Comparing Handoff Performance of Mobile IP, Fast Handoff and mSCTP in Wireless Networks*

Ken C.K. Tsang, Cho-Li Wang and Francis C.M. Lau
Department of Computer Science
The University of Hong Kong, Hong Kong
{cktsang, clwang, fcmlau}@cs.hku.hk

Abstract

We compare the performance of three handoff protocols, namely Mobile IP, Fast Handoff and mSCTP. Among the three schemes, Mobile IP suffers from the lowest data throughput and longest handoff latency. Fast Handoff can perform better, provided that the mobile node can handoff to the new base station at an appropriate time instant when data forwarding between network routers begins. mSCTP supports multihoming; the mobile node does not need to determine the exact handoff time. Nevertheless, packet reordering and the subsequent fast retransmission degrades its handoff performance. To avoid these problems, adding some flow control operations in the transport layer is necessary. For flow control to be carried out in-sync with the handoff operations, the transport layer needs to be handoff-aware. We therefore conclude a way in designing a handoff scheme, which is to centralise the handoff and flow control operations in the transport layer.

1. Introduction

Mobile devices can access the Internet anywhere anytime via wireless access points in the vicinity, thanks to the extensive wireless network coverage including wireless LANs (e.g., 802.11b-based WLANs) and third-generation cellular networks (e.g., GPRS, UMTS, etc.) in a mobile environment. As a mobile device roams and changes its connectivity to the access point, the existing IP connections to the device need to be terminated and reconnected. This *handoff* process hinders a smooth data transfer and results in performance degradation in applications. A handoff is *vertical* if it happens across different wireless technologies. An *upward* vertical handoff is a handoff to a mobile network with a larger coverage area and lower bandwidth, e.g., from WLAN to 3G; a *downward* vertical handoff is a handoff to a mobile network with a smaller coverage area and higher bandwidth, e.g., from 3G to WLAN [11].

Various handoff protocols have been proposed to support an efficient IP handoff in wireless networks. In the network layer, Mobile IP [12] (MIP) uses IP tunnels to forward packets to a mobile device. Fast Handoff IPv6 (or simply Fast Handoff) [9] eliminates the triangle routing problem in MIP and reduces packet loss at the network routers through buffering. In the transport layer, the recently proposed Mobile SCTP (mSCTP) [10] makes use of SCTP's (the Stream Control Transmission Protocol [14]) multihoming feature to facilitate an efficient handoff. While there exists several researches that suggest mSCTP for handoff [7, 15], it remains uncertain how well does mSCTP perform. In [16], a thorough performance analysis on MIP, SIP [5] and Migrate [3] is presented. Nevertheless, to the best of our knowledge, there is no existing work that compare mSCTP's handoff performance with the well-studied MIP or Fast Handoff.

We can classify the data transfer between a fixed host and a mobile node as follow. First, we consider whether reliable data transfer is required. Second, we consider whether the mobile node, which experiences a handoff, is the data destination (a mobile receiver's handoff) or the data source (a mobile sender's handoff). As an example, large file transfer requires reliable data transfer, and the mobile node can be the data destination (file download) or the data source (file upload). On the other hand, multimedia services like real-time movie retrieval or video conferencing do not have stringent reliability requirement. The mobile node is the data destination during an online movie retrieval, while it is the data source during an Internet conferencing session when voice and image data are sent to the fixed host.

In this research, we conduct simulation experiments to compare the handoff performance of MIP, Fast Handoff and mSCTP under the following types of data transfer: (1) reliable data service to a mobile receiver, (2) reliable data service from a mobile sender, (3) multimedia data service to a mobile receiver, and (4) multimedia data service from a

*This work is supported in part by by National Natural Science Foundation of China (NSFC) Grant No. 60533040 and 60773089.

mobile sender. The first two types are tested in an upward vertical handoff, whereas the last two types are tested in a downward vertical handoff. Data throughput, the amount of packet reordered or loss, and the handoff latency are the performance metrics used in our comparison. Our goals are to study the merits and deficiencies of the three handoff schemes, and how well does mSCTP perform when compared with MIP and Fast Handoff. Furthermore, through the comparison, we aim for some new ideas concerning the design of an efficient handoff protocol in wireless networks.

Section 2 gives an overview on MIP, Fast Handoff and mSCTP. Performance evaluation of the schemes are described in Section 3. Section 4 discusses our ideas on the design of a handoff scheme, and Section 5 concludes.

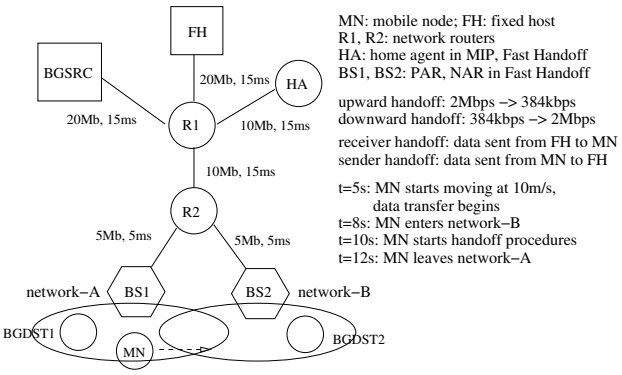


Figure 1. Simulation network topology

2 Preliminaries

1) Mobile IP: Mobile IP [12] is a network layer solution which uses a *home agent* to intercept and forward packets that are sent from the correspondent host to the mobile node. Each mobile node has two addresses, a static home address under its *home network* as its identifier, and a care-of address for packet routing. When it moves to a *foreign network*, it notifies its home agent the new care-of address using the *binding update* message. The home agent then updates its *binding cache entries*—the home address to care-of address pairs of the mobile node. Data are destined to the home address of the mobile node, intercepted by the home agent, and forwarded to the care-of address of the mobile node through IP header encapsulation.

2) Fast Handoff: Fast Handoff follows Mobile IPv6 [4] to avoid the triangle routing problem in MIP. The correspondent host also keeps the binding cache, so that data packets can be sent to the mobile node's care-of address directly. In Fast Handoff, the handoff latency is further reduced by "hiding" the address configuration process (the process by the mobile node to acquire the new care-of address) from handoff [9]; the mobile node configures a new care-of address before handoff while data transfer continues. Fast Handoff also reduces packet loss by (1) using an "IP tunnel" between the access router in the original network (PAR) and that in the new network (NAR), and (2) buffering packets at NAR. Packets reaching PAR are not dropped, but forwarded and buffered at NAR. The mobile node, once attaches to the network, will notify NAR for the buffered packets.

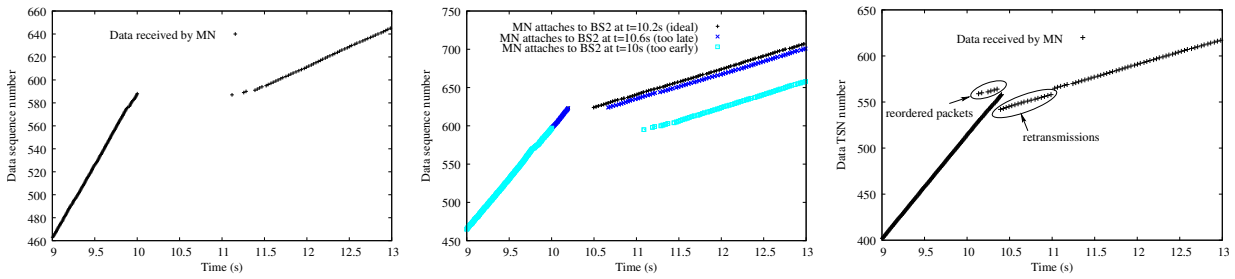
3) mSCTP: Mobile SCTP (mSCTP) [10] extends the multihoming feature of the Stream Control Transmission Protocol (SCTP) [14] to facilitate an IP handoff. SCTP's multihoming feature enables an endpoint to set up an association (the end-to-end connection in SCTP) that spans across its IP addresses: the primary address for packet transfer and other backup addresses. An SCTP sender keeps a separate congestion control variable for each path to the receiver's primary or backup addresses. Thus, data transmission rate of each path can be adjusted independently based on the path's network condition. When the receiver receives an out-of-order data, it returns a selective acknowledgment (SACK) with *Gap Ack blocks*, which reports to the sender the group of data that are missing. If and when the sender receives the third SACKs from the receiver, the missing data are then considered "lost", which are retransmitted during *fast recovery*. At this stage, the sender does not adjust the congestion window size even if an acknowledgment is received until fast recovery finishes [14].

mSCTP defines the handoff steps for a multihomed mobile node to communicate with a fixed host [10]. During handoff, the primary destination for data transfer is changed dynamically. When the mobile node enters a new IP subnet, it obtains a new IP address. It then communicates with the fixed host by exchanging mSCTP control messages, to first add the new address into the association, and then change the new address as primary. After these steps, data are sent to the new address of the mobile node.

3 Handoff Schemes Evaluation

Ns-2 [2] simulation experiments are used to evaluate the performance of the three handoff schemes. We construct the network topology as shown in Fig. 1. Background traffic, the Pareto ON/OFF UDP flows [8], is generated by the node BGSRC to two static destinations BGDST1 and BGDST2. Similar to [8], BGSRC sends packet of size 200 bytes in a constant bit rate of 24kbps during the ON times. The average ON and OFF times are set to 200ms and 100ms respectively. Data packets from BGSRC are sent to BGDST1 from t=5s to t=10s, and to BGDST2 from t=10s to t=15s.

To simulate the operations of Mobile IP and Fast Handoff, we patched the ns-2 package with the MobiWan IPv6 extension [1] and the Fast Handoff protocol based on [9]. To simulate the operations of mSCTP, we use the SCTP module included in the ns-2 package. MN is multihomed



(a) Mobile IP (b) Fast Handoff (c) mSCTP

Figure 2. Upward vertical handoff: data received by MN

while FH is single-homed. The "set-primary-destination" method in SCTP package is used to initiate handoff—when MN connects to another network, the new IP address is set as the primary destination. We consider both upward and downward vertical handoff of mobile receiver and mobile sender. In upward vertical handoff, reliable data transfer is investigated. For Mobile IP and Fast Handoff, FTP over TCP is chosen as the application and the transport protocol. For mSCTP, FTP/mSCTP is used. In downward vertical handoff, multimedia data transfer is investigated. In Mobile IP and Fast Handoff, the CBR (constant bit rate) over UDP is chosen. On the other hand, since mSCTP supports reliable data transfer, under the limitation of ns-2 we still choose FTP as the application, which is considered as multimedia data transferring over SCTP.

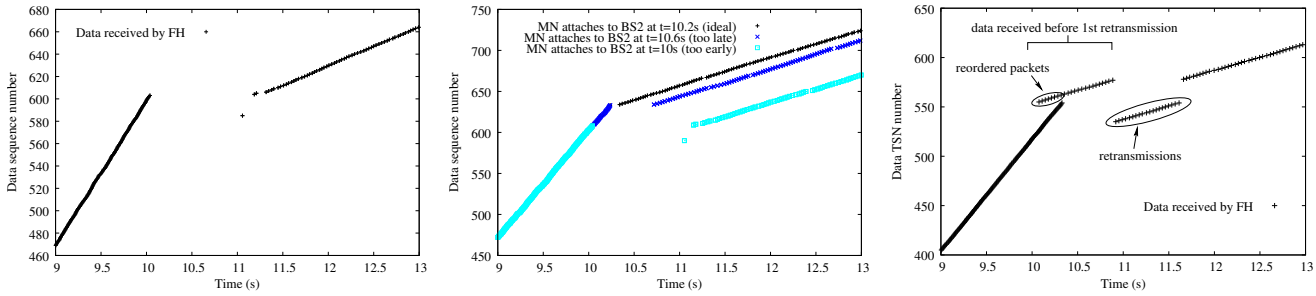
3.1 Reliable Data and Mobile Receiver

In MIP, there is a period during handoff which no data arrives at MN (a "blackout" period, as shown in Fig. 2(a)). MN switches connectivity to network-B and performs address configuration with HA at t=10s. Before the configuration completes, data sent to MN through BS1 are lost. The "blackout" period persists until MN receives the first retransmission (at t=11.1s), which is sent after the retransmission timeout (RTO) at FH expires. In Fast Handoff, three data sequence are plotted depending on the time when MN changes its connectivity to network-B (Fig. 2(b)). MN starts the fast handoff procedures at t=10s. Ideally, it should switch connectivity to BS2 *just after* BS1 starts packet forwarding. However, because of the "timing ambiguity" problem [13]—the uncertainty in the time that MN should attach to NAR (BS2), sub-optimal performance is obtained. If MN attaches to BS2 too early, fast handoff procedures have not completed; packets are not buffered in BS2 but lost at BS1. This results in a performance similar to that in MIP. In contrast, if MN attaches to BS2 too late, it would stay connected with BS1 for too long and during the period no packets are received as they have been forwarded to BS2. In the ideal case, a short "blackout" period still exists; packet forwarding to MN begins only after BS2 is notified, and the "blackout" period is the traveling time of the notification message.

Fig. 2(c) shows the handoff result using mSCTP. Unlike Mobile IP or Fast Handoff, there is not any "blackout" period as MN is multihomed; while the control messages are exchanged between MN and FH, data transfer can continue via the old network. However, the reordered packets and retransmissions undesirably affect the handoff performance—there is a time period that MN only receives the retransmission but not any new data. Packets are reordered due to the sudden decrease in their end-to-end delay when FH starts sending data via the new network path. As the bandwidth utilization of a network path varies according to the amount of traffic in the network, there is no guarantee packets traversing a network with a higher bandwidth would experience a smaller end-to-end delay. Therefore, as an example in the simulation, a decrease in packets' end-to-end delay during handoff is possible even if the available bandwidth of the old network is larger. Packet reordering causes unnecessary fast retransmission as a result.

3.2 Reliable Data and Mobile Sender

In MIP, acknowledgments replied to MN are lost at BS1 during address configuration (Fig. 3(a)). Thus, no data is sent by MN which results in a "blackout" period. In our simulation, all acknowledgments of data belonging to the sending window just before handoff have been lost. MN thus retransmits after RTO expires (at t=10.99s). FH receives the first retransmission at t=11.05s, and replies with an acknowledgment which cumulatively acknowledges all the received data. Data transfer then resumes via network-B. In Fast Handoff, the packet forwarding from BS1 to BS2, and buffering at BS2 avoids acknowledgments loss; BS2 can forward the buffered acknowledgments to MN after MN's handoff. Again, the handoff performance depends



(a) Mobile IP (b) Fast Handoff (c) mSCTP

Figure 3. Upward vertical handoff: data received by FH

Table 1. Handoff schemes comparison in upward vertical handoff

Handoff Scheme	Throughput (kB/s)	Reordered pkts
Mobile IP	74.57 / 76.32	1 / 1
Fast Handoff (early)	75.57 / 76.33	1 / 1
Fast Handoff (ideal)	80.60 / 82.27	0 / 0
Fast Handoff (late)	80.16 / 80.81	0 / 0
mSCTP	98.36 / 97.70	6 / 8

* Data shown as: mobile receiver handoff / mobile sender handoff

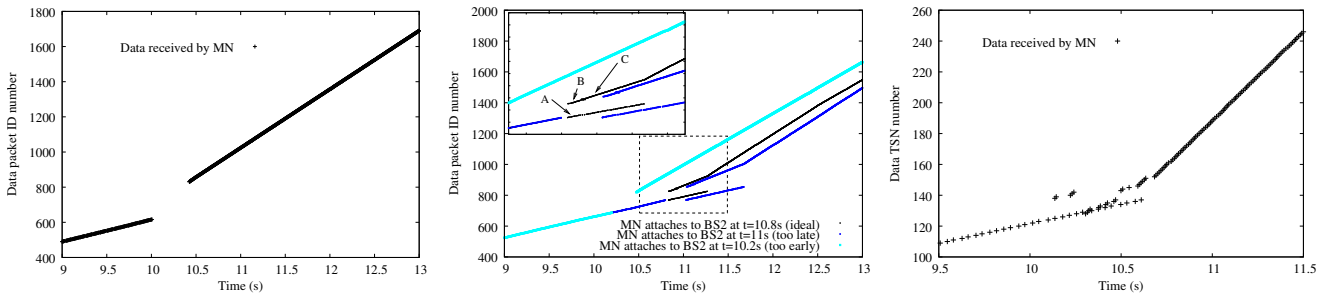
on the time when MN switches connectivity from BS1 to BS2. If it is carried out at an inappropriate time, a similar "blackout" period as in MIP is observed (Fig. 3(b)).

In mSCTP, the multihoming capability in MN avoids acknowledgments loss. Nevertheless, packet reordering and the ensuing fast retransmission affects its handoff performance (Fig. 3(c)). Compared with mobile receiver's handoff, here FH receives more data before the first retransmission. This is because MN sends more data via network-B before fast retransmission begins. It only keeps one congestion window variable for the single destination address in FH, despite the fact that two network paths exist for data transfer. Thus, after handoff, the data sending rate is based on the congestion window of the old path, which has been increased after some time of data transfer according to the condition of the old network. In contrast, in mobile receiver's handoff, FH sends data slowly according to the small congestion window variable for the new path. The transmission is independent of the traffic in the old network.

Next, we compare the data throughput achieved by various protocols during an upward vertical handoff (Table 1). The data throughput achieved by mSCTP is the highest because of the following two reasons. First, while the retransmission in mSCTP reduces the data throughput, the time spent in retransmission is shorter than the "blackout" period found in MIP or Fast Handoff. Second, SCTP adopts the delayed acknowledgment scheme which reduces the amount of acknowledgment traffic in the network. Table 1 also shows the number of reordered packets resulted by various handoff schemes. In our analysis, a reordered packet is one which causes the receiver to reply a duplicate acknowledgment (TCP), or a SACK with Gap Ack blocks (mSCTP), to the sender. In MIP or Fast Handoff with an early handoff time, one reordered packet is found, which is the first retransmission after the "blackout" period. No retransmission occurs in other Fast Handoff cases and hence no reordered packets is found. Packet reordering exists in mSCTP, and unnecessary fast retransmission occurs as a consequence.

3.3 Multimedia Data and Mobile Receiver

In MIP, data from 669 to 742 inclusive are forwarded to BS1 during address configuration (Fig. 4(a)). They are lost at BS1 and not resent, as UDP is unreliable. Nevertheless, the "blackout" period is shorter compared with that if TCP is used, as the sender does not wait for RTO expiry to retransmit the lost data. In Fast Handoff, if MN attaches to BS2 too late or too early, a long "blackout" period is observed (Fig. 4(b)). With an ideal handoff time, packets arrive at MN out-of-order just after MN changes its connectivity to BS2. This is because MN receives three sets of data at the same time: packets that were buffered at BS2 (set A), packets that are forwarded to BS2 from BS1 (set B), and packets that are sent directly from FH to BS2 without passing BS1 (set C). Data transfer is reliable when mSCTP is used (Fig. 4(c)), but packet reordering and fast retransmission occurs as in upward vertical handoff. It is interesting to note that although data are sent via the higher bandwidth network-B during fast retransmission, the receiving rate of these retransmissions at MN is limited and close to that of network-A's data transfer. Such phenomenon can be explained as follow. Since the bandwidth of network-A



(a) Mobile IP (b) Fast Handoff (c) mSCTP

Figure 4. Downward vertical handoff: data received by MN

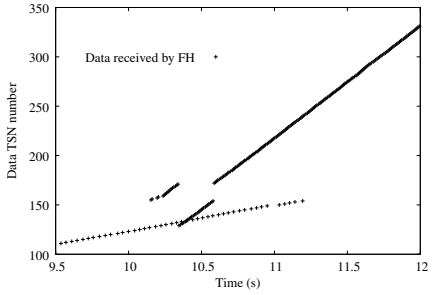


Figure 5. Downward vertical handoff: data received by FH under mSCTP

Table 2. Handoff schemes comparison in downward vertical handoff

Handoff Scheme	Handoff latency (ms)	No. of lost pkts
Mobile IP	421.12 / –	214 / –
Fast Handoff (early)	274.97 / –	132 / –
Fast Handoff (ideal)	38.37 / –	0 / –
Fast Handoff (late)	236.03 / –	0 / –
mSCTP	42.56 / 39.91	0 / 0

* Data shown as: mobile receiver handoff / mobile sender handoff

is smaller, just before handoff data is received by MN at a slower rate, so as the acknowledgments replied by MN. While FH can only retransmit one data after it receives each of these slow acknowledgments, the retransmission is thus carried out sluggishly. Moreover, the congestion window is not advanced and remains small during fast retransmission. These lead to a slow retransmission arrival rate at MN.

3.4 Multimedia Data and Mobile Sender

In UDP over MIP or Fast Handoff, no acknowledgments is replied to MN. In other words, MN does not receive any data nor acknowledgments, and the handoff operations are not necessary. For mSCTP, a downward vertical handoff is simulated, and packet reordering and retransmissions are observed (Fig. 5). Different from the mobile receiver's handoff, retransmissions are received at a faster rate. This is because MN sends data with a larger congestion window. It keeps one congestion window variable for both network paths, which has been increased based on the traffic condition of network-A. The large number of retransmission received after the out-of-order packets could introduce additional data buffering requirement by applications, in order to avoid the loss of the out-of-order packets during handoff.

Next, we compare the handoff protocols in terms of the amount of packet loss and the handoff latency resulted from a downward vertical handoff. The handoff latency is measured as the longest time for the receiver to wait for the next new data packet during the handoff process. As shown in Table 2, many packets are lost in MIP and in Fast Handoff with an early handoff time. The handoff latency is the shortest in mSCTP, which is almost equal to the inter-packet arrival time during normal data transfer using network-A.

4 Discussion

Based on our simulation, we come up with the following ideas concerning the design of an efficient handoff scheme:

1) Multihoming at mobile node facilitates handoff: Without the multihoming capability in the mobile node as in MIP or Fast Handoff, there would be a "blackout" period in the handoff process, during which data transfer is halted. Its duration can be minimized in Fast Handoff if the mobile node can switch connectivity to the new network at the correct time instant. Nevertheless, the exact handoff time is often difficult to determine in practice, which previous work reveal as the timing ambiguity problem [13]. Our simula-

tion further quantifies its adverse effect—a difference in the unit of 0.1 second in the handoff time would be sufficient to produce a reduction of 7.2% in data throughput (in the upward vertical handoff of a mobile sender). An approach to overcome the timing ambiguity problem is to allow multihoming at the mobile node. It can additionally "hide" the address configuration latency—while address configuration is done in the new network, data transfer can continue at the same time in the old network. Previous work suggest multihoming for handoff [6, 15], but we further quantify its advantage; it can avoid a "blackout" period of more than one second due to the TCP's RTO retransmission mechanism.

2) Packet reordering degrades handoff performance: Packet reordering is possible in a handoff scheme, which attempts to prevent the "blackout" period through simultaneous data transfer via the routes before and after handoff. In Fast Handoff, the effect of reordering could be more significant than that shown in the simulation. Although the mobile node changes its connectivity to another base station which is close to the original one geographically, it is possible that the two base stations are topologically far away from each other. Packets traversing the two routing paths would experience a large difference in the end-to-end delay and arrive out of order at the destination. TCP would then begin unnecessary fast retransmission. Packet reordering and unnecessary retransmissions also exist in mSCTP, due to the difference in network condition in the heterogeneous wireless networks. It also causes two implicit behaviours in downward vertical handoff: (1) an undesirable reduction in the transmission rate during retransmission in a mobile receiver's handoff; and (2) additional buffering requirement in the application, for preventing any loss of reordered data in a mobile sender's handoff. For better performance, the reordering should be avoided.

3) Transport layer should be aware of handoff: Transport layer reacts incorrectly in response to either the "blackout" period or packet reordering during the handoff process. For example, in MIP the TCP sender can resume data transmission only after the RTO of the last transmission expires whereas in mSCTP, packet reordering causes unnecessary retransmission after the receiver receives the third out-of-order data. As a result, the transport layer should be handoff-aware; it should recognise *when* do the handoff procedures start and finish, so as to control the data transfer during handoff. Flow control operations can then be carried out in-sync with the handoff operations. Hence, both handoff and flow control operations should be centralised in the transport layer, as in mSCTP. In contrast, if the handoff process involves participation from the network routers, as in MIP or Fast Handoff which the routers forward packets, it becomes uneasy for the transport layer at the end system to control the data transfer in response to the data forwarding.

5 Conclusion

We compare the performance of three handoff schemes namely Mobile IP, Fast Handoff and mSCTP. From the simulation results, we observe each of the schemes suffers from some deficiencies. The performance degradation is due to the incorrect response by the transport layer to either the "blackout" period or the reordered packets during the handoff process. The transport layer therefore needs to be aware of the mobile device's handoff and be responsible for the handoff operations. As the future work, we plan to further examine mSCTP handoff. We plan to design a protocol which can avoid the packet reordering problem, and to solve any further deficiencies that may exist in mSCTP handoff.

References

[1] MobiWan. http://www.inrialpes.fr/planete/pub/mobiwan.

[2] The Network Simulator - ns2. http://www.isi.edu/nsnam.

[3] A. C. Snoeren et al. An End-to-End Approach to Host Mobility. In *Proc. of the 6th Annual International Conference on Mobile Computing and Networking*, pages 155–166, Boston, MA, USA, August 2000.

[4] J. Davies. *Understanding IPv6*. Microsoft Press, 2003.

[5] H. Schulzrinne et al. Application-Layer Mobility Using SIP. *ACM SIGMOBILE Mobile Computing and Communications Review*, 4(3):47–57, July 2000.

[6] H. Y. Hsieh et al. An End-To-End Approach for Transparent Mobility Across Heterogeneous Wireless Networks. *Mobile Networks and Applications*, 9(4):363–378, August 2004.

[7] I. Aydin et al. Cellular SCTP: A Transport-Layer Approach to Internet Mobility. In *Proc. of the 12th International Conference on Computer Communications and Networks (ICCCN)*, pages 285–290, October 2003.

[8] K. Harfoush et al. Robust Identification of Shared Losses. In *Proc. of the 2000 International Conference on Network Protocols*, pages 22–33, November 2000.

[9] R. Koodli. Fast Handovers for Mobile IPv6. IETF Internet Draft, October 2003. Work in Progress.

[10] M. Riegel et al. Mobile SCTP. IETF Internet Draft, July 2005. Work in Progress.

[11] M. Stemm et al. Vertical Handoffs in Wireless Overlay Networks. *Mobile Networks and Applications*, 3(4):335–350, 1998.

[12] C. Perkins. IP Mobility Support. Request For Comments 2002, October 1996.

[13] R. Hsieh et al. S-MIP: A Seamless Handoff Architecture for Mobile IP. In *Proc. of IEEE Infocom 2003*, pages 1774–1884, March 2003.

[14] R. Stewart. Stream Control Transmission Protocol. Request For Comments 4960, September 2007.

[15] S. J. Koh et al. mSCTP for Soft Handover in Transport Layer. *IEEE Communications Letters*, 8(3):189–191, March 2004.

[16] S. Mohanty et al. Performance Analysis of Handoff Techniques Based on Mobile IP, TCP-Migrate, and SIP. *IEEE Transactions on Mobile Computing*, 6(7):731–747, July 2007.