauil, S. Baba, and H. Schulzrinne, "Application layer mobility management scheme for wireless Internet," in *3Gwireless*, (San Francisco), p. 7, May 2001.

[15] A. Jonsson, E. Gustafsson, and C. E. Perkins, "Mobile IPv4 regional registration," internet draft, Internet Engineering Task Force, Oct. 2002. Work in progress.

[16] C. E. Perkins and K. H. Wang, "Optimized smooth handoffs in mobile IP," in *IEEE Symposium on Computers and Communications*, (Red Sea, Egypt), July 1999.

[17] R. Caceres and V. N. Padmanabhan, "Fast and scalable handoffs in support of mobile Internet audio," *ACM Mobile Networks and Applications (MONET) Journal*, 1999.

[18] A. Misra, S. Das, A. Dutta, A. McAuley, and S. Das, "IDMP based fast-handoff and paging in IP based 4G mobile networks," *IEEE Communications Magazine*, Mar. 2002.

[19] F. Vakil, D. Famolari, S. Baba, and D. Famolari, "Virtual soft hand-off in IP-Centric wireless CDMA networks," in *3G Wireless 2001*, (San Francisco), Delson, Delson, May 2001.

[20] H. Schulzrinne, "SIP registration," internet draft, Internet Engineering Task Force, Apr. 2001. Work in progress.

[21] J. Lennox, H. Schulzrinne, and J. Rosenberg, "Common gateway interface for SIP," RFC 3050, Internet Engineering Task Force, Jan. 2001.