Efficient Mobility Management for Vertical Handoff between WWAN and WLAN

Qian Zhang, Chuanxiong Guo, Zihua Guo, and Wenwu Zhu,
Wireless and Networking Group, Microsoft Research Asia

ABSTRACT

As we move toward next-generation all-IP wireless networks, we are facing the integration of heterogeneous networks, such as WWAN and WLAN, where vertical handoff is required. In vertical handoff between WWAN and WLAN, mobile hosts should be able to move freely across different networks while satisfying QoS requirements for a variety of applications. In order to achieve seamless handoff and maintain continuity of connection, in this article we propose a novel mobility management system that integrates a connection manager to detect network condition changes in a timely and accurate manner, and a virtual connectivity manager that uses an end-to-end principle to maintain a connection without additional network infrastructure support. The prototype system was built to test the effectiveness of the proposed system. Experiments show that seamless roaming between WLAN and WWAN can be achieved, and much better performance can be obtained than with the traditional scheme.

INTRODUCTION

The increasing number of Internet users in combination with the evolution of IP-based applications has created a strong demand for wide-area broadband access to IP services. A future wireless Internet is expected to consist of different types of wireless networks, each providing varying access bandwidth and coverage level. Today, the natural trend is to utilize high-bandwidth wireless local area networks (WLANs) such as IEEE 802.11 in hotspots and switch to wireless wide area networks (WWANs) such as General Packet Radio Service/Universal Mobile Telecommunications System (GPRS/UMTS) networks when the coverage of WLAN is not available or the network condition in WLAN is not good enough. We refer to such a procedure as *vertical handoff*. By combining the wide coverage of next-generation cellular systems with the advantage of high bandwidth in WLANs, users can make the most of wireless IP communication.

In the next-generation heterogeneous wireless systems, one of the major challenges is seamless vertical handoff. Here, by seamless we mean that the handoff procedure should be transparent to upper-layer applications. Notice that here there are two directional handoffs, one from WLAN to WWAN and the other in the reverse direction. In order to achieve seamless handoff, several issues, such as handoff metrics and handoff decision algorithms, and mobility handling to maintain ongoing user connections, need to be addressed.

To date handoff schemes that only deal with the switch between base stations (BSs) or access points (APs) in the homogeneous wireless system have been well studied [1]. These are usually called *horizontal handoff*. However, in the case of vertical handoff, we face the following challenges:

- When a user moves from WWAN to WLAN, since the WWAN is usually always on, the handoff cannot be triggered by signal decay of the current system, as in horizontal handoff [2].
- In vertical handoff between WWAN and WLAN, there is no comparable signal strength available to aid the decision as in horizontal handoff.
- During a handoff procedure, the metrics upper-layer applications are really interested in are network conditions (available bandwidth and delay, user preference, etc.), rather than the physical layer parameters such as received signal strength and signal-to-interference ratio.

Some work on vertical handoff has been reported in the literature. In [3] a roaming scheme that considered the relative bandwidth

0163-6804/03/$17.00 © 2003 IEEE

of WLAN and GPRS was proposed. But no information on how to obtain the available bandwidth is given. In [4] a detailed vertical handoff signaling procedure was presented. However, no details on a handoff decision algorithm were provided.

Once the vertical handoff decision has been made, the other key issue for a roaming system is the mobility management scheme, which can maintain a connection's continuity after a vertical handoff. Mobile IP [5, 6] is the most widely studied approach to mobility handling, where packets from and to the mobile host are tunneled through a home agent at its home network, so the corresponding host that communicates with the mobile host can be shielded from the mobility of the mobile host. To improve the routing performance and resolve certain scalability problems associated with Mobile IP, several new schemes such as [7] have been proposed. All of these solutions, including Mobile IP, significantly rely on newly introduced network infrastructures such as the home agent.

Targeting reduction of the performance degradation from Mobile IP and, most important, allowing an easier deployment path than Mobile IP, end-to-end solutions such as Migrate [8] have been proposed. Although Migrate has many attractive characteristics, it has two major limitations:

- Since a host notifies its peer directly of the IP address change, this scheme may not make mobility transparent to applications. That is, legacy UDP-based applications need to be recoded to support mobility.
- As a pure end-to-end approach, it cannot maintain user connections under several cases, such as when both hosts move simultaneously and when hosts are behind a Network Address Translation (NAT) box.

To provide an efficient seamless roaming service to end users, we propose a completely IP-centric approach to address all the above mentioned issues. The architecture of this approach is illustrated in Fig. 1.

In this approach a *connection manager* (CM) is introduced to intelligently detect the conditions of the different types of networks and the availability of multiple networks. The CM can handle two directional handoffs between WWAN and WLAN: from WWAN to WLAN and from WLAN to WWAN. When a user in a WWAN moves into a WLAN, medium access control (MAC)-layer sensing and traditional physical-layer sensing for the WLAN is performed to ensure better quality of service (QoS). On the other hand, when the user moves out of a WLAN area, we should promptly detect the unavailability of the WLAN and switch the connection from WLAN to WWAN seamlessly. Therein a fast Fourier transform (FFT)-based signal decay detection scheme is used to reduce the ping-pong effect, and an adaptive threshold configuration approach is proposed to prolong the time the user stays in WLAN.

Meanwhile, a *virtual connectivity manager* (VC) is proposed to maintain connection's continuity using an end-to-end argument when handoff occurs. By utilizing the information provided by the CM, not only can the VC maintain the connection unbroken, but it also always achieves the best possible communication quality. In the VC, in order to address the aforementioned problems faced by end-to-end schemes:

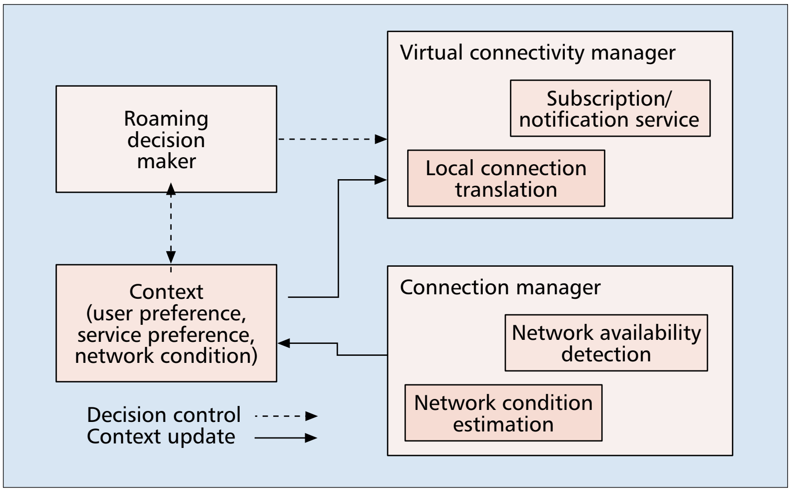

Virtual connectivity manager
Subscription/ notification service
Roaming decision maker
Local connection translation
Context (user preference, service preference, network condition)
Connection manager
Network availability detection
Network condition estimation
Decision control
Context update

Figure 1. *Architecture for seamless handover between WLAN and WWAN.*

- We introduce a local connection translation (LCT), which maintains a mapping relationship between the original connection information and the current connection information for each active connection, therefore making this mobility solution transparent to upper applications.
- A subscription/notification (S/N) service, which provides a bridge between two communicating parties, is proposed to support the NAT and simultaneous movement cases.

In this architecture a *roaming decision maker* and a *context* database are introduced to act as the interconnection between CM and VC. The purpose of this context database is to make roaming context-aware. The context may consist of user/service preferences and technical parameters, such as access delay, available bandwidth, and capabilities of the terminal. The roaming decision maker module makes the roaming decision after taking the entire related context into consideration.

The collaboration between these four components accomplishes efficient end-to-end mobility for vertical handoff. In the next two sections we describe the CM and VC modules in detail.

Connection Manager

In the CM, we consider two handoff scenarios. When moving from WWAN to WLAN, since WLAN is optional, the objective of the handoff is to improve the QoS. When moving out of WLAN, we need to have a timely and accurate handoff decision to maintain the connectivity before the loss of WLAN access.

Handoff from WWAN to WLAN

When a user who is connected to a WWAN system steps into a WLAN area, he/she would like to change the connection to WLAN to obtain

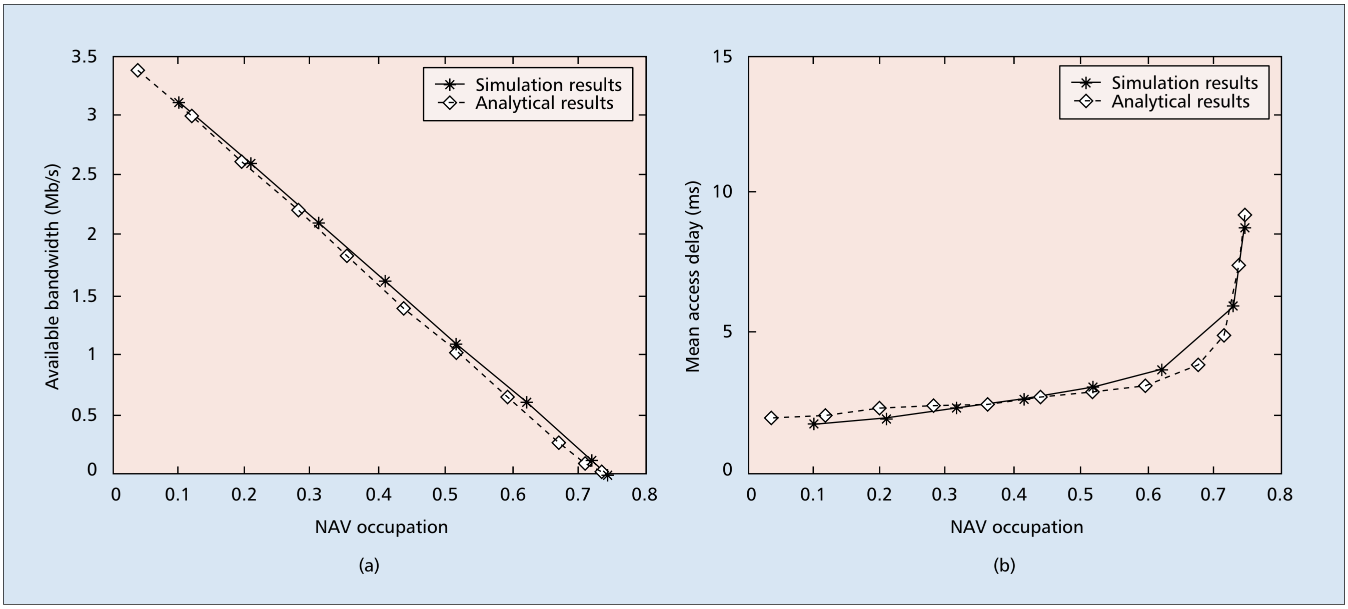


Figure 2. *Network performance under different NAV occupation (packet size =1 kbyte): a) available bandwidth; b) mean access delay.*

possible larger bandwidth and less cost. To ensure better QoS, we need to perform the sensing in both the physical and MAC layers of WLAN. More specifically, physical layer sensing is used to detect the availability of the stable WLAN signal, while MAC layer sensing is used to detect the network conditions of the WLAN system, such as access delay and available bandwidth. Since physical layer sensing is well studied in literature, we only present MAC layer sensing next.

MAC Layer Sensing — In this work we propose to listen and collect the network allocation vector (NAV) in the MAC layer to estimate network conditions (e.g., available bandwidth and access delay). It is known that NAV is the major mechanism in the MAC of the IEEE 802.11 WLAN to avoid collision. Once a station hears other stations' transmission, it will set the NAV to busy state and keep silent for a time duration equal to the duration ID in the packet header [9]. From the above description, we can see that the NAV busy state can well reflect the media's busy state or traffic load. The higher the traffic, the larger the NAV busy occupation will be, and vice versa. With comprehensive simulations and analysis, we have found that there is a fixed relationship between NAV and the available bandwidth and access delay, which is illustrated in Fig. 2a and 2b, respectively. This relationship is insensitive to user number and traffic pattern. That is, once we observe a NAV value in a time window, the available bandwidth and access delay can be estimated given a certain packet length.

In summary, the MAC sensing scheme has several characteristics. First, a network-condition-aware handoff is achieved. Second, the QoS information can facilitate the adaptation of upper transport and application layers. Third, among multiple APs, the one with the best QoS can be selected.

Handoff from WLAN to WWAN

Since WLAN has a smaller coverage range, when the user steps out of a WLAN area, we should quickly detect the unavailability of the WLAN and switch the connection to WWAN seamlessly. Therefore, the goal for this directional handoff is to switch to WWAN before the WLAN link breaks, while staying in the WLAN as long as possible due to lower cost and better QoS. There are two key issues in detecting the WLAN unavailability:

- How to accurately detect the signal decay. Although on average the mean received signal strength indication (RSSI) will decrease when a user leaves, there is a great deal of variation in the sampled RSSI (up to 10 dB).
- How to determine if the signal is weak. Note that different manufacturers have different card implementation techniques, although the standard defines a lower bound for WLAN signal receiver sensitivity.

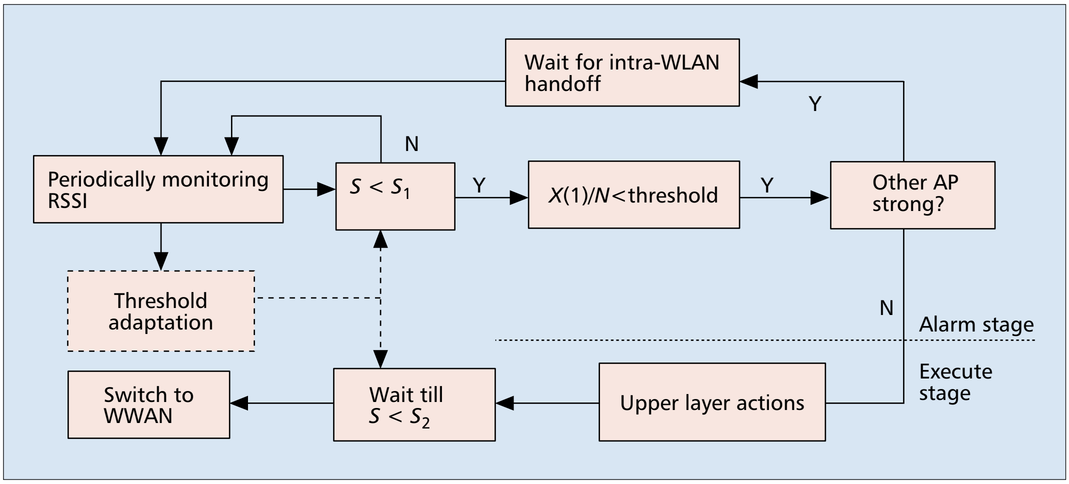
To address these two problems, an FFT-based approach is proposed to detect signal decay. Meanwhile, an adaptive configuration scheme is proposed to set an appropriate RSSI threshold.

FFT-Based Decay Detection — To detect signal decay quickly and accurately, we propose a signal decay detection approach referring to the following property of FFT: the fundamental term of the FFT of a statistically decreasing sequence $x(n)$ with length N always has a negative imaginary part. That is,

$$E\left[X(1)=\sum_{n=0}^{N-1} x(n)\sin(-\frac{2\pi n}{N})\right]<0.$$

Based on this property, we can set a threshold for $X(1)/N$. If the result is smaller than this threshold, the signal should be decaying.

The advantages of using FFT to detect the



■ **Figure 3.** *The diagram for the CM in handoff from WLAN to WWAN.*

Transparent to application is very important, especially for those UDP applications that use the IP address and port number of the received packet to identify which user this packet belongs to.

signal decay are as follows. First, the FFT fundamental term is much more sensitive than the normal threshold-based method since the first $N/2$ signals should have a large difference between the second $N/2$ signals even with variation. Second, if we regard $\sin(-2\pi n/N)$ as a linear filter applied to the sequence $x(n)$, it is obvious that $X(1)$ is the smoothest metric because $\sin(-2\pi n/N)$ is the filter with the least high-frequency component. This will reduce the variation of $X(1)$ even if $x(n)$ varies severely.

Adaptive Threshold Configuration — As different manufacturers have different implementation techniques, different WLAN cards may have different performance. Therefore, it is necessary to set the threshold for a weak signal adaptively according to each card's performance. In this work we apply a *max-min* method to approach a card's performance, which is described as follows:

Step 1. When sampling the RSSI, record the current RSSI if the association with WLAN is valid.

Step 2. If the sampled RSSI < S_2 for some duration (e.g., 1 s), update S_2 with the maximum RSSI within this time duration. Correspondingly, set $S_1 = S_2 + \Delta$, where Δ is a margin.

The handoff procedure from WLAN to WWAN is depicted in Fig. 3, which is a two-stage procedure. When the user is in WLAN, the RSSI is sampled periodically. Once the RSSI is less than S_1, intensive sampling will begin. If the RSSI is consistently less than S_1 for a period of time, we perform FFT and compare with the threshold. If $X(1)/N<$ threshold, it means the currently associated AP is weak and will lose connection. Then we will check if there are other strong APs nearby. If yes, we will simply wait for an intra-WLAN handoff. Otherwise, we need to pass the WLAN_WEAK message to a decision maker and switch to WWAN when finally the average RSSI is less than S_2. In our implementation, the initial value of S_1 for adaptive threshold configuration is set to –76 dBm and Δ = 6 dBm. In addition, the number of RSSI samples in FFT calculation, N, is set to 60, and the RSSI sampling interval is 100 ms.

Virtual Connectivity Manager

It is known that when a mobile host moves, its IP address may change, and the existing connections will be broken. By using the end-to-end approach, which is network-infrastructure-independent, mobile users can maintain their ongoing connections even if the physical connections have been changed. By this method, the connections between peers are virtualized. The basic end-to-end operation of the *virtual connectivity manager* (VC) can be illustrated by the following example. Suppose peer A is communicating with peer B via GPRS provided by ISP_1. After A moves into a WLAN provided by ISP_2, it will notify B of its new connection information (e.g., IP address and port number) directly. Then they can continuously keep the previous connection using WLAN, which has much better bandwidth than GPRS. Moreover, in a VC, an S/N service is further introduced to overcome the problems brought by NAT and simultaneous movement. VC can address the following problems that are not addressed (or not fully addressed) by previous end-to-end approaches:

Transparency to application: By transparency to application, we mean that an upper-layer application should be unaware of changes of address, port, and routing-related information. Transparency to application is very important, especially for those UDP applications that use the IP address and port number of the received packet to identify to which user this packet belongs. We use the following example to illustrate this problem. Suppose mobile host A is communicating with server B; the IP address of A is IP_A. The server program at B buffers the address (and port) of A and uses it as the identity of A in the program. When A moves to a new place and gets a new IP address, say, $IP_{A'}$, the following two things happen:

- When A sends a packet to B using the new IP address $IP_{A'}$, B cannot know it is a packet from A; thus, the packet cannot be processed correctly.
- When B sends a packet to A, it still uses the old IP address IP_A. Hence, the packet will be delivered to the wrong receiver.

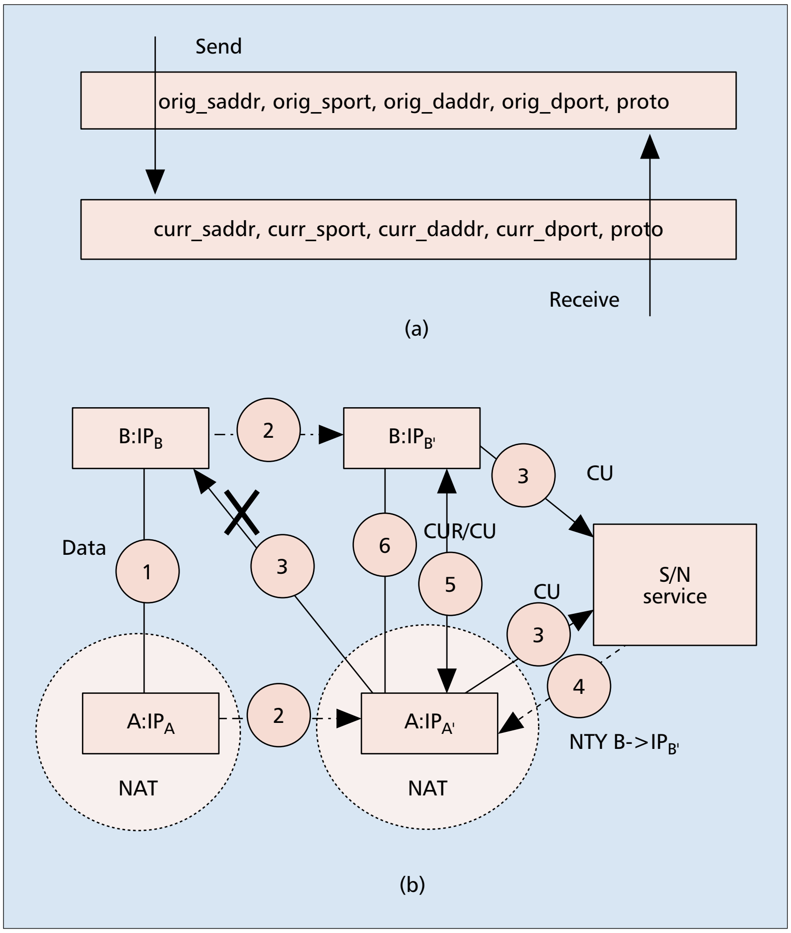
The technique we use to solve this problem is LCT, which will be described later.

Transparency to NAT: We note that NAT may break the end-to-end connection under the following scenario: suppose hosts A and B are communicating; host A is behind a NAT box, and B is in the public domain. If B moves to a new place and gets a new IP address, it will not be able to notify A of its new connection information due to the separation of the NAT box, so the connection breaks.

Simultaneous movements: We also note that connectivity may be broken with simultaneous movements. If mobile hosts A and B move to new network attach points simultaneously, they cannot know the exact address of their peer. Therefore, the update information cannot be sent to the right place.

To solve the problems of NAT and simultaneous movements, a third party service, S/N service is proposed in this work.

Besides LCT and S/N service, we also design a lightweight VC protocol to exchange connection-related information between the end hosts and between the end host and the S/N service. This VC protocol includes several key components: *peer negotiation* is used by two peers to agree on things, such as shared key and original connection information, for secure and accurate mobility management before mobility events happen; *connection maintenance* is used to exchange connection information between peers when mobility events occur, and an *S/N protocol* is designed for the end host to exchange subscribe/notify messages with the S/N service. Due to space limitations, we cannot describe them in detail. In the following, we describe how the LCT and S/N service work.



■ **Figure 4.** *Two key components of VC: a) local connection translation; b) subscription/notification service.*

Local Connection Translation

Local connection translation (LCT) is the technique we use to address the application transparency problem. The idea is to maintain a mapping relationship between the original connection information (e.g., IP address and port number) and the current connection information for each active connection, as illustrated in Fig. 4a. The original connection information does not change during the connection's lifetime, while the current connection information changes each time the host or peer gets a new IP address. The upper layer applications only see the original connection information. Therefore, mobility is made transparent to the upper-layer applications.

The following actions are performed by the VC when the host sends and receives packets:

- Packet sending. When an application sends a packet, the VC looks up the LCT table to substitute the original connection information for the current information, and then delivers the packet to the lower layer.
- Packet receiving. When the VC receives a packet from outside, it also looks up the LCT table to substitute the current connection information of the packet for the original packet information, and then delivers the packet to the upper layer.

The mapping item is created when a new connection is established, and the item is updated when the source and/or destination hosts move to new network access points.

The Subscription/Notification Service

The subscription/notification (S/N) service is introduced to overcome problems (i.e., under NAT and simultaneous movement) that cannot be solved by traditional end-to-end approaches. The idea is to introduce a third-party S/N service to bridge the two communication parties so that they can exchange address information via this S/N service to resume their otherwise broken end-to-end connections. We assume an S/N server is publicly addressed and never moves.

Using the previous NAT scenario, both A and B are connected to an S/N server, and A subscribes to B's IP address changes via the S/N server. If B (the host with a public address) moves to a new network AP, it will inform S/N of its new address; then S/N notifies A of this new address of B. In this way, A can resume the connection with B.

As to the simultaneous movement scenario, A and B subscribe to the IP address changes of each other; therefore, when A and B move simultaneously, they inform S/N of their new IP addresses, and then the S/N server notifies A and B of the new address of each peer. After that, A and B can resume the connection.

Here we use the example illustrated in Fig. 4b to show how S/N works under NAT and simultaneous movement scenarios. Note that both A and B maintain a connection with the S/N server. The connection between A and B is maintained using the following steps.

1. Mobile hosts A and B are communicating with each other, A with private address IP_A and B with public address IP_B.
2. A and B move simultaneously and get new IP addresses $IP_{A'}$ and $IP_{B'}$, respectively.
3. A and B send a connection update (CU) message to S/N to update the connection with S/N, and A also sends a CU to IP_B to update the connection with B; this message will be lost since B has moved away (B does not send a CU to IP_A since A is privately addressed).
4. S/N notifies A of B's new IP address via a notify message (NTY B->$IP_{B'}$).
5. A issues a connection update request (CUR) message to B's new IP address and B replies with a CU message to A's new IP address.
6. The connection between A and B is resumed.

PERFORMANCE EVALUATION

To demonstrate the feasibility and effectiveness of our approach, we have built a prototype system based on the architecture illustrated in Fig. 1 in the Windows 2000 operating system. In our implementation, a CM manipulates and monitors all wireless network interfaces via the NDIS device interface, and provides related information (WLAN_WEAK and WLAN_AVAILABLE, available bandwidth, delay, etc.) to the *roaming decision maker*, which runs in the system as a background service. The decision maker then triggers the VC to perform connection maintenance functionalities when the handoff condition is satisfied. The VC is implemented in the system together with the TCP/IP stack, and is located naturally between the network and transport layers.

We use the following experiment to demonstrate the ability of our system to maintain a connection's continuity and achieve better performance (higher TCP throughput in this case) than a traditional scheme. The network setup for our experiment is shown in Fig. 5. Specifically, there are two 802.11b APs connected to the same local network segment. Mobile host A is equipped with an IEEE 802.11b WLAN card and a GPRS card, whereas desktop PC B has an Ethernet interface. The GPRS network (provided by China Mobile) is always on. That is, host A can always maintain the connection with host B via the GPRS network even if the WLAN is not available. There is an always-on TCP session between A and B to transmit data as fast as possible. For comparison, we also implement the traditional handoff approach. In the traditional approach (TA) as used in [10], a handoff from WWAN to WLAN will be triggered when the WLAN RSSI is above than a threshold; likewise, a handoff from WLAN to WWAN will be triggered once the WLAN RSSI is lower than a threshold. According to the IEEE 802.11b standard requirement, we fix the threshold in the TA to be –76 dBm. During the experiment, host A first moves around within the WLAN area covered by the two APs and then steps out of the WLAN. The connection may be switched between AP_1 and AP_2 in WLAN at the link layer or between WLAN and GPRS at the network layer. The switch at the link layer is transparent to the VC and does not trigger VC actions. We compare the performance of the traditional solution (TA + VC) and our solution (CM + VC). In TA+VC, even when A moves around within the WLAN, several vertical handoffs between WLAN and WWAN are triggered due to the severe RSSI fluctuation and slow intra-WLAN handoff. Figure 6 shows the TCP throughput comparison of the connection between A and B under the two approaches.

During the whole testing period, TCP connection is maintained successfully under both cases due to the effect of the VC. From Fig. 6 we observe that by using CM+VC, much higher throughput can be obtained since, with the help of CM, a handoff decision can be made more accurately. In this experiment the TA triggers five vertical handoffs (three for WLAN to GPRS

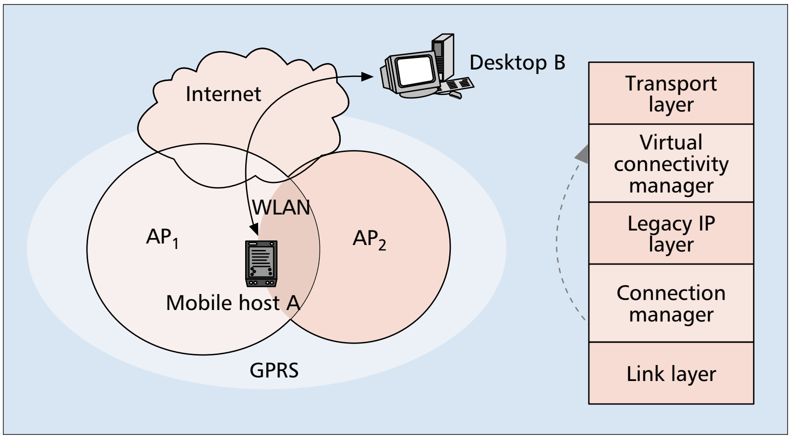


Figure 5. *Network setup of the experiment.*

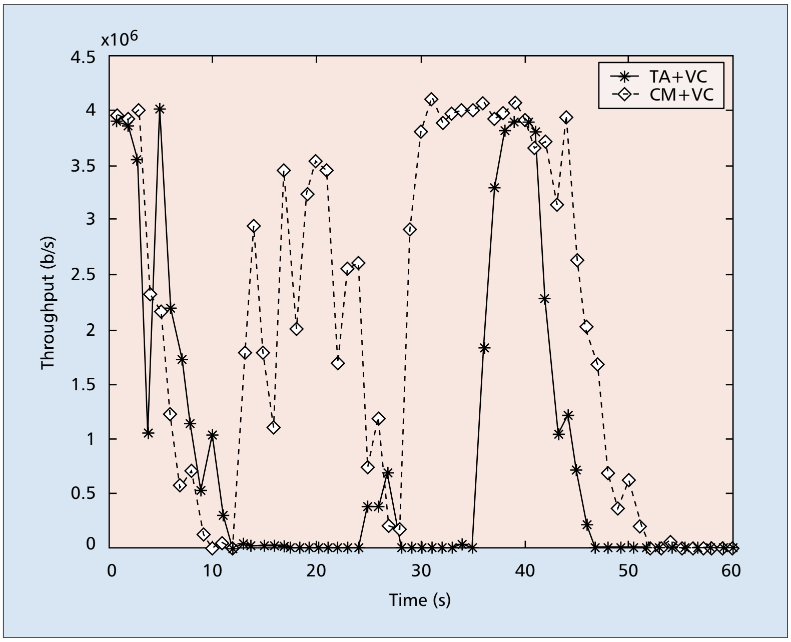


Figure 6. *Throughput comparison of the TCP connections.*

and two for GPRS to WLAN), while the CM only triggers one vertical handoff when finally stepping out of WLAN. The mean TCP throughputs are 2.5 Mb/s and 1.1 Mb/s under CM+VC and TA+VC, respectively. Note that the GPRS network has a low bit rate, and only achieves approximately 20 kb/s throughput for TCP. Moreover, due to the low RSSI and unknown interference, sometimes the transient throughput may also be low even in WLAN, but still larger than that in GPRS.

Our prototype system demonstrates that seamless roaming between WWAN and WLAN can be achieved and much higher throughput can be obtained compared to the traditional scheme.

Conclusions

In this article a novel mobility management system is proposed for vertical handoff between WWAN and WLAN. The system integrates a *connection manager* that intelligently detects the wireless network changes and a *virtual connectivity manager* that maintains connectivity using the end-to-end principle. Collaboration between the CM and VC accomplishes seamless handoff between WWAN and WLAN. More specifically, in the CM, MAC sensing, FFT detection, and adaptive threshold configuration algorithms are proposed to significantly reduce the handoff rate and ping-pong effect. In the VC, an end-to-end connection maintenance mechanism is used to make it independent of additional network infrastructure. Moreover, local connection translation and subscription/notification service introduced in the VC provide advantages of application transparency, working under NAT, and successful handling of simultaneous movement. Our prototype system demonstrates that seamless roaming between WWAN and WLAN can be achieved, and much higher throughput can be obtained than with the traditional scheme.

References

[1] G. Corazza, D. Giancristofaro, and F. Santucci, "Characterization of Handover Initialization in Cellular Mobile Radio Networks," *Proc. IEEE VTC '94*.

[2] M. Gudmundson, "Analysis of Handover Algorithms," *Proc. IEEE VTC '91*.

[3] K. Pahlavan *et al.*, "Handoff in Hybrid Mobile Data Networks," *IEEE Pers. Commun.*, Apr. 2001.

[4] J. McNair, I. Akyildiz, and M. Bender, "An Inter-System Handoff Technique for the IMT-2000 System," *Proc. IEEE INFOCOM 2000*.

[5] C. Perkins, Ed., "IP Mobility Support for IPv4," IETF RFC 3344, http://www.ietf.org/rfc/rfc3344.txt.

[6] C. Pekins *et al.*, "Mobility Support in IPv6," http://www.ietf.org/html.charters/mobileip-charter.html

[7] R. Ramjee *et al.*, "HAWAII: A Domain-Based Approach for Supporting Mobility in Wide-Area Wireless Networks," *IEEE/ACM Trans. Net.*, vol. 10, no. 3, June 2002.

[8] A. C. Snoeren and H. Balakrishnan, "An End-to-End Approach to Host Mobility," *Proc. Mobicom 2000*.

[9] IEEE Standard for Wireless LAN-Medium Access Control and Physical Layer Specification, part 11, 1999.

[10] EURESCOM Project P1013-FIT-MIP, "First Steps towards UMTS: Mobile IP Services, A European Testbed," http://www.eurescom.de/public/projects/P1000-series/p1013/default.asp