Performance of TCP and UDP during Mobile IP Handoffs in Single-Agent Subnetworks

N. A. Fikouras, K. El Malki, S. R. Cvetkovic, C. Smythe
The University of Sheffield, Department of Computer Science,
211 Portobello Street, Sheffield, S1 4DP, UK

Abstract-In this paper we analyse the performance of Mobile IP (MIP) handoffs with respect to the three available MIP movement detection methods, namely Lazy Cell Switching, Prefix Matching and Eager Cell Switching. The network scenario considered involves single-agent subnetworks and Mobile Nodes that do not have the capability to contemporarily participate in multiple networks (e.g. IEEE802.11). This analysis will experimentally consider Transmission Control Protocol (TCP) communications over MIP, but the results can also be used to draw conclusions on the performance of User Datagram Protocol (UDP) communications. Since MNs are expected to perform multiple handoffs within the lifetime of a communication, it is important to analyse the impact of a handoff on the overall communication. It will be shown that the disruption caused by the MIP handoff may result in an unsatisfactory service for both TCP and UDP applications. It will be identified that the combination of the inherent characteristics of TCP and MIP handoffs may result in a service disruption interval of more than ten seconds. It will also be shown that UDP communications may suffer a service disruption interval up to 6 seconds.

I. INTRODUCTION

There have been significant technological advancements in recent years in the areas of laptop and notebook computers and Wireless LANs (WLANs), which have resulted in a movement towards total mobility. Users wanting to access their networks and services on-the-move required support for mobility. This functionality was not existent in the IP suite which was designed with the assumption that hosts are stationary with respect to the subnet they are connected to.

In TCP/IP (Transmission Control Protocol/Internet Protocol) and UDP/IP (User Datatgram Protocol/Internet Protocol) communications, the greatest resistance to mobility resides in the IP layer, which is responsible for data routing. IP assumes that a node's IP address uniquely identifies the point of attachment of the node to the Internet. Therefore, a node must be located in the network indicated by its IP address in order to receive datagrams destined to it. This IP prerequisite means that MNs with a tendency to vary their point of attachment to the Internet (i.e. between their home and other IP networks) will not be able to receive traffic directed to them.

In order to achieve this mobility function, the Internet Protocol (IP) was extended to become the Mobile Internet Protocol (MIP). MIP provides hosts with the ability to change their point of attachment to the network without compromising their ability to communicate. The mobility support provided by MIP is transparent to the other protocol layers so as not to affect the operation of applications which do not have the mobile capability. MIP introduces three new entities required to support the protocol: the Home Agent (HA), the Foreign Agent (FA) and the MN (MN). Further information on MIP functionality may be found in [1]. With the introduction of IPv6, MIPv6 excludes the FA, since FA functionality has been distributed amongst the MNs and Dynamic Host Configuration Protocol (DHCP) enabled hosts. Therefore we will often refer to the word “agent” (i.e. mobility agent) as a generalisation to identify HAs and FAs in MIPv4 and HAs in MIPv6.

Mobility in a wireless MIP network involves handoffs. A handoff is the process during which a node is “handed over” between two designated mobility agents. In terms of wireless local area networks (WLANs) this involves the movement of a MN between the areas of coverage of two wireless base stations (BS). At the MIP level, a handoff occurs when a MN switches its point of attachment to a new IP subnetwork and is required to register its new location. This will be described in more detail in later sections.

This paper experimentally analyses the performance of communications involving the Transport Control Protocol (TCP) over MIPv4 during handoffs. The results obtained are also applicable to the User Datagram Protocol (UDP). The efficiency of MIP handoffs is measured in terms of service disruption duration. This time duration is crucial for maintaining a level of quality during handoffs. By service disruption we mean the state in which the MN has moved but has not yet performed a MIP handoff and therefore is not able to send or receive traffic.

In our experimental tests, the wireless medium is replaced by Ethernet. This is because the WLAN IEEE802.11 standard [2] is used as a reference, where a handoff is considered as being a minor disruption. In fact this should be true for most WLANs. For example the Lucent WaveLAN IEEE802.11 product takes approximately 10 ms to perform a wireless handoff [3]. Therefore it is assumed that high-level mobility and its interaction with the inherent characteristics of TCP form the greatest part of the service disruption period.

II. MOBILITY

Fig. 1 illustrates a single-agent subnetwork scenario. This is

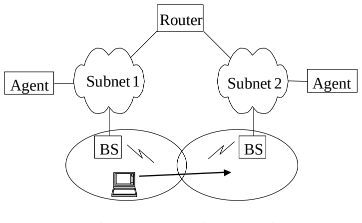

Fig. 1. Single-Agent Subnetwork Scenario

0-7803-5669-1/$10.00 (c) 1998 IEEE

typical in networks where each group of computers (e.g. department) is assigned a subnetwork and where a subnetwork prefix will identify all its nodes. It should be noted that only one agent exists per subnetwork and that the base stations (BSs) are WLAN bridges.

The performance of MIP handoffs depends on the interaction between the MIP and the wireless access protocol. By wireless access protocol we identify layer 2 of the OSI 7-layer model, which defines the Medium Access Control (MAC) behaviour of the wireless station. It is therefore necessary to distinguish between low-level and high-level mobility.

In the field of wireless communications, mobility is usually understood as low-level mobility, that is the movement of a MN between base stations (BSs) and the relative wireless hand-off mechanism which occurs as a result. The term low-level mobility denotes node mobility that is handled by the MAC and its support for roaming. Its aim is to establish and maintain reliable link connectivity despite node movement and the lossy nature of the wireless medium. With respect to Fig. 1, low-level mobility would involve the MN leaving the region of coverage of the BS in subnet 1, entering the region of coverage of the BS in subnet 2 and registering with it at link layer level. However, when considering a MIP network it is necessary to define another type of mobility, namely high-level mobility. High-level mobility occurs at the IP (i.e. MIP) level. This involves the MIP movement detection methods which enable a MN to take decisions on whether it should switch between agents. This is not normally dependent on low-level mobility, although low-level mobility must precede high-level mobility. With reference to Fig. 1, high-level mobility would take place after the MN has exited subnet 1 and joined subnet 2. It would include detecting movement and re-registering its current position with the HA. The HA is just an agent with which the MN shares a security association and may authenticate the MN during registration.

Throughout this paper we will be considering a wireless access technology which behaves like the IEEE 802.11 standard for WLANs will be considered. The most important characteristics of this standard, with respect to the research in this paper, is that MNs that are in the coverage areas of two WLANs will not be able to contemporarily participate in both networks. It is clear that in order for high-level mobility to take place, low-level mobility must have completed. Therefore we have introduced two underlying assumptions. The first assumption is that only one agent is required per subnetwork. Although this is a common scenario, it is not always applicable, and further study is underway to identify the performance issues involved in mobility for multiple-agent subnetworks. However, for the purpose of research into typical TCP/UDP performance issues over MIP, we will apply this assumption.

The second assumption is that the MN may not participate simultaneously in multiple networks. This is dependent upon the capabilities of the low-level mobility mechanisms.

Considering the network scenario presented in Fig. 1, this means that the MN will not be able to receive advertisements contemporarily from the agents in subnet 1 and subnet 2, when between the area of coverage of the two Base Stations (i.e. WLAN bridges). This is typical of current WLANs. This point is of great importance for a particular MIP movement detection method, which uses the receipt of agent advertisements as an indication of movement.

A final assumption is that the duration of low level mobility is insignificant compared to that of high level mobility. This is because low level mobility requires at most tens of milliseconds, while high level mobility requires a number of seconds. Therefore we define the service disruption interval as the duration of high level mobility and more specifically the sum of the MIP movement detection and registration intervals.

III. MIP MOVEMENT DETECTION METHODS

A. *Lazy Cell Switching (LCS)*

This method is also referred to as Algorithm 1[1]. It uses agent advertisement lifetime expiration as an indication of MN movement. Under normal conditions, agent lifetime expiration takes place only if the node has left the agent's area of coverage and the boundaries of the subnet. For the duration of the advertisement lifetime, the node is forced to experience disruption in service. Meanwhile, incoming advertisements from other mobility agents are ignored. After the advertisement expiration, the MN may attempt registration with a known mobility agent or solicit for the discovery of other agents.

An advantage of this method is that it avoids handoffs between mobility agents in the same subnet. Assuming that broadcast traffic reaches all nodes in the same subnet it is highly unlikely that a MN will loose contact with its current agent for the duration of the advertisement lifetime.

B. *Pattern Matching (PM)*

This method is presented in [1] as Algorithm 2. In this method the MN compares the subnet prefixes of mobility agents in order to determine new agents and eliminate agents within the same subnet. Normally, agent advertisements do not contain information pertaining to the agent's subnet prefix size. Therefore in order for a MN to use this movement detection scheme, all agents are required to include the prefix-length extension in their advertisements. This extension contains the prefix length of the agent's address. If this method indicates that the MN has moved, rather than re-registering with the old agent, a MN will instead register with a newly discovered agent after the lifetime expiration of the node's current binding

This method is useful in occasions where the node can receive advertisements from multiple agents located either in the same or in different subnets. However, given the underlying assumptions of single agent subnets and inability of the mobile node's subnetwork technology to participate in multiple networks simultaneously, the advantage of this method is in effect cancelled. As will be presented by the corresponding results, the LCS and PM mechanisms tend to behave similarly in the aforementioned conditions.

C. *Eager Cell Switching (ECS)*

This movement detection method assumes that MNs tend to change their direction of movement very slowly. That is, if they are moving forward in one direction, it is unlikely that they will stop and turn back. Hence, it is appropriate for nodes to handoff immediately upon encountering a new agent. As we will experimentally demonstrate in a single-agent subnet, its eagerness to register with every newly encountered agent allows the completion of handoffs faster than the other two methods.

IV. EXPERIMENTAL SETUP

In this paper we have experimentally tested the performance of MIP high-level mobility in a single-agent subnetwork scenario, using the three movement detection methods: LCS, PM and ECS. We aim to identify the impact of MIP handoffs in terms of service disruption time, which is considered to be far greater than that caused by low-level mobility.

In our tests we have used the Sunlabs MIP [4] implementation for Redhat 5.x Linux. This implementation is fully compliant with [1] and [5] and implements all three movement detection methods.

For the tests, traffic has been generated using *ttcp* benchmark [6]. In every experiment we have set the rate of agent advertisement transmissions to the recommended value of one per second [1]. As a consequence, the agent advertisement lifetime, whose recommended value is three times the agent advertisement transmission rate, is set to 3 seconds. This means that a MN will delete an agent from its list of valid agents after missing three successive advertisements.

The testbed used in our experiments is presented in Fig. 2. It is a flexible and efficient topology which can represent a variety of typical LAN scenarios. The computers are Pentium intel machnines that are running Redhat 5.2 Linux. Connectivity between the PC's is 10Mbps Ethernet.

V. TEST RESULTS

In the following sections we are going to present the three test cases and results using the testbed shown in Fig. 2.

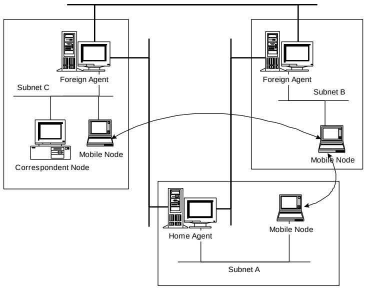

Fig. 2. Testbed Setup

Experimentally we will focus on handoffs between subnets A and B (i.e. HA to FA) in Fig. 2. Given the underlying assumptions stated in the section 2, all combinations of handoffs between networks will produce the same results. That is, handoffs between subnets A and B (i.e. HA to FA) will represent handoffs between subnets A and C (i.e. HA to FA) and subnets B and C (i.e. FA to FA).

A. *Lazy Cell Switching Test*

For this experiment the agent advertisement rate is 1 per second and the agent advertisement lifetime is 3 seconds. The MN is originally located in its home network (HN), subnet A. The Correspondent Node (CN) in subnet C initiates a *ttcp* transfer, which involves a TCP connection to the MN for the transfer of 100 Mbyte. The MN then enters subnet B. In a wireless network this would involve low-level mobility. Only after the LCS method has acquired positive indication of movement (i.e. the expiration of agent advertisement lifetime) can the MIP protocol complete the handoff. It is noted that, for this method, the receipt of a new agent's advertisements is not considered as an indication of movement.

High-level mobility causes a period of service description. This interval affects to TCP traffic due to Exponential backoff of the TCP timeout timers [7] and HA Tunnel MTU (Maximum Transmission Unit) discovery [5].

When a timeout occurs, TCP resorts to exponential backoff. This means that TCP doubles the size of its timeout interval with each successive timeout. This mechanism can be seen in Fig. 3.

In Fig. 3 it is shown that, during MIP handoff, TCP goes through 4 successive timeouts and retransmissions. The packet with sequence number 6615260 lies in the CN's TCP retransmission list and is transmitted at exponentially increasing intervals. At time 12.51 seconds, the MIP handoff has completed. However, packet 6615260, which is again retransmitted after that time, fails to go through to the MN. Hence the CN's TCP layer is forced to suffer another timeout, the biggest in this series. The latest packet rejection has nothing to do with service disruption imposed by the MIP handoff but is

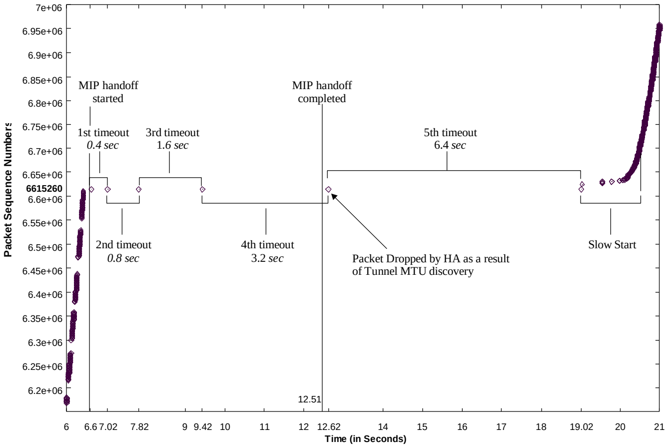

Fig. 3. TCP Packet Trace of MIP handoff with LCS

caused as a result of the HAs Tunnel MTU discovery [5]. The purpose of this mechanism is to prevent packets, tunnelled by the HA to the MN's current location, from getting fragmented within the tunnel. Therefore, oversized incoming packets at the HA, having the DF (Don't Fragment) bit set, are dropped and an ICMP "Datagram Too Big" message is returned to the packet's source indicating the tunnel's transmission unit maximum size. However, all tunnelling schemes increase the packet size by at least a minimum amount during the encapsulation process. The IPinIP [5] encapsulation scheme, used in these experiments, increases the packet size by 20 octets, (i.e. the size of an IP header without any options). In Fig. 3, the fifth TCP timeout is caused by the first encapsulated packet which is dropped.

Following the MIP handoff interval is the slow start period ([8],[9]), which takes effect after the final retransmission in Fig. 3. Slow start is a congestion avoidance mechanism that is initiated by a TCP timeout. This mechanism reduces the communication's usable window to as little as the size of a single full packet. Based on the receipt of acknowledgements it then exponentially increases it. The purpose of this mechanism is to detect the current congestion state of the transmission pipe that connects the source with the destination and avoid exceeding it. The slow start period is visible in Fig. 4, after the final TCP retransmission. It can be seen that the packet sequence numbers starts increasing at time zero, level off due to the handoff and then increase again.

It can be seen from Fig. 3 that the total service disruption period lasted 12.42 seconds in addition to slow-start, of which 5.91 seconds were due to high-level mobility and 6.51 seconds were due to TCP characteristics. It can be concluded that a UDP communication would experience a total of 5.91 seconds of service disruption due to the handoff.

B. *Prefix Matching Test*

For this experiment the testbed agents have been configured to include the prefix extension in their advertisements. The agent advertisement rate is 1 per second and the agent advertisement lifetime is 3 seconds. Again the *ttcp* application on the Correspondent Node (CN) will initiate a TCP connection for the transfer of 100 Mbytes of data to the MN.

The MN is initially located in subnet A. MN movement will lead the node into subnet B. It is assumed that low-level mobility has taken place to maintain link connectivity. High-level mobility will be in turn initiated according to the movement detection scheme. In the absence of multiple network participation capability, Prefix Matching and LCS operate in a similar fashion. They both rely on the expiration of their old agent's advertisement lifetime, in this case the HA, in order to register with a new agent.

The similarities between Prefix Matching and LCS are not confined to their high-level mobility handoff intervals, but extend to their impact on transport layer communications. For UDP traffic, the service disruption will cause the loss of all the packets transmitted during that interval. However, beyond that point communications should commence without any further drawbacks. The biggest impact of the MIP handoff interval is imposed on TCP connections.

Fig. 4 illustrates the packet trace of the Prefix Matching trial. In this experiment the MIP handoff is initiated while the transmission of a whole window of packets is in progress. Although all of these packets will reach the HA, only a few will be delivered to the MN. The first packet that is caught in the service disruption is the packet with sequence number 4789260. In Fig. 4 we can see that the particular packet is retransmitted 6 times after a corresponding number of successive timeouts. TCP exponential backoff doubles the duration of the timeout interval after every timeout. It can be seen from Fig. 4 that the final timeout interval lasted for 6.4 seconds.

The MIP handoff is completed at time 10.13-seconds, only 6.1 milliseconds before the fifth timeout and packet retransmission. Again, this retransmission will fail to reach the node. This is due to the Tunnel MTU discovery mechanism of the HA. After the sixth retransmission the packet is delivered and the TCP communication enters the slow start period. Hence, in total, the TCP communication has suffered a 12.5 second interval of service disruption before the slow start period, of which 5.63 seconds are due to the handoff and 7.02 seconds are due to TCP characteristics. It is therefore possible to conclude that a UDP connection would experience a total of 5.63 seconds of service disruption due to handoff.

C. *Eager Cell Switching Test*

For this experiment the MN is initially placed in the foreign network (subnet B). The MN is already registered with the HA and maintains a binding for its current location. Incoming traffic from the CN, destined for the MN, is intercepted by the HA and tunnelled to the FA of subnet B. It is the FA's responsibility to decapsulate and deliver the traffic to the MN. At the CN, a TCP connection is initiated to the MN using the *ttcp* benchmarking software for the transfer of 100Mbytes of data. At some point MN mobility will bring the node into subnet A. It is noted that, with this movement detection method, the discovery of any new agent will cause an

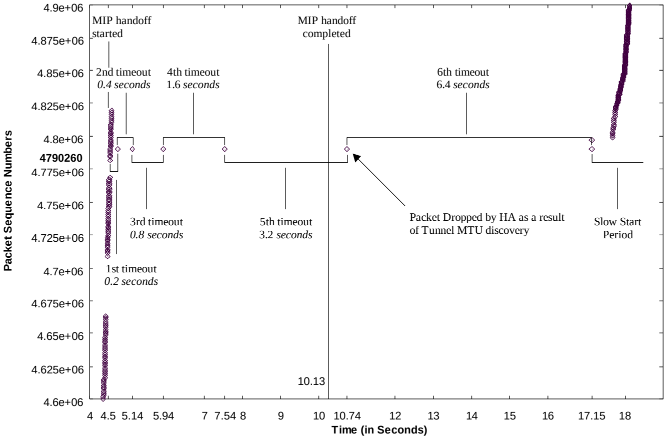

Fig. 4. TCP Packet Trace of MIP handoff with Prefix Matching

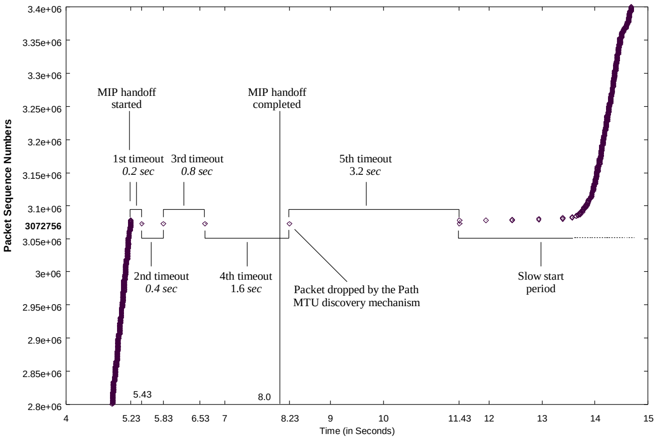

Fig. 5. TCP Packet Trace of MIP handoff with ECS

immediate high-level handoff.

For UDP traffic the MIP handoff interval simply means a service disruption equivalent in length to the handoff interval. It is on TCP traffic that the MIP handoff interval has the greatest impact. Fig. 5 illustrates the packet trace of the TCP communication. The interval during which only retransmissions of the same packet are observed (i.e. where the curve is flat) is due to the service disruption caused by the MIP handoff, TCP exponential backoff and Path MTU discovery [10].

It can be seen from Fig. 5 that the total service disruption period lasted 7.77 seconds in addition to slow-start, of which 2.77 seconds were due to high-level mobility and 5 seconds were due to TCP mechanisms. It is therefore possible to conclude that a UDP connection would experience a total of 2.77 seconds of service disruption due to the MIP handoff. This improvement in performance is observed due to the eagerness of ECS to handoff to every agent that is newly encountered without suffering further delay. This same advantage of ECS that enhances its performance can also prove to be a disadvantage because it will force a node that is entering a new subnet with multiple agents to handoff to every agent in the subnet prior to reaching a stable state. This is only the case in multiple-agent subnetworks or those capable of multiple network participation.

VI. CONCLUSIONS

In this paper we have analysed the MIP handoff in single-agent subnetworks without multiple network participation capability. We have presented three methods of movement detection, introduced the notion of high and low-level mobility and demonstrated how their interaction can speed up or delay the completion of MIP handoffs. We have focused on the disruption period caused by high-level mobility which has been shown to be higher than that introduced by low-level mobility. Overall the results indicate that no movement detection method can offer a MIP handoff without first suffering some period of service disruption. For the TCP protocol, service disruption is perceived as an indication of congestion that requires TCP exponential backoff and slow-start. Successive TCP timeouts will increase the TCP timeout interval so much that, even after the MIP handoff has completed, TCP will not be in the position to immediately resume the communication. Finally the interaction of the Path/Tunnel MTU discovery mechanism and TCP slow-start have been shown to cause further TCP timeouts and further delay.

The results indicate that Eager Cell Switching (ECS) provides for the most efficient movement detection method with respect to the performance of TCP and UDP traffic. Prefix Matching has intermediate performance and Lazy Cell Switching (LCS) has the worst performance of the three methods. However, it should be noted that we have only considered single-agent subnetworks with MNs which are not capable of participating contemporarily in multiple networks.

The overall conclusion of this paper is that the performance of current movement detection mechanisms, and consequently of MIP handoffs, is inappropriate for most applications and especially those which are time-critical. By this we mean that a web or file transfer application (over TCP) will suffer much disruption during a handoff, potentially more than 10 seconds. Also, an Internet Telephony application running over UDP would suffer a disruption period between 3 (ECS) and 6 (LCS) seconds with every handoff. On top of these intervals we will have to add a further amount due to low-level mobility, although this will be comparatively small. This is unacceptable since current mobile telephony technologies (e.g. GSM) allow handoff disruption intervals in the range of tens of milliseconds. In addition to these considerations, it is not unreasonable to expect multiple handoffs within the disruption interval. This will cause even greater disruption and the delay may cause the application to terminate the connection.

REFERENCES

[1] Perkins E. Charles, 'IP Mobility Support', RFC 2002, October 1996.

[2] IEEE Std 802.11-1997. "Wireless LAN Medium Access control (MAC) and Physical Layer (PH) specifications"

[3] Caceres, R. and Padmanabhan V. N., "Fast and Scalable Wireless Handoffs in Support of Mobile Internet Audio", ACM MONET Journal, Vol. 3, No. 4, December 1998

[4] Sunlabs MobileIP,ftp://playground.sun.com/pub/mobile-ip

[5] Perkins E. Charles, 'IP Encapsulation within IP', RFC 2003, October 1996

[6] "ttcp: Test TCP" U. S. Army Ballistics Research Lab, December 1984

[7] Stevens W.Richard, "TCP/IP Illustrated Volume 1". Reading: Addison Wesley Professional Computing Series.

[8] Jacobson Van, Karels J. Michael (1988). "Congestion Avoidance and Control". pp:8., http://www-nrg.ee.lbl.gov/papers/congavoid.pdf

[9] Stevens W. Richard. "TCP Slow Start, Congestion Avoidance, Fast Recovery Algorithms', RFC 2001, January 1997.

[10] Mogul J.C., Deering S.E., 'Path MTU Discovery', RFC 1191, November 1990.