Mobile Networks and Applications 6, 67–79, 2001
© 2001 Kluwer Academic Publishers. Manufactured in The Netherlands.

Intelligent Handoff for Mobile Wireless Internet *

JON CHIUNG-SHIEN WU **,***
Philips Research East Asia – Taipei 24FA, 66, Sec. 1, Chung Hsiao W. Rd., P.O. Box 22978, Taipei 100, Taiwan, R.O.C.
E-mail: jon.cs.wu@philips.com

CHIEH-WEN CHENG and NEN-FU HUANG
Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan, R.O.C.

GIN-KOU MA
Electronics Research & Service Organization, Industrial Technology Research Institute, Chutung, Hsinchu, Taiwan, 310 R.O.C.

Abstract. This paper presents an intelligent mobility management scheme for Mobile Wireless InterNet – MWIN. MWIN is a wireless service networks wherein its core network consisting of Internet routers and its access network can be built from any Internet-capable radio network. Two major standards are currently available for MWIN, i.e., the mobile IP and wireless LAN. Mobile IP solves address mobility problem with the Internet protocol while wireless LAN provides a wireless Internet access in the local area. However, both schemes solve problems independently at different layers, thereby some additional problems occur, e.g., delayed handoff, packet loss, and inefficient routing. This paper identifies these new problems and performs analyses and some real measurements on the handoff within MWIN. Then, a new handoff architecture that extends the features of both mobile IP and wireless LAN handoff mechanism was proposed. This new architecture consists of mobile IP extensions and a modified wireless LAN handoff algorithm. The effect of this enhancement provides a linkage between different layers for preventing packet loss and reducing handoff latency. Finally, some optimization issues regarding network planning and routing are addressed.

Keywords: mobile Internet, wireless Internet, wireless data networks

1. Introduction

Following the rapidly expanding markets of cellular phone services, mobile high-speed data communications are now becoming the next candidate of new targeting business. Mobile Internet services will be the major sources of traffic in the future. Basically there are two points of view on the infrastructure for supporting mobile wireless services over the Internet. The first one is called *Internet via mobile cellular dial-up*. This comes from cellular telecommunication concept and is currently available, for example, GSM (Global System for Mobile communications), IS-54, and IS-95 [9,10,25]. They use circuit-switched cellular phone networks as their infrastructure and relay the Internet packet to/from the Internet gateway. Figure 1 shows a typical example using GSM. The radio access network is controlled by a core network consisting of non-Internet devices such as Base Station Controller (BSC), Mobile Switching Center (MSC), Gateway Mobile Switching Center (GMSC), and Public Telephone Switched Networks (PTSN). The core network probably uses protocols other than TCP/IP for the transport of voice calls. To support mobile Internet services, the core network is attached onto the Internet via one or more point of attachments. Users need to dial up to the core network first, and then obtain the Internet service later via point-to-point protocol (PPP) and the TCP/IP protocol suite. One advantage of this architecture is that mobility management can be independently achieved by the core network and no other control mechanism is necessary. However, the limita-

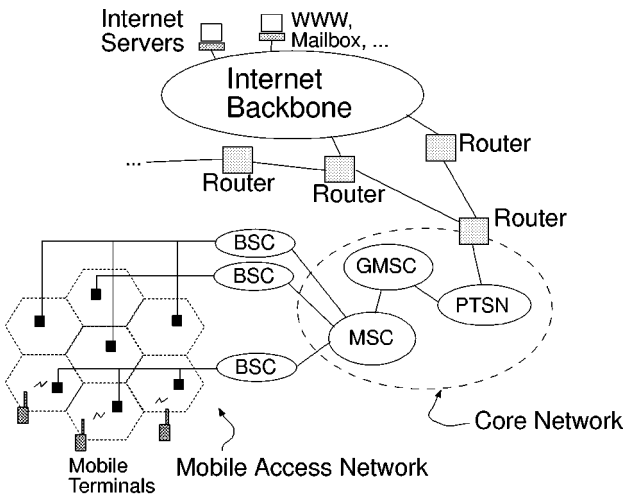

Figure 1. Internet via mobile cellular networks.

* Supported by the R.O.C. Ministry of Economic Affairs under the project No. 3P12200 conducted by ITRI.
** This work was done while first author was with Computer and Communications Research Labs., ITRI, Taiwan.
*** Corresponding author.

tion is that most of the communication sessions are mobile initiated due to the temporary IP address assignment. Furthermore, the bit rate of this type of networks is usually very low, around tens of Kbps, and the bandwidth efficiency is not high.

Besides circuit-switched mobile data services, there are packet-switched mobile data networks on the shelf, i.e., CDPD (Cellular Digital Packet Data), ARDIS (Advanced Radio Data Information Services), Mobitex Packet Radio Data [26,27,30]. CDPD is built upon the existing US mobile phone standard AMPS (Advanced Mobile Phone Systems) and is using TCP/IP protocol suite. Thus, it can be recognized as the first mobile Internet system as its pure Internet-compliant feature. ARDIS and Mobitex are also packet-switched mobile data networks, however, they use proprietary protocols other than TCP/IP. To provide Internet services, protocol encapsulation or tunneling can be used. Basically, all these mobile data networks are macro-cellular based and work at low data rate in licensed RF band.

The focus of this paper is also on the packet-switched mobile data networks, however, with more advanced features required, i.e., high data rate, micro/pico-cellular, TCP/IP compliant, unlicensed RF band. We refer to the target network as *Mobile Wireless InterNet (MWIN)*. MWIN is required to be a promising architecture for future mobile multimedia services. Figure 2 shows a straightforward approach of MWIN wherein the Internet backbone (consisting of routers) serves as the core network. This direct approach gives mobile users an Internet-friendly environment, for example, to support wireless Internet services in a campus, organization, enterprises, and the residential areas. For the radio access part, there are many radio LANs suitable for this scenario, i.e., IEEE 802.11 wireless LAN [14], wireless ATM [1]. Most of these wireless access technologies work in low-power, unlicensed band. Currently, wireless LAN is a standardized product for indoor use. In the future, outdoor broadband wireless solutions that aim at mobile multimedia will be available, i.e., Wideband CDMA [11], cdma2000 [18].

Although MWIN is attractive, it is incomplete and requires more enhancements and optimization. For example, the TCP connection in MWIN extends across both wired and wireless segments. Due to different characteristics at different link layers, TCP will adapt to the link with poor performance and result in a throughput degradation of the whole connection. Several proposals have been published by other researchers to improve TCP connection performance in a mobile wireless network [5]. The split connection approach called Indirect TCP was proposed by splitting the TCP connection into two segments such that the data transport at radio part can be optimized independently [3,4]. Another approach called fast retransmit scheme was invented to solve the handoff delay problem by having the mobile host send a certain threshold number of duplicate acknowledgements to the sender [7]. This causes TCP at the sender to immediately reduce its window size and retransmit packets quickly. In [2], a link layer retransmission scheme was proposed to improve the reliability of TCP connection. In [6], a caching approach was proposed to improve the TCP connection performance by keeping a packet cache at the base station such that the retransmission can be performed very quickly. Basically, the unreliability of TCP connection in a mobile wireless network partially comes from the handoff delay. In [8], a hierarchical mobility management scheme was proposed to improve the handoff latency, thereby improving the TCP performance at the upper layer. However, all the above presented approaches solve the problems independently at one protocol layer and there is no proposal yet to consider the cooperation of multiple protocol layers.

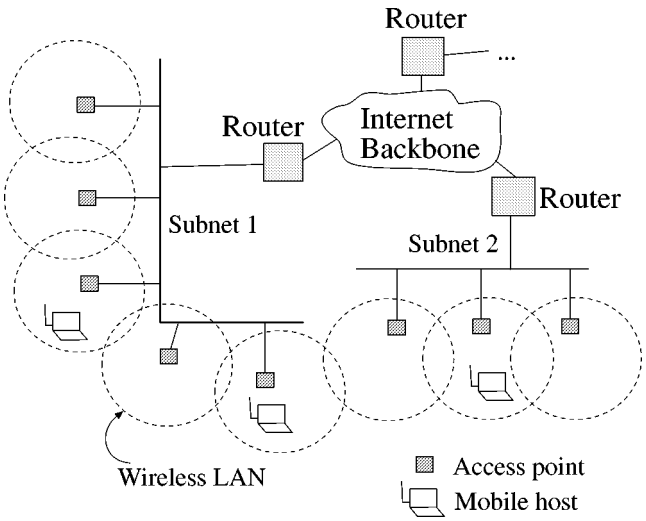

Figure 2. The mobile wireless Internet (MWIN).

Mobility management in MWIN is an importance issue and is the main focus of this paper. Two major tasks regarding mobility in MWIN are the roaming at layer 3 and the handoff at layer 2. Handoff at layer 2 is related to the change to a new radio LAN while the mobile station moves across the radio boundary. Typical handoff algorithms are based on the measurement of radio signal strength and data error rate. Roaming is related to the IP address manipulation and routing as the mobile station moves into a new Internet sub-network. A proposal called *Mobile IP* for current IP version 4 has been approved by IETF [21]. The major idea is simply the use of temporary IP address called care-of-address or a proxy-based routing agent when the mobile host has moved away from its original subnetwork. Recently, the mobility support on IP version 6 is also under consideration [17,29].

Problems with MWIN arise even at the availability of mobile IP and wireless LAN. One of them is that mobile IP and wireless LAN solve their problems independently at different layers, that is, mobile IP works on layer 3 and does not talk to layer 2, and vice versa. Another drawback is that there is no standard or guidelines on cell planning for Internet in the wireless environment. Cell planning is essential on cellular phone networks such as GSM. With cell planning, each base station is given a neighbor lists consisting of the network candidates to which a mobile handset can be handed

over. Without such planning, a mobile host does not know where is the neighbor and which radio channel to tune into at lower layers. It will waste a lot of time in searching a right channel where a suitable neighbor proxy agent is located.

In this paper, problems with MWIN are identified and analyzed. In the next section, an overview on mobile IP will be presented. Problems with MWIN are described in section 3. In section 4, we verify the problems stated in section 3 by analysis and real traffic measurement. In section 5, a new handoff architecture that extends the features of both mobile IP and wireless LAN handoff mechanism was proposed. This new architecture consists of mobile IP extensions and a modified wireless LAN handoff algorithm. The effect of this enhancement provides the linkage and coordination between different layers for preventing packet loss and reducing handoff latency. Finally, optimization issues regarding network planning and routing are addressed.

2. Overview of mobile IP

Mobile IP was proposed to support Internet host mobility, and thus, it is very useful in MWIN's roaming at layer 3. Two versions of mobile IP have been proposed, one for current IP version 4 and one for future IP version 6 [17,21]. Basically, the schemes of both versions are very similar. For simplicity, we consider the case for IP version 4. Throughout this text, we will use the term network to refer to an IP subnet. According to mobile IP standard, any IP packet addressed to this mobile host should be re-routed to the new network that is currently being visited [23,28]. Detailed mechanisms are described in the following subsections.

2.1. Basic mobile IPv4

The basic concept of mobile IP defined in RFC2002 is that the *mobile host (MH)* has a permanent IP address called *home address*. When it enters its home network, it will register itself to its *home agent (HA)* that is also located in the home network. When MH moves from home network to another foreign network, it will detect an available mobility agent called *foreign agent (FA)* and attempt to register to HA via it. The FA will notify HA that the MH has moved and all packets addressed to this MH should therefore be forwarded to this foreign network via a *care-of-address* which is either the address of FA or a temporary address colocated in MH. HA is responsible for forwarding packets addressed to the MH. The forwarded packets will be re-encapsulated with an additional IP header that contains the care-of-address as the destination [22]. This is referred to as *IP tunneling*, as explained in figure 3. The tunnel is formed between HA and FA or the MH itself.

2.2. Move detection

Move detection is a mechanism that a mobile host use to detect that it has moved away from its home network and now

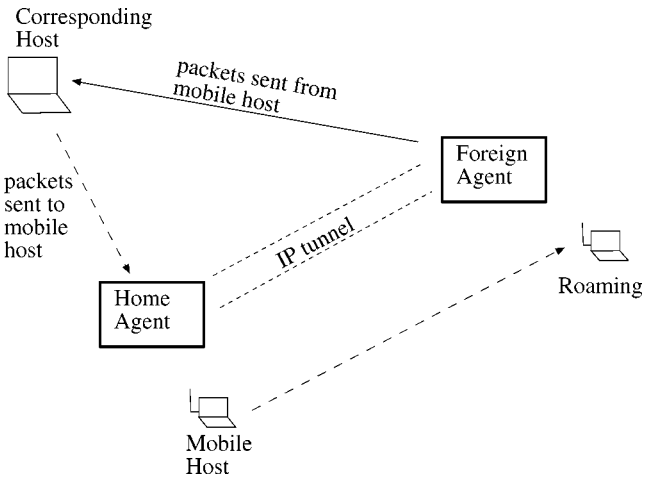

Figure 3. The mobile IP tunneling.

is in a foreign network. Problems with this technique are the major focus of this paper. The move detection mechanism is tightly coupled with another mechanism in mobile IP called *agent discovery*. A mobility agent can be discovered using only layer 3 information or via the aid of a link-layer protocol. Mobile IP allows both schemes but only the mechanism using layer 3 is included in its specification.

The agent discovery and move detection schemes using only layer 3 information are described as follows. Let us assume that a MH is now located at home network. First, this MH should be able to determine at any time that the home network is still reachable. If the home network is unreachable, it assumes that it has moved away from the home network. Then, this MH must determine the existence of a foreign agent. The following three mechanisms can be used.

Agent advertisement. A mobility agent will periodically broadcast ICMP *advertisement* message in the subnet. Each advertisement message carries a life time. For simplicity, the timer is called *Agent Advertisement Renew (AAR)* timer. If, within the life time of an advertisement message, another advertisement is received by the mobile host, then this mobility agent can be confirmed to be alive and reachable. Otherwise, when the AAR timer expires, this agent is assumed to be unreachable from this MH. At this moment, the MH should start to search a new network with a mobility agent and attach to it. In [21] and in most of the mobile IP implementation, AAR timer is set to be three times of the interval in which an advertisement is sent by an agent.

Network prefix. According to [21], the prefix length extension within a agent advertisement message can be used with mobility agent's IP address in order to identify a subnet. This is especially useful when the MH is moving around two networks that have the same network ID but different subnet IDs. The network prefix length is carried within the advertisement to indicate the subnet mask. By calculating the subnet prefix whenever an agent advertisement is received, a move can be found if the subnet ID is different. Thus, it is possible to use this field to detect a move. However,

this scheme cannot be used sometimes in a wireless environment. If a MH within an IP subnet is able to receive different agent advertisements all the time, e.g., when the mobile host is located in a wireless environment with two or more overlapping radio signals, it is not appropriate to make an immediate decision to switch into another network. This will result in the so-called "ping pong" effect which makes a MH switches between two networks. To avoid this oscillation, the decision to attach to a new network is primarily based on the agent advertisement lifetime as mentioned previously.

Router solicitation. A mobile host can also send agent solicitation message in the subnet to find a new agent. If there is any agent reachable, it will reply. However, it is not possible for a mobile host to periodically send solicitation messages to an agent due to limited wireless resources. This is often used to find a new router when one of the above two conditions has happened or is triggered by the information from the link layer.

In MWIN, HA should be able to determine at any time whether the MH has moved away from its current visiting network or not. In Mobile IP, the only mechanism that achieves this is when a new agent carrying a new registration message is communicating to HA. Usually, this will take a significant amount of time for the MH to notify the HA (via the FA) that the MH has in fact moved.

The above-presented architecture of mobile IP is a general overview based on RFC 2002 [21]. However, it requires more extension and enhancement to make MWIN function optimally. Fortunately, mobile IP does not limit itself to work with any extension, i.e., to cooperate with link layer protocols. Before introducing our proposal, some problems with MWIN are addressed in the next section.

3. Some problems with mobile wireless networks

In this section, we will state the potential problems with MWIN using the current mobile IPv4. Most of the problems presented here are essentially due to the independency of layer 2 and layer 3, i.e., no information is exchanged across layer 2 and layer 3. In some cellular mobile phone networks, the handoff process is accomplished by the cooperation of two or more layers, e.g., signal strength measurement, neighbor channel selection etc. We first present a typical link-layer handoff algorithm and then describe the problems in the following subsections.

3.1. Pure layer-2 handoff

Current wireless LAN products support handoff at layer 2 with the aid of information from layer 1. The handoff is mobile-initiated. Basically, for each MH equipped with a wireless LAN adapter, a serving channel will be selected for carrying all the data packets between MH and the corresponding AP (access point). The MH, while staying in the radio coverage of the AP, will periodically check the current RSSI (Receiving Signal Strength Indicator) and calculate the current FER (Frame Error Rate) on the serving channel. Besides, the MH actively scans all other channels for their receiving signal strength, also periodically. At any time, if the quality of the serving channel falls down under some predefined threshold, the MH will decide to start a handoff process and look for a new serving channel.

The pure layer-2 handoff procedure is given as follows.

Pure layer-2 handoff algorithm

/* Parameters:
RSSI-S: the RSSI of the serving channel,
FER-S: the FER of the serving channel,
RSSI-N1: the RSSI of the best neighboring channel N1,
RSSI-N2: the RSSI of the second best neighboring channel N2,
...
RSSI-Nk: the RSSI of the kth best neighboring channel Nk,
RSSI-X: the threshold value of RSSI,
FER-X: the threshold value of FER,
Δ: the smoothing factor. */

Periodical procedures
Check and update RSSI-S, FER-S, RSSI-N1, RSSI-N2, ..., RSSI-Nk.

Handoff conditions
1. If RSSI-N1 $\geqslant$ RSSI-X, issue a handoff command when RSSI-S $<$ RSSI-X or FER-S $>$ FER-X.
2. If RSSI-N1 $<$ RSSI-X, issue a handoff command when RSSI-N1 $>$ RSSI-S + Δ.

Handoff procedures
(1) If one of the handoff conditions occurs, make a handoff decision and start to attach to the network on channel N1 at layer 2.
(2) If the attachment to the network on channel N1 fails, repeat the step (1) on channel N2, N3, ..., until a successful attachment is achieved.
(3) If no attachment attempt on any neighboring channel succeeds, report a connection termination.

/* The smoothing factor Δ is a positive value that avoids the handoff oscillation between two neighboring base stations. */

3.2. Delayed mobile IP roaming

The first problem encountered in MWIN is the latency caused by a handoff or roaming process. At layer 2, the wireless interface will measure the radio quality and determine to switch to another radio channel. However, at layer 3, the MH will not know that it has lost contact with home network until one of the following conditions occurs:

1. AAR timer expires.

2. Receive a different IP subnet prefix. (However, this will not trigger the MH to re-register via a new agent. The reason has been explained in section 2.2.)

It is also possible that a mobile host can receive a different router advertisement even when it is still located in its home, for example, there are two wireless interfaces in the mobile host. But, this does not mean that the mobile host should immediately handoff to the new network. In order to avoid the *ping-pong* effect (i.e., the mobile host is located in an area where two radio signals overlap), the mobile host will try to stay in its original network until communication is lost with the network. According to [21], the mobile host keeps waiting for the original router advertisement message until it is timed out. Then, it will begin to search for a new agent and register to it.

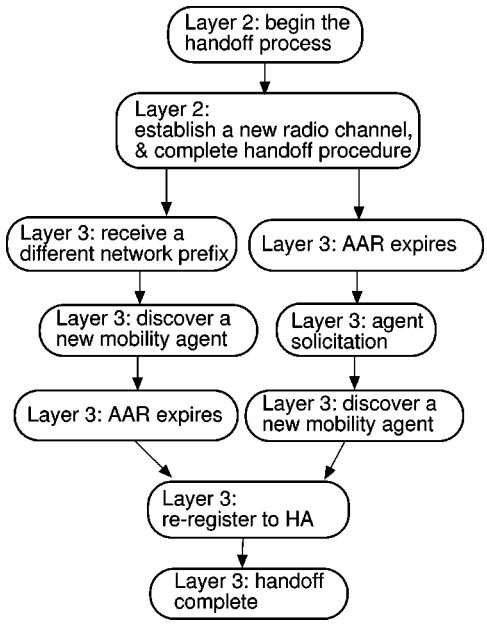

Figure 4. The handoff/roaming flow in MWIN.

It was suggested in [21] that the agent advertisement should be sent every 1 s and the AAR timer is set to three times of that, i.e., 3 s. To avoid the network congestion, this interval can be set longer, for example, 2 or 3 s. In any case, it would take several seconds since the router advertisement message will not be sent too frequently (at most once in 1 s). In figure 4, the flow chart of handoff and roaming in MWIN is shown. As introduced in the pure layer 2 handoff process, the MH will switch into a new radio channel if one of the handoff conditions occurs. At layer 3, the agent advertisement packet(s) may be lost and new agent advertisement packets may be received. Note that the new agent advertisement massage may contain informations such as the different network prefix. But the MH will not make a decision to switch into the new network unless the AAR timer expires. Without the aid of link layer information, all the roaming decisions at layer 3 should rely only on the AAR timer and this will cause additional delay. If the overall latency for handoff and roaming is too long, some applications at higher layer would be also timed out and the program will be terminated.

To explain the problem with more details, a message sequence chart of handoff and roaming in MWIN is given in figure 5. This example assumes that the radio quality becomes unacceptable after the receipt of the second data packet. Thus, the third data packet and those after the third will be lost. They will be recovered only when the MH successfully re-registers to HA via a new FA. As shown in the figure, there are two major factors besides the exchange of control messages that contribute to the handoff/roaming delay. The first one is the time spent in the move detection of mobile IP, i.e., AAR timer expiration. The second one is caused by TCP retransmission timer if we assume the upper layer is using TCP (this will be explained later in section 4).

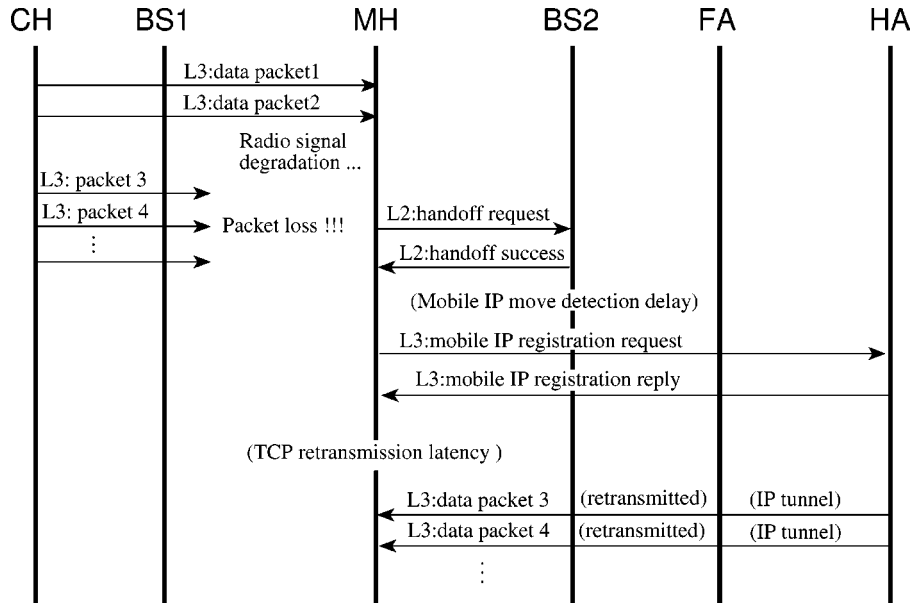

Figure 5. The message sequence chart of handoff/roaming in MWIN.

3.3. Packet loss problem

Following the problem described previously, there may be a packet loss problem with MWIN using mobile IP. In most of the cellular phone networks, handoff decision is made by the base station and there is a pre-setup on the neighbor channel. Thus, the base station has a chance to duplicate the data on the neighbor base station and avoid data loss. However, in mobile IP, the roaming decision at layer 3 is made by MH. And the handoff decision at layer 2 is made independently from layer 3. Referring to figure 5, when the radio quality degrades down to an unacceptable level, the mobility agent may not be aware of this event immediately and thus packets addressed to this MH are dropped. This will be recovered when the MH has successfully re-registered to HA via a new FA.

Recently, some research has been made on the prevention of packet loss due to handoff. For example, the use of multicast-based handoff has been proposed in [12,20]. However, the proposed method still does not take into consideration the layer 2 behavior and, therefore, the delay problem still exists. If the delay is longer, the redundant multicast packets will keep on wasting the limited wireless bandwidth.

3.4. Isolated subnet

In conventional cellular phone networks, radio cell planning is a necessary procedure before the networks becoming acceptable. At layer 2 or layer 1, the radio coverage of each base station should be tuned carefully such that the networks are optimized, for example, the reduction of co-channel interference, the increase of frequency reuse, neighboring cell assignment, etc. At layer 3, some parameters must be assigned optimally such that the network can operate efficiently, for example, the handoff threshold, minimum acceptable signal strength, or maximum acceptable bit error rate.

However, in MWIN using mobile IP, there is no cell planning for mobile Internet. It is possible that a MH will enter an unexpected situation. In figure 6, subnet 3 is an isolated network, or alternatively, subnet 3 does not have a mobility agent or router. Since the handoff process at layer 2 is performed independently from layer 3. The MH may be forced to enter subnet 3 at layer 2, instead of subnet 2 which supports Mobile IP. At layer 3, the MH will fail to find a new default router or agent within subnet 3. Unless the MH has moved into a location where the signal strength of subnet 2 is stronger, or the wireless interface is forced to be switched into subnet 2 (triggered by layer 2), the MH cannot make any communication.

3.5. Routing problem

Following the isolated subnet problem, a MH may fail to find a good network when it is searching for a new mobility agent. In figure 7, an example shows that a MH chooses a bad route. There are many cases when a MH moves to a location with more than one mobility agent available. Let subnet 2 and 3 be the new network that the MH will be moving into. If subnet 3 is selected, the route distance to its home network is much longer the that starting from subnet 2. If route optimization option in mobile IP is not used, this will cause big problems in delay and bandwidth utilization.

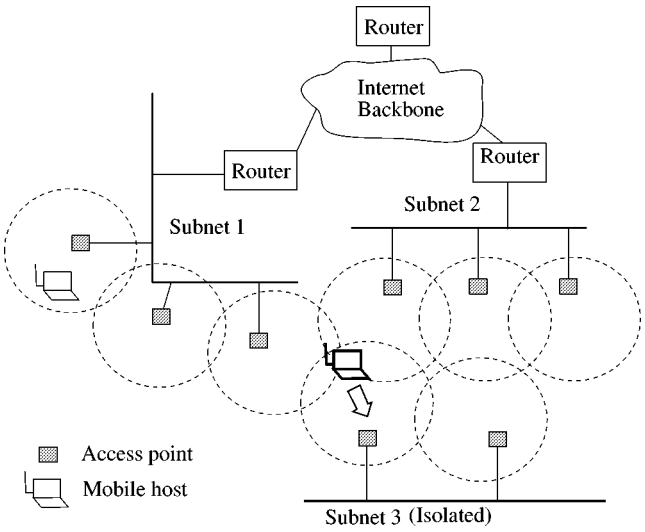

Figure 6. The isolated subnet problem.

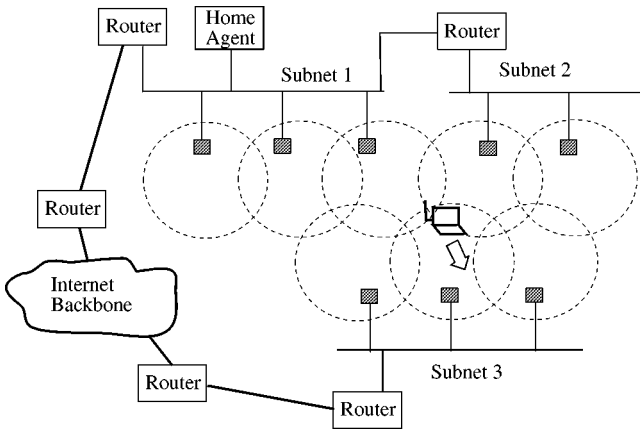

Figure 7. The bad Mobile IP routing problem.

If route optimization option is used, this problem will become a minor issue. But, there are still some necessary traffic between the MH and its home agent. For example, if the binding update message required by the route optimization can arrive earlier at the home network, the route can be optimized earlier. Furthermore, it is often the case that a MH tends to communicate with the server located at its home, i.e., e-mail server, ftp server, WWW server, etc. Thus, it makes sense that a MH chooses a FA that is "closer" to the home of the MH.

4. Analysis and measurement

In MWIN, the most common source of traffic is naturally traditional Internet traffic, i.e., TCP/IP. To explore the draw-

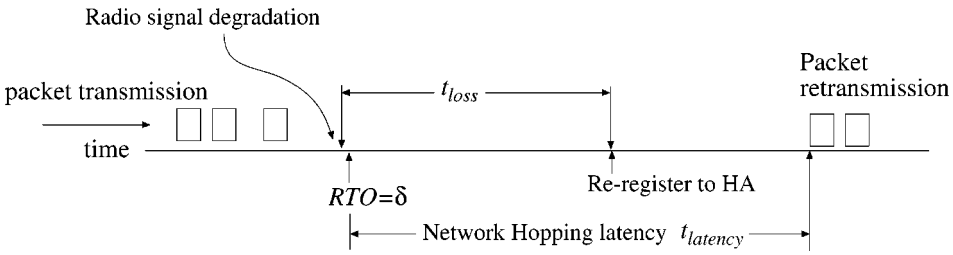

Figure 8. The network hopping latency and packet loss interval.

backs introduced in section 3, the *network hopping latency* at TCP layer and the file transfer throughput during network hopping without the proposed scheme are analyzed and measured. By network hopping we mean that a MH is moving from one subnet into the other subnet in MWIN. Thus, the network hopping latency at TCP layer includes the roaming delay introduced by mobile IP at layer 3 and the handoff delay caused by the radio LAN at layer 2.

4.1. Network hopping latency

When a mobile station moves from one subnet to another, there is a time interval during which this mobile station is out of touch from the HA and any other mobility agent. This interval is referred to as *packet loss interval* because packets will be lost during that interval. In a network hopping process, the packet loss interval is defined as the time interval between a MH's detachment from its original radio LAN at layer 2 and this MH's successful re-registration to its HA. In figure 8, the packet loss interval is denoted as t_{loss}. In a subnet of MWIN, the agent will send an agent advertisement message in every fixed interval. We call this interval as agent advertisement interval, denoted by t_{advert}. According to mobile IPv4 [21], the AAR timer duration (i.e., the duration of waiting for the original agent advertisement) is suggested to be three times of t_{advert}. Therefore, $t_{\text{loss}} \geqslant 3t_{\text{advert}}$. In addition, there are other factors that contribute to the value of t_{loss}. Let t_{L2} represent the time required for a MH to complete the handoff process at the link layer. Let t_{dis} represent the time required for a MH to discover a new mobility agent and let t_{req} represent the time required for a MH to re-register to HA via a new mobility agent. Then, we have

$$t_{\text{loss}} = 3t_{\text{advert}} + t_{\text{reg}} + t_{\text{L2}} + t_{\text{dis}}. \tag{1}$$

Since packets may be lost during t_{loss}, TCP retransmission will be started to recover the error. TCP retranssmision mechanism relies on a retransmission timer, denoted by *RTO*. Each transmitted packet is associated with a *RTO*. Once the timer expired and the corresponding acknowledgement is not received yet, then the corresponding packet will be retransmitted. According to TCP standard [24] or Van Jacobson's algorithm [15], *RTO* will be multiplied by 2 each time the corresponding acknowledgement is not received. Thus, an exponential waiting time can be expected if the packet will be lost during a period of time. In figure 8, we have shown this situation. Let t_{latency} represent this period of waiting time. Then, we have

$$t_{\text{latency}} \geqslant t_{\text{loss}}. \tag{2}$$

Let the first retransmission in t_{loss} occur in RTO $= \delta$, as shown in figure 8. Assuming that the RTO timer expired n times within t_{loss}, then we have

$$\begin{aligned} t_{\text{latency}} &= \delta + \delta \cdot 2 + \delta \cdot 2^2 + \cdots + \delta \cdot 2^n \\ &= \delta \cdot 2^{n+1} - \delta. \end{aligned} \tag{3}$$

Therefore, n equals to the minimum integer k such that

$$\delta \cdot 2^{k+1} - \delta > t_{\text{loss}}. \tag{4}$$

Hence,

$$k > -1 + \log_2\left(1 + \frac{t_{\text{loss}}}{\delta}\right). \tag{5}$$

Thus,

$$k = \left\lfloor \log_2\left(1 + \frac{t_{\text{loss}}}{\delta}\right) \right\rfloor. \tag{6}$$

In equation (3), let $n = k$. Then we have[1]

$$\begin{aligned} t_{\text{latency}} &= \delta \cdot 2^{k+1} - \delta \\ &= \delta \cdot 2^{\lfloor \log_2(1+t_{\text{loss}}/\delta) \rfloor + 1} - \delta \\ &\leqslant 2\delta\left(1 + \frac{t_{\text{loss}}}{\delta}\right) - \delta \\ &= 2t_{\text{loss}} + \delta. \end{aligned} \tag{7}$$

From equations (2) and (7), we conclude that

$$t_{\text{loss}} \leqslant t_{\text{latency}} \leqslant 2t_{\text{loss}} + \delta. \tag{8}$$

Assuming a MH is roaming in a foreign network, the performance of t_{latency} can be estimated as follows. t_{L2} for a typical wireless LAN is usually in the range from several hundred μs up to several ms. t_{dis} is also in the range of several ms. However, t_{reg} depends on the traffic load and the distance from the MH to its HA. We use *trace route (tracert)* on Windows 95 to estimate the packet transfer time between our research center within Computer Communications Research Laboratories (140.96.89.59) and the Department of Computer Science at National Tsing Hua University (140.114.78.68). These two organizations are about 9 hops

[1] According to RFC 793 [24], there is an upper bound for RTO. But this constraint makes no change to equation (7).

Table 1
The relation of t_{loss}, t_{advert}, and t_{latency} (in s), when $\delta = 200$ ms, $t_{\text{reg}} + t_{\text{L2}} + t_{\text{dis}} = 200$ ms.

t_{advert}	t_{loss}	Max t_{latency}
1	3.2	6.6
2	6.2	12.6
3	9.2	18.6
4	12.2	14.6
5	15.2	30.6
6	18.2	36.6

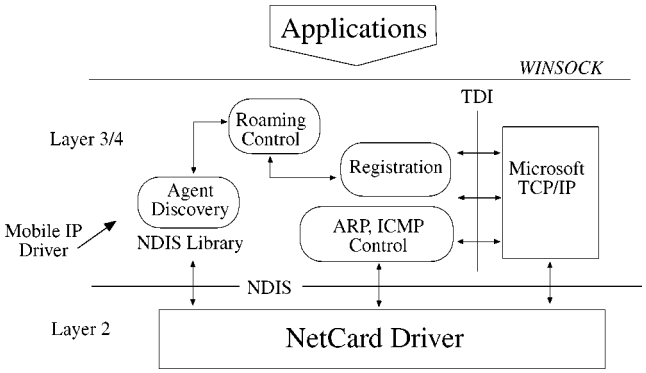

Figure 9. The mobile host design on Windows 95.

away. The result of trace route ranges from 100 ms up to around 600 ms.

In table 1, we present the value of t_{latency} for different t_{advert}. Note that δ is the TCP retransmission timer at the beginning of the network hopping process. According to RFC 793, TCP retransmission timer is set to twice of the estimated round trip delay. We assume that the round trip delay is 100 ms; thus, $\delta = 200$ ms. We also assume that $t_{\text{reg}} + t_{\text{L2}} + t_{\text{dis}} = 200$ ms. From table 1, the resulting network hopping latency is still very large even when t_{advert} is as small as 1 s. Basically, a very large percentage of the networking hopping latency is due to the AAR timer.

4.2. File transfer throughput measurement

We have implemented mobile IPv4 on Windows 95/98[2]. Both MH and mobility agent are designed. The MH is implemented as a Windows 95 VxD protocol driver. Figure 9 shows the program architecture which has two interfaces. ARP control and related ICMP functions are interfacing with NDIS at link layer. Roaming control and registration functions are interfacing with Microsoft TCP/IP. The mobility agent is also implemented on Windows 95 using a similar architecture. The design of the mobility agent is based on the Linux version program developed by the State University of New York, Binghamton [16]. For our experiment, we assignment different values of t_{advert} in the programs for the agent and mobile host.

In figure 10, the experimental configuration of a MH moving from its home to a neighboring network is shown.

[2] For obtaining the software or more information, please contact Dr. Chiung-Shien Wu at cwu@atc.ccl.itri.org.tw, or chiungwu@acm.org.

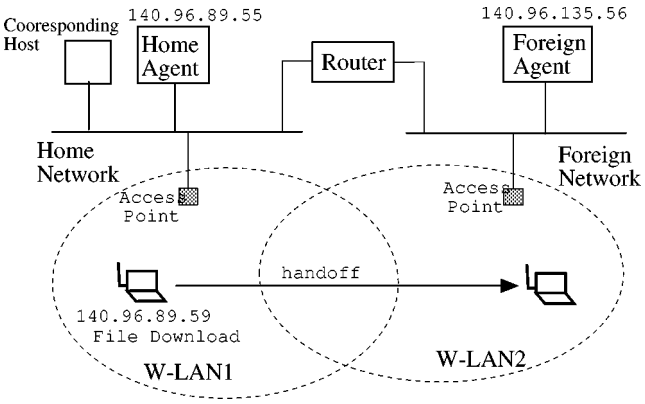

Figure 10. The experimental configuration.

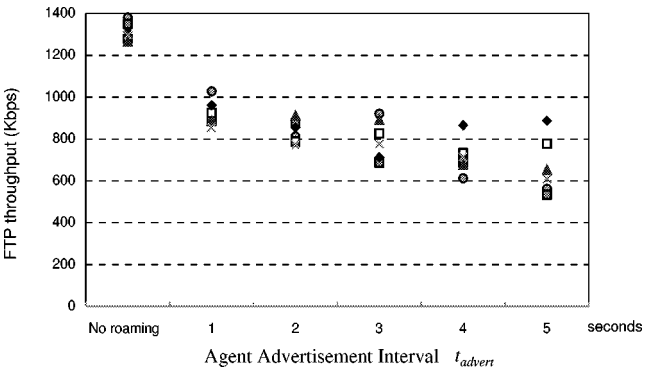

Figure 11. The measured FTP throughput of a MH moving from its home network to a neighboring foreign network.

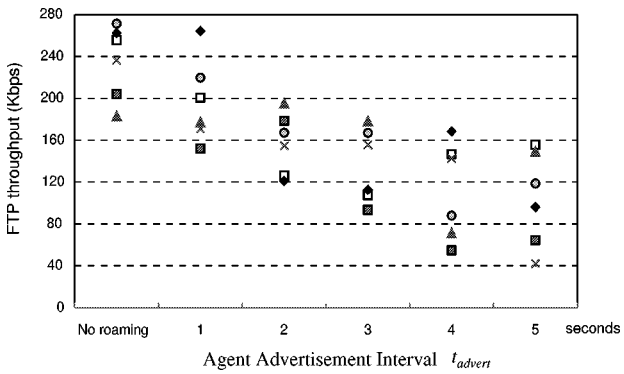

Figure 12. The measured FTP throughput of a MH moving from a foreign network to another neighboring foreign network.

The IEEE 802.11 wireless LAN developed by Computer Communication Research Laboratories is used. The moving speed of the mobile host is around 8 km/h (walking speed). We assume that the MH is downloading a file of 8650752 bytes from a corresponding host that is also located in the home network of MH. The measured FTP throughput with respect to different t_{advert} are shown in figures 11 and 12. Figure 11 shows the result obtained from a configuration that the MH (140.96.89.59) is moving from its home network (140.96.89.XXX) to the neighboring network (140.96.135.XXX). The result corresponding to "no roaming" means the MH keeps staying in the home network dur-

ing the FTP session. It can be observed that the average throughput is degraded during a network hopping process because there is additional latency. When t_{advert} is larger, the variation of performance becomes more serious and unpredictable. For example, it takes about 55 s to download the above 8 Mbytes file at 1.3 Mbps in a "no roaming" condition. In the case of network hopping process, and assuming the latency is 10 s, the download throughput will degrade down to 1 Mbps. If the latency is as large as 40 s, the download throughput will reduce to around 700 Kbps.

Figure 12 shows another measurement. In this case, the home network of the MH is set at the Department of Computer Science at National Tsing Hua University (140.114.78.XXX) which is 9 hops away from the foreign network (140.96.89.XXX). The measurement is made when the MH is moving from the network (140.96.89.XXX) to the neighboring network (140.96.135.XXX). The FTP throughput in "no roaming" case is around 270 Kbps since now the MH is downloading the file from the server located in National Tsing Hua University (140.114.78.XXX). The average FTP throughput is similar but is lower than the result shown in figure 11.

5. Intelligent handoff architecture

The preceding analysis demonstrates a problem with wireless Internet that occurs from the fact that layer 2 does not synchronize with layer 3 to support handoff/roaming. In addition, there is no global network planning to optimally allocate and arrange the radio resources. For example, at layer 3, there is no high-layer radio network planning to accommodate the handoff process at layer 2. Fortunately, those missing functions are not forbidden neither in Mobile IP specifications nor wireless LAN standards [14,17,21]. Therefore, we propose an intelligent handoff architecture that extends the handoff features of the above mentioned specifications.

The proposed architecture is made compatible with Mobile IP and most layer 2 wireless networks. The solution consists of three major extensions: packet buffering, neighbor list update, and layer 2 handoff notification, as described below.

Packet buffering. The first extension is a new ICMP packet that notifies the mobility agent to buffer the packet temporarily when the MH is about to handoff at layer 2. This can prevent packet loss due to the network hopping process. After the MH's successful roaming to a new network, the HA will receive a re-registration request from the new FA. As original design of mobile IP, the HA will then tunnel the following packets to this new FA. However, with this extension, the HA should send a re-route message to notify the old FA to forward the buffered packets to the new FA. The re-route message is agent-to-agent communications and therefore can be carried either by ICMP or UDP.

Neighboring mobility agent's IP	Link layer network type	Link layer RF infomation
140.114.78.68	IEEE 802.11	2.4 GHz DSSS, Ch1
140.96.89.55	IEEE 802.11	2.4 GHz DSSS, Ch5
140.96.135.56	IEEE 802.11	2.4 GHz DSSS, Ch7
140.96.87.222	IEEE 802.11	2.4 GHz DSSS, Ch11
.....	....	....

Figure 13. An example of neighbor list.

Neighbor list update message. The previous extension is an example that layer 2 helps the layer 3 to prevent packet loss. In this extension, layer 3 can also help layer 2 to make the handoff faster. Each mobility agent should keep a neighbor list that contains very useful informations for the MH. Each entry in the neighbor list has the following informations: (1) *neighboring mobility agent IP address*, (2) *related link layer network type*, and (3) *channel information at link layer*. All entries in the list are candidates that a MH may possibly roam into from its current location. Figure 13 shows an example of a neighbor list that is kept in a mobility agent. When a MH is staying in a mobility agent, it can use the information in the neighbor list to quickly handoff to a neighboring network and quickly make the re-registration at layer 3. In our proposal, there are two ways that a MH can obtain the neighbor list. First, the neighbor list can be attached in the agent advertisement ICMP packet as a new extension. Secondly, the MH can actively send a request to its serving mobility agent for downloading the neighbor list, possibly via a new ICMP packet. Using the neighbor list can also prevent the isolated network problem presented earlier. It can also help the optimization of mobile IP routing problem. Note that the agent's IP address in each entry of the neighbor list may be the same as the current serving agent. This means that the sub-network contains a radio coverage formed by several radio LANs. In this case, a network hopping process is not needed and a control packet called *buffer release* (may be also carried by ICMP) will be sent from the MH to the current serving mobility agent. The establishment of the neighbor list for each mobility agent relates to the global network planning of the MWIN which will be discussed in a later subsection.

Layer 2 handoff notification to layer 3. In the MH, it's nice for the layer 3 program to know exactly when there is a new and successful handoff at layer 2, thereby avoiding the extra waiting until AAR timer expires. Thus, a message may be sent from the layer 2 up to layer 3. And this will trigger the layer 3 to send a registration request in the new subnet. Note that the MH does not need to send an agent solicitation as there is agent information carried in the neighbor list. The MH can directly send a re-registration request to HA via the new FA. This will prevent the long delay problem at the layer 3.

Figure 14 summarizes the new control messages in the proposed architecture.

Control messages	Direction
Packet buffering request	MH -> serving FA
Packet buffering confirm	serving FA -> MH
Packet buffering release	MH -> serving FA
Packet re- route command	HA -> previous FA
Neighbor list request	MH -> serving FA or HA
Neighbor list download	serving HA or FA -> MH
Link layer handoff notification	Leyer 2 -> layer 3 (within a MH)

Figure 14. The proposed new control messages.

5.1. The proposed handoff algorithm

The proposed algorithm that combines layer 2 handoff and layer 3 roaming for the MH is given as follows.

Modified handoff algorithm (layer 2 + layer 3)

/* Parameters :
- RSSI-S: the RSSI of the serving channel,
- FER-S: the FER of the serving channel,
- RSSI-N1: the RSSI of the best neighboring channel N1 in the neighbor list,
- RSSI-N2: the RSSI of the second best neighboring channel N2 in the neighbor list,
- ...
- RSSI-Nk: the RSSI of the kth best neighboring channel Nk in the neighbor list,
- RSSI-X: the threshold value of RSSI,
- FER-X: the threshold value of FER,
- Δ: the smoothing factor. */

5.1.1. Periodical procedures

Check and update RSSI-S, FER-S, RSSI-N1, RSSI-N2, ..., RSSI-Nk.

Handoff conditions

1. If RSSI-N1 $\geqslant$ RSSI-X, issue a handoff command when RSSI-S $<$ RSSI-X or FER-S $>$ FER-X.
2. If RSSI-N1 $<$ RSSI-X, issue a handoff command when RSSI-N1 $>$ RSSI-S $+ \Delta$.

Handoff procedures

(1) If one of the handoff conditions occurs, send a packet buffering request to the original serving agent at layer 3. /* Packet buffering. */

(2) After the success of the packet buffering request, make a handoff decision and start to attach to the network on channel N1 at layer 2 (note that channel N1 is in the neighbor list). /* Neighbor list. */

(3) If the attachment to the network on channel N1 fails, repeat procedure (2) on channel N2, N3, ..., until a successful attachment is achieved.

(4) After the successful attachment on channel Ni ($1 \leqslant i \leqslant k$), send a mobile IP registration request via the new agent on channel Ni (network hopping case), or send a packet buffer release to the original serving mobility agent (layer 2 handoff only). /* Layer 2 handoff notification to layer 3. */

(5) If no attachment attempt on any neighboring channel succeeds, or the registration request is not granted, report a connection termination.

/* The smoothing factor Δ is a positive value that avoids the handoff oscillation between two neighboring base stations. */

The message sequence chart of using the proposed handoff algorithm is shown in figure 15. Due to the cooperation of layers 2 and 3, the network hopping latency is efficiently reduced. Using the new scheme, it can be found that the packet loss interval consists only t_{reg}, t_{dis}, and t_{L2}. Thus, equation (1) can be rewritten as

$$t_{\text{loss}} = t_{\text{reg}} + t_{\text{dis}} + t_{\text{L2}}, \tag{9}$$

and equation (8) can be changed into

$$t_{\text{reg}} + t_{\text{dis}} + t_{\text{L2}} \leqslant t_{\text{latency}} \leqslant 2(t_{\text{reg}} + t_{\text{dis}} + t_{\text{L2}}) + \delta. \tag{10}$$

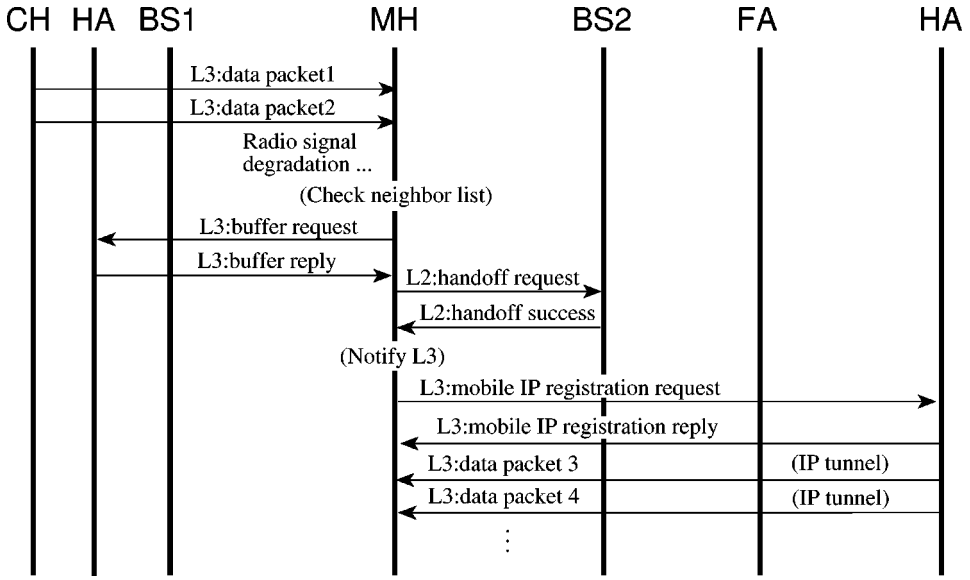

Figure 15. The message sequence chart of using intelligent handoff in MWIN.

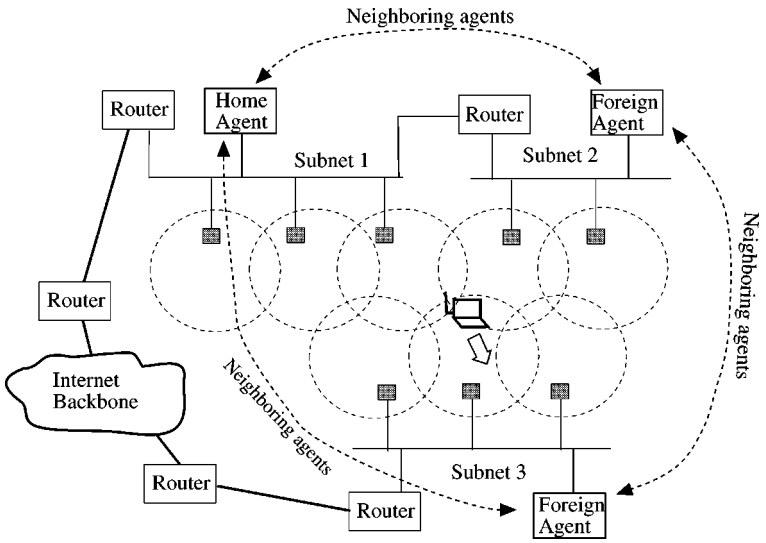

Figure 16. The neighboring agents in MWIN.

In the example of table 1, the network hopping latency can now be bounded by

$$200 \text{ ms} \leqslant t_{\text{latency}} \leqslant 600 \text{ ms}. \tag{11}$$

Note that the proposed scheme uses the packet buffer extension during the network hopping process. Thus, there will be no packet loss. The above observation shows that the HA requires only a small space to buffer the packet temporarily while the MH is hopping to another network.

In MWIN, it is worth to note that a handoff at layer 2 may not necessarily cause a network hopping at layer 3. But in the proposed architecture, a packet buffer request will be sent at the time when the radio signal degrades suddenly. Thus, after the successful handoff at layer 2, the MH should judge whether there is a network hopping or not. By examining the neighbor list entries, the MH can distinguish whether the new radio LAN is located in the same network or not. The establishment of a suitable neighbor list is described in the following section.

5.2. Forming the neighbor list

The neighbor list stored in each mobility agent provides the MH a global information related to what the network topology looks like and how the radio resources are allocated. The establishment of the neighbor list is a very important task related to the so-called *cell planning* or *mobile network planning*. In most of the existing mobile networks, network planning is a very important and difficult task. Although there are computer tools that aid the above process, a lot of efforts were still done manually. This is primarily because of the unpredictable characteristics of radio performance. For MWIN, the first requirement in generating a neighbor list is the knowledge of the global network topology. Then, in order to guarantee the correct mobility, four principles in defining the neighbor list for each agent must be followed.

1. *Connectivity.* Agents that are not connected to the Internet backbone should not be selected in the neighbor list. This is to avoid the isolated subnet problem stated in section 3.3.
2. *Overlap at layer 3.* If two radio LANs located in two different sub-networks have overlapped areas, then the two agents, each located on one of the two sub-networks, should be set to neighbors with each other.
3. *Overlap at layer 2.* If two radio LANs located in the same sub-networks have overlapped areas, then the link layer information of these two radio LANs should be included in the neighbor list.
4. *Mutuality.* If agent A is a neighbor of agent B, then agent B must be selected as a neighbor of agent A. This is to prevent the one-way handoff problem in mobile networks.

An example satisfying the above principles is shown in figure 16.

5.3. Routing consideration

The neighbor list sent by a agent is basically the same for all MH within the subnet. At each MH, different considerations may possibly be taken, for example, the *distance-to-home* optimization. It is very likely that a MH will communicate with the servers at home network very often. One common example is e-mail. People may want to check their e-mail from time to time while they are roaming in a foreign network. Another example is the file server that contains useful software programs. Thus, one may want to keep its distance to home network as short as possible.

It is achievable for a MH to choose a neighboring network with a shorter distance to home during the process of network hopping. Given a neighboring agent's IP address, there are four ways to obtain the distance-to-home value, as follows.

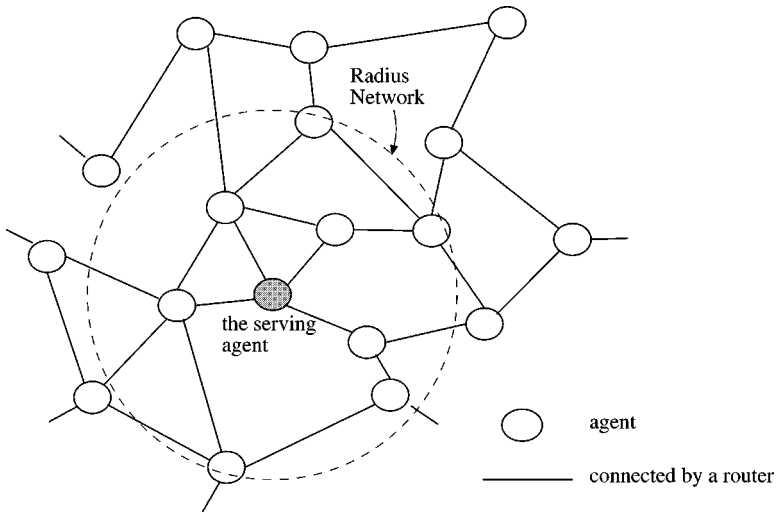

Figure 17. The radius network of a Mobile IP agent.

1. The simplest way to obtain the distance-to-home is to download the routing table from the router. Routers may obtain the hop counts to a targeting network via routing information protocol (RIP), or open shortest path first (OSPF) [13,19].

2. In the network planning stage, prepare a topology configuration of a special network called *radius network*, *RN*. RN is the network that uses the corresponding agent as a center and covers the region within a predefined radius. The distance-to-home value can be simply obtained by calculating the all-pair shortest path of RN. That is to assume that a mobile host is likely to move within the range of the predefined radius. Besides, it is unlikely to put the calculation on the whole Internet even if one assumes that a mobile host may be roaming anywhere. An illustrative description is shown in figure 17.

3. To compensate the shortcoming when a mobile host is moving too far, the ICMP trace route scheme can be used. However, this is very time-consuming since the response of the ICMP trace route can cause a significant delay.

4. *Estimation*. When a mobile host is moving from an agent to another, a worst-case estimation on the distance-to-home can be easily obtained by increasing every old distance-to-home value by 1 before the correct value is received. This estimation is simply based on an assumption that the mobile host is moving farther from its home.

6. Conclusions

This paper presents an extension on the mobile IP in MWIN, thereby enabling the interaction of layer 2 and layer 3 during the network hopping process. This is very important as in most of the mechanism proposed for the mobile cellular phone system, the handoff process is carried out by layer 1 through layer 3. For Internet to serve as a good infrastructure for mobile network, it is necessary to provide a reliable and efficient handoff/roaming service. One advantage of our method is the proposed method is an add-on feature on the current mobile IP and Internet router. It is compatible with the current network. This enables our scheme to serve as an interim solution for real mobile Internet.

Acknowledgement

The authors would like to thank the guest editors and referees for their valuable comments, which have helped greatly to improve the presentation of the paper.

References

[1] G.A. Awater and J. Kruys, Wireless ATM – An overview, Mobile Networks and Applications 1(3) (December 1996) 235–243.
[2] E. Ayanoglu, S. Paul, T.F. LaPorta, K.K. Sabnani and R.D. Gitlin, AIRMAIL: A link-layer protocol for wireless networks, Wireless Networks 1(1) (1995) 47–60.
[3] A. Bakre and B.R. Badrinath, I-TCP: Indirect TCP for mobile hosts, Technical report DCS-TR-314, Rutgers University (1994).
[4] A. Bakre and B.R. Badrinath, Handoff and system support for indirect TCP/IP, in: *Proc. Second USENIX Symp. on Mobile and Location-Independent Computing* (1995).
[5] H. Balakrishnan, V.N. Padmanabhan, S. Seshan and R.H. Katz, A comparison of mechanisms for improving TCP performance over wireless links, IEEE/ACM Transactions on Networking 5(6) (December 1997) 756–769.
[6] H. Balakrishnan, S. Seshan and R.H. Katz, Improving reliable transport and handoff performance in cellular wireless networks, Wireless Networks 1(4) (1995) 469–481.
[7] R. Caceres and L. Iftode, Improving the performance of reliable transport protocols in mobile computing environments, IEEE Journal on Selected Areas in Communications 13(5) (1994) 850–857.
[8] R. Caceres and V.N. Padmanabhan, Fast and scalable wireless handoffs in support of mobile Internet audio, in: *Proc. Second Annual International Conference on Mobile Computing and Networking (MOBICOM'96)*, New York (November 1996) pp. 56–66.
[9] Cellular System, IS-54 (EIA/TIA 553), Dual-mode mobile station–base station compatibility standard, EIA, Engineering Department, PN-2215 (December 1989).

[10] Cellular System, IS-95 (EIA/TIA 553), Dual-mode mobile station–base station wideband spread spectrum compatibility standard, EIA, Engineering Department, PN-3118 (December 1992).
[11] E. Dahlman, B. Gudmundson, M. Nilsson and J. Skold, UMTS/IMT-2000 based on wideband CDMA, IEEE Communications Magazine 36(9) (September 1998) 70–80.
[12] T.G. Harrison, C.L. Williamson, W.L. Mackrell and R.B. Bunt, Mobile Multicast (MoM) protocol: Multicast support for mobile hosts, in: *Proceedings of 3rd ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'97)* (1997) pp. 151–160.
[13] C.L. Hedrick, Routing information protocol, Internet document RFC 1058 (June 1988).
[14] IEEE Standard, Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, IEEE P802.11, D6.1 (May 1997).
[15] V. Jacobson, Congestion avoidance and control, ACM Computer Communication Review, ACM SIGCOMM (1988) 314–329.
[16] V. Jacobson, Mobile IP implementation, `http://anchor.cs.binghamton.edu/mobileip/`.
[17] D.B. Johnson and C. Perkins, Mobility support in IPv6, Internet draft, Mobile IP Working Group.
[18] D.N. Knisely, S. Kumar, S. Laha and S. Nanda, Evolution of wireless data services: IS-95 to cdma2000, IEEE Communications Magazine 36(10) (October 1998) 140–149.
[19] J. Moy, OSPF Version 2, Internet document RFC 1247 (July 1991).
[20] J. Mysore and V. Bharghavan, A new multicast-based algorithm for Internet host mobility, in: *Proceedings of 3rd ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'97)* (1997) pp. 161–172.
[21] C. Perkins, IP mobility support, Internet document RFC 2002 (October 1996).
[22] C.E. Perkins, IP encapsulation within IP, Internet document RFC 2003 (October 1996).
[23] C.E. Perkins, *Mobile IP: Design Principles and Practice* (Addison-Wesley, 1998).
[24] J. Postel, Transmission control protocol, Internet document RFC 793 (October 1981).
[25] S.M. Redl, M.K. Weber and M.W. Oliphant, *An Introduction to GSM* (Artech House Publishers, 1995).
[26] D. Saha and S.E. Kay, Cellular digital packet data network, IEEE Transactions on Vehicular Technology 46(3) (August 1997) 697–706.
[27] A.K. Salkinzis and C. Chamzas, Mobile packet data technology: An insight into MOBITEX, IEEE Personal Communications 4(1) (February 1997) 10–18.
[28] J.D. Solomon, *Mobile IP: The Internet Unplugged* (Prentice-Hall, 1998).
[29] F. Teraoka, K. Uehara, H. Sunahara and J. Murai, VIP: A protocol providing host mobility, Communications of the ACM 37(8) (August 1994) 67–75.
[30] E.K. Wesel, *Wireless Multimedia Communications: Networking Video, Voice and Data* (Addison-Wesley, 1998).

Jon Chiung-Shien Wu received his B.S. degree in computer science and information engineering from National Taiwan University in 1989, and his Ph.D. degree in computer science from National Tsing Hua University in 1994. Since 1994, he joined Computer Communication Research Labs., Industrial Technology Research Institute (ITRI), where he works as a researcher on broadband wired and wireless communication systems. He has been involved in the design of several prototype systems such as PC-based video-on-demand, mobile Internet system, and RLC/MAC protocol for GPRS/WCDMA/UMTS. Dr. Wu conducted ITRI's activity of participating in the 3GPP WG RAN2 and made 3 contributions to the committee (2 of them were included in the technical specification). He has published over 60 papers in refereed journals and conferences and was on the Marquis 1999 Who's Who in the World and the 1999 International Who's Who in Professional Management. He received the research achievement award from ITRI in 1998 for his contribution on mobile Internet research. Dr. Wu joined Philips Research East Asia in October 1999, working on protocol and architecture of home networking and mobile Internet access. His current research interests include wireless connectivity for home and away environment (Bluetooth & UMTS), mobile Internet protocols and mobile multimedia systems and protocols. Dr. Wu is a member of ACM and IEEE computer and communication society.
E-mail: jon.cs. wu@philips.com

Chieh-Wen Cheng received the B.S. degree in applied mathematics from National Chung Hsing University, Taichung, Taiwan, in 1991 and the M.S. degree in computer science from National Tsing Hua University in 1993. He is currently working toward the Ph.D. degree at National Tsing Hua University. His main research interests are in the area of mobile communication networks.
E-mail: dr828304@cs.nthu.edu.tw

Nen-Fu Huang received the B.S.E.E. degree from National Cheng Kung University, Taiwan, R.O.C., in 1981, and the M.S. and Ph.D. degrees in computer science from National Tsing Hua University, Taiwan, R.O.C., in 1983 and 1986, respectively. In 1986–1994, he was an Associate Professor of Department of Computer Science at National Tsing Hua University, Taiwan, R.O.C., in 1994–1997, he was a Professor and the Chairman of the same department. His current research interests include ATM networks, mobile networks, and high-speed multi-layer switching routers. Dr. Huang is the Guest Editor of the special issue on Bandwidth Management on High-speed Networks for the Computer Communications Journal. Since 1997, he serves as the Editor of the Journal of Information Science and Engineering. He received the Outstanding Teaching Award from National Tsing Hua University in 1993 and 1998, and the Outstanding University/Industrial Cooperating Award from the Ministry of Education, R.O.C., in 1998. Dr. Huang is a member of IEEE.
E-mail: nfhuang@cs.nthu.edu.tw

Gin-Kou Ma received the Ph.D. degree in electrical engineering from University of Florida, Gainesville, in 1989. He joined the Athena Group Inc., USA, during 1985–1989, where he was responsible for the MONARCH-DSP CAD tool project. Since 1990, he was working as an R&D Manager of High-speed Broad-band Information Networks: Communication Technologies and Services for the Computer Communication Research Laboratories (CCL) of Industrial Technology Research Institute (ITRI), Taiwan, R.O.C. He is currently the R&D Manager of Advanced Microelectronics System Technologies for Electronics Research & Service Organization (ERSO) of ITRI. He obtained the 1997 National Outstanding Information Engineer Award, Taiwan. He is interested in the researches of high-speed broad-band multimedia communication and digital signal processing technologies. Dr. Ma is a member of ACM and IEEE communication, signal processing, circuits & systems, information theory, and computer societies.
E-mail: gkma@erso.itri.org.tw