

Handoff Management for Mobile Devices in Hybrid Wireless Data Networks

Riaz Inayat, Reiji Aibara, and Kouji Nishimura

Abstract: Today's wireless access networks consist of several tiers that overlap each other. Provisioning of real time uninterrupted communication to mobile users, anywhere and anytime through these heterogeneous overlay networks, is a challenging task. We extend the end-to-end approach for the handoff management in hybrid wireless data network by designing a fully mobile-controlled handoff for mobile devices equipped with dual mode interfaces. By handoff, we mean switching the communication between interfaces connected to different subnets. This mobile-controlled handoff scheme reduces the service disruption time during both horizontal and vertical handoffs and does not require any modification in the access networks. We exploit the IP diversity created by the dual interfaces in the overlapping area by simultaneously connecting to different subnets and networks. Power saving is achieved by activating both interfaces only during the handoff period. The performance evaluation of the handoff is carried out by a simple mathematical analysis. The analysis shows that with proper network engineering, exploiting the speed of mobile node and overlapping area between subnets can reduce service disruption and power consumption during handoff significantly. We believe that with more powerful network interfaces our proposal of dual interfaces can be realized.

Index Terms: Dual-mode interfaces, hybrid network, mobile IP, seamless handoff, ubiquitous computing.

I. INTRODUCTION

Ubiquitous mobile computing is a coming reality, fueled in part by continuing advances in wireless transmission technologies and handheld computing devices. The systems and technologies that can adapt to the dynamic heterogeneous environment are the main force behind this drive toward ubiquitous wireless communications [1]. The future scenario for global connectivity, where a mobile host could have network connection anywhere and anytime at any device, requires from the mobile hosts to be more intelligent and informative about their environments. Simultaneous network connection capability in mobile network architectures, i.e., a mobile host using multiple network interfaces connected to different networks at the same time, is useful for load balancing, redundancy, and disaster recovery. We believe and show in this paper that the use of this IP diversity is not limited to these applications only. With the widespread deployment of wireless infrastructure, this capabil-

ity can be exploited to achieve high mobility, too.

The paradigm of ubiquitous networking, i.e., to be and to remain connected anywhere at anytime on any device, has focused the research on enabling mobile networking across an extremely wide variety of real world networks and mobile devices. The fundamental issues that restrict ubiquitous networking keep getting more complicated with wide variety of devices having different sizes and constraints, disparate wireless links and technologies, highly mobile environments, ever demanding user applications, and Internet protocols that were never meant for this [2]. Ubiquitous connectivity is definitely real to users however; transparent mobility is certainly a challenge. Though with the current technological advancements, a user can access information on a range of devices (e.g., laptops, PDAs, cell phones, etc.) with a range of access technologies (e.g., W-CDMA, GPRS/3G, Bluetooth, PHS, WLAN, Satellite, etc.), but true seamless roaming in this hybrid environment demands transparent integration between diverse network systems, applications, and services. In particular, for this to happen, it demands seamless intra-network and inter-network handoff mechanisms with existing connectivity as devices continue to move across various environments, while still minimizing any disruption to ongoing flow during switchovers [3]. A handoff mechanism that enables this has to exhibit a low handoff latency, incur little or no data loss (even in highly mobile environments), scale to large internetworks, adapt to different environments, and act as a juncture between heterogeneous environments and technologies without compromising on key issues related to security and reliability.

As advances are being made in the different standardization areas (IETF, 3GPP, 3GPP2, and ITU) to define global mobile network architecture, it is becoming obvious that the core network of the next generation mobile network will be purely IP based. Internet protocol's most recent version added with mobility support (MIPv6) [4], [5] has recently gained a lot of interest as a solution for global mobility. A number of research projects in the last few years [2], [6]–[8] have worked on overlay networks with IP as core network protocol. The results show that handoff management is an essential component of mobile communication in these overlay networks. Handoff is generally defined as a process or mechanism of switching coverage responsibilities between respective access nodes. At data link layer, this may cause inter-technology (vertical) or intra-technology (horizontal) handoff. Handoff can also be categorized as break-before-make (hard) or make-before-break (soft). Hard handoffs are usually very simple but for reducing packet loss, they require buffering of packets in the access nodes. This introduces delay and jitter during handoff. Hard handoff also suffers from ping-pong effect [9]. On the other hand, though the soft handoffs reduce delay and jitter as well as require less power but result in

Manuscript received November 18, 2003; approved for publication by Wha Sook Jeon, Division III Editor, July 26, 2004.

R. Inayat is with the Graduate School of Engineering, Hiroshima University, Japan, email: riaz@hiroshima-u.ac.jp.

R. Aibara and K. Nishimura are with the Information Media Center, Hiroshima University, Japan, email: {ray, kouji}@hiroshima-u.ac.jp.

higher complexity of the network system architecture. The vertical handoff may involve switching of network interfaces [10], [11] or switching of access nodes [12] in a single interface at the mobile node. These schemes require major modifications in the overlay infrastructure, especially in the access networks, which is obviously very problematic, as the infrastructure is usually not in the jurisdiction of a single party. This modification will become more problematic with the introduction of new access technologies and protocols. Therefore, instead of waiting for the network to be adaptive to the dynamic environment it is more feasible and practicable to make the mobile node more adaptive to various disparate access technologies. However, as power drained by the network interface constitutes a large fraction of the total power consumed by the mobile device [13], using multiple air interfaced devices for handoff demands a proper mechanism for power saving.

We believe and show in this paper that with dual interfaces, even without modifying the overlay infrastructure, service disruption can be minimized during vertical and horizontal handoffs. With dual interfaces, we can have advantages of both soft and hard handoffs by using a hard handoff mechanism at the data link layer level but switching the communication between already connected interfaces during overlapping area before the data link layer handoff. In the overlapping area, the two interfaces connect to different subnets. This IP diversity can be used to cover up the handoff latencies due to the address acquisition and registration processes as far as mobile IP is concerned [14]. We also devise a network interfaces switching criterion for optimal power consumption. Though we mainly explain and discuss the handoff scheme in the overlay networks that consist of IEEE 802.11 WLAN and GSM based systems (GPRS) but the procedure may also be used for other systems, like Bluetooth, CDMA, or PHS, as the algorithm is transparent to the access technologies.

The rest of the paper is organized as follows. Section II outlines the overlay network characteristics. In Section III, the detailed description of our proposed handoff scheme is given. Section IV discusses some solutions to realize our proposed handoff mechanism. Then Section V describes how the IP macro-mobility protocols can be adopted in the core network to support mobility. The analytical model and performance evaluation with numerical results and considerations in Section VI are followed by conclusions in Section VII.

II. CHARACTERISTICS OF THE HETEROGENEOUS NETWORK

Heterogeneous networks have some distinct properties. Firstly, wide area data networks are usually not owned and administered by a single party. Therefore, we can't directly modify or control the overlay infrastructure.

Secondly, the network's service areas are overlapped. For example, GPRS network acts as an umbrella network to the WLAN network. Therefore, if MN is not under the coverage area of WLAN, then it can be at least reached through GPRS network. Even the different cells in the WLANs are overlapped. We can utilize this overlapping area to reduce service disruption, by simultaneously connecting to different subnets of the

same access technology during horizontal handoff or connecting to the different networks at the boundary of one network during vertical handoff.

Thirdly, the overlaid networks support different data rates. The WLAN can support data rates in Mbps, while GPRS can only supply tens of kbps. Therefore, the priority of handoff is usually from GPRS to the WLAN.

Fourth, because of the heterogeneous networks, the signal powers received from the base stations of the different networks are incomparable.

Fifth, due to IP layer mobility protocols, whenever there is a change in IP address of the mobile node, the correspondent nodes or mobility agents have to be informed about this new MAdr. During this address registration process the interface becomes unavailable for the communication that causes service disruption. IP layer handoff usually lags data link layer handoff. We can compensate this by using two interfaces.

Furthermore, the power consumed by the interfaces is directly related to the transmitted power. GPRS's transmitted power is higher than that of WLAN. Thus if two different interfaces are used to facilitate roaming in heterogeneous networks then the power consumption is very significant if the GPRS interface is left activated for long time durations.

III. DUAL-INTERFACE HANDOFF MECHANISMS

Handoff is generally defined as a mechanism of switching coverage responsibilities between respective access nodes. *Vertical (inter-technology) handoff* involves communication switching between two physically separate network interfaces or mode switching on a single multi-mode interface, whereas *horizontal (intra-technology) handoff* only involves a single air interface. In this paper, by horizontal as well as vertical handoff, we mean switching of communication between two separate air interfaces that are connected to different access nodes. To increase the overlapping area we prolong the data link layer handoff on a single interface until the threshold is less than the carrier detect threshold. In this paper, we discuss only those handoffs that are accompanied by change of IP address, i.e., between subnets for WLAN and between GPRS systems owned by different cellular operators.

Keeping in view the properties of the heterogeneous networks stated in previous section, we propose a fully mobile-controlled handoff by using two interfaces with common control. Our proposal does not require any modification of the overlay infrastructure. Coverage area of an access node is generally divided into normal region and handoff region. We further divide the handoff region into two distinct regions as horizontal and vertical handoff regions. While roaming in the overlay networks, the MNs follow the following policy.

- Priority of the handoff is to the high-bandwidth access technology, i.e., to the lower overlay network.
- When the MN is in the normal region, only one interface will be in active mode. Second interface will be in OFF state. However, the MN through this second interface periodically searches for the high bandwidth network for downward vertical handoff.
- In the horizontal handoff region, the MN searches for an-

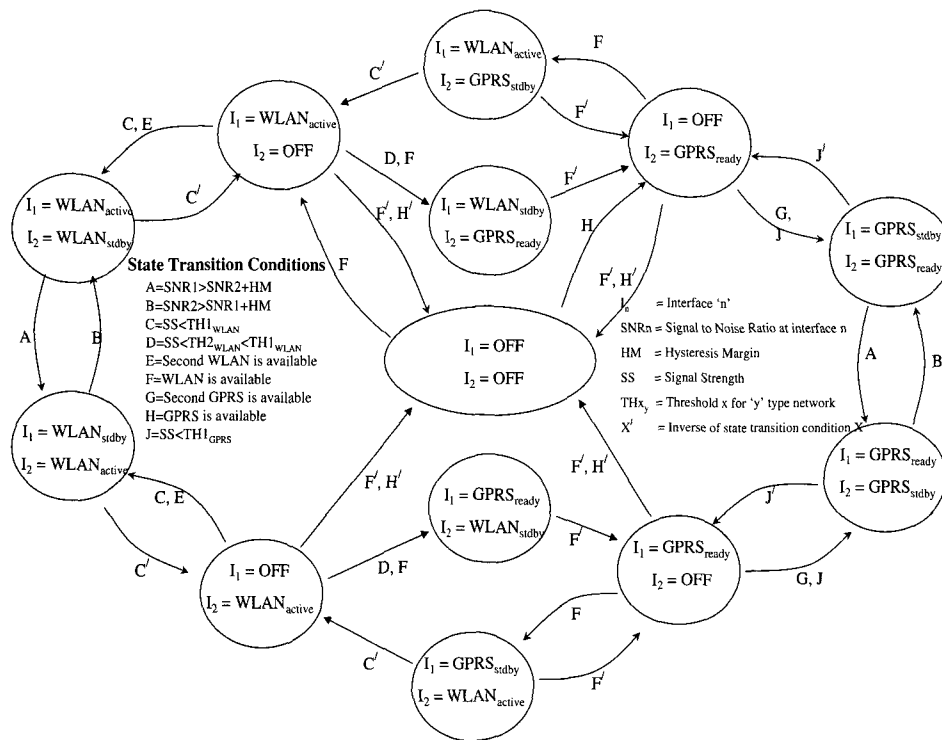


Fig. 1. Mobility management state model of mobile node multi-mode interfaces while communicating.

other subnet of the same access technology, and if available, connects the second interface to such subnet in standby mode. The interfaces change their states depending upon their signal to noise ratios (SNR).

- If the MN reaches the vertical handoff region without finding a second subnet of the same access technology, it looks for the higher low-bandwidth network and connects to such network in active mode. The first interface switches to the standby mode and consequently to OFF state.

We can implement this idea with multi-mode interfaces by using software radio [15]. However, software radio is neither mature nor widely available. Fortunately, even nowadays we have a number of dual-mode network interfaces, like WLAN/CDMA, WLAN/GPRS, and GPRS/UMTS that are widely available and cover a large geographical area. To elaborate this policy further now we assume an overlay network consisting of WLAN and GPRS. Due to high-bandwidth support, the priority of handoff is given to WLAN. To differentiate between the handoff regions, we use two thresholds ($TH1_{WLAN}$, $TH2_{WLAN}$) of received signal strength (SS). We prefer to use SS as the metric rather than SNR, which is prone to random fluctuations due to noise. $TH1_{WLAN}$ is greater than $TH2_{WLAN}$. While in horizontal handoff region, when SS is between these thresholds, only WLAN subnet is scanned and connected. Whereas if the $SS < TH2_{WLAN}$ (vertical handoff region), then the MN looks for the GPRS, too. According to the fourth property as stated in previous section, we do not compare the SS of different networks. As the priority is to WLAN, so the vertical handoff decisions are also based on the WLAN thresholds. Only for the horizontal handoff between GPRSs, we use GPRS threshold (TH_{GPRS}). We switch interfaces from active to idle state

for optimal power consumption. The two interfaces are active only in handoff period. Interfaces can be in OFF, WLAN active, WLAN standby, GPRS ready, and GPRS standby states. The mobility management state model of mobile node's interfaces is given in Fig. 1. It should be noted that this model presents the states of dual interfaces. The state transition conditions are based on the signal strength received by the dual interfaces of mobile node. A more sophisticated access selection algorithm [16], based on network, radio and user defined parameters as perceived by a single interface, can be used to complement this interface switching criteria. In the state model depicted in Fig. 1, by OFF state, we mean that the interface is not attached to any network. GPRS ready means that the MN is updating its location every time it changes the cellular cell. This is to avoid the paging delay to find the location of GPRS. MN is in GPRS ready state when it is transmitting data through GPRS network. GPRS standby means that the GPRS is sending update only when it changes a larger routing area (area consisting of a number of cells). This is to save battery power. When the interface is in GPRS ready or WLAN active states it can receive as well as transmit the data whereas when it is in GPRS standby or WLAN standby states, the interface is only in receiving mode. It can't receive or transmit data when in OFF state.

Interfaces follow the following policy.

- If the MN is in the coverage area of two WLAN subnets and $SS > TH1_{WLAN}$, then only one interface will be connected to the WLAN and other will be OFF. However, if $SS < TH1_{WLAN}$, then both interfaces will be connected with different subnet prefixes. One of the interfaces will be in standby mode.
- If the MN is in the coverage area of only one WLAN subnet

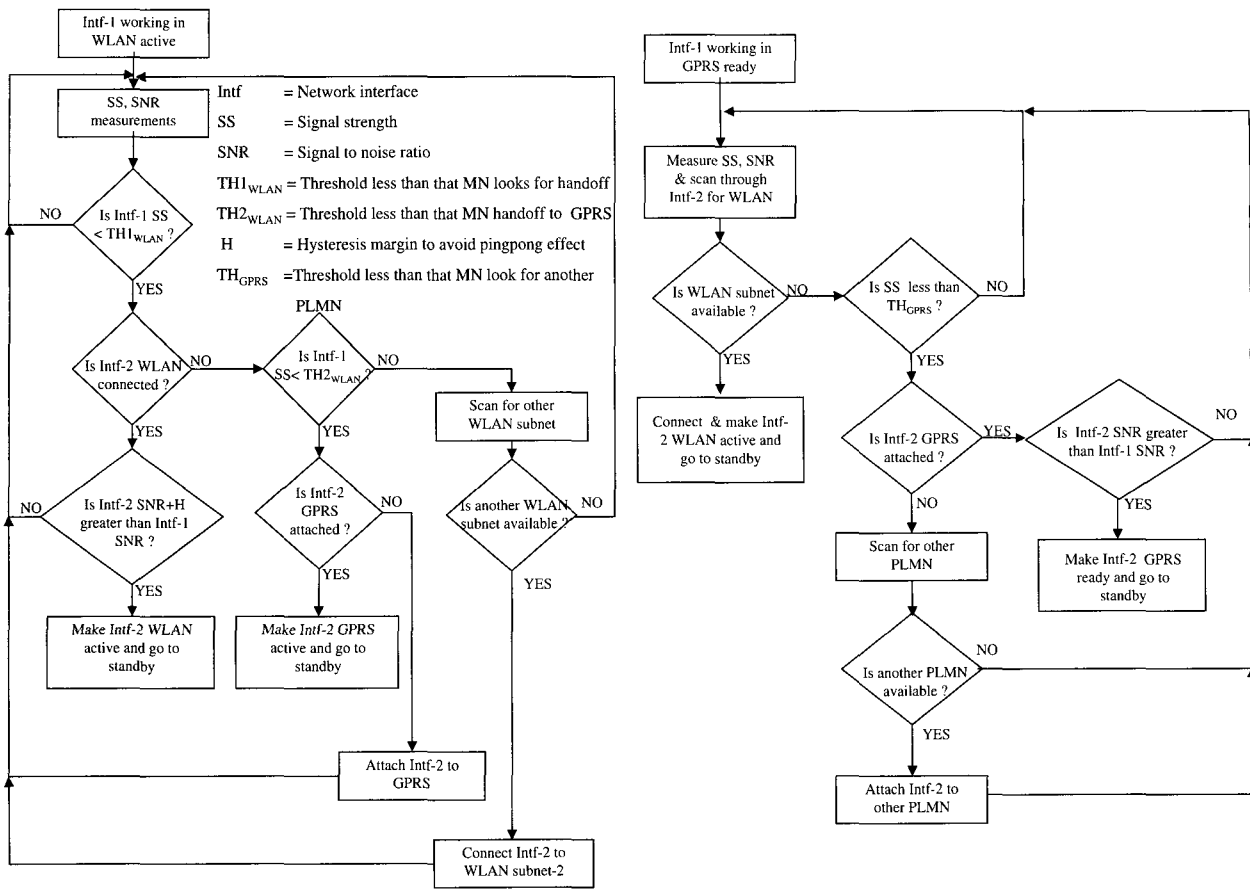


Fig. 2. Handoff algorithm for an interface in WLAN active and GPRS ready mode.

and $SS < TH2_{WLAN}$, then one interface will be connected to the WLAN in standby mode and other will be connected to the GPRS in ready mode.

- If no WLAN is available and only GPRS is available and $SS > TH_{GPRS}$, then one interface is connected to the GPRS in ready mode and other will be OFF. In this case, the MN scans periodically for WLAN through this interface. But when $SS < TH_{GPRS}$, then MN will also search for another GPRS system. By another GPRS, we mean another public land mobile network (PLMN). Intra-GPRS handoff is handled at data link layer.

It should be noted that by handoff we mean to switch the communication from one interface to another. When the direction of communication is from MN (sender) to the correspondent node (CN), then the MN itself switches the communication between interfaces. However, when the communication is from CN (sender) to the MN (receiver), then the handoff is achieved by registering the new MAdr to the HA and CN in mobile IP or by changing the address priorities at IMS in MAT. The previous interface remains connected to the previous subnet as long as possible but in receiving only mode. Therefore, both interfaces are accessible in handoff area.

Handoff algorithms from active states are given in Fig. 2. horizontal handoff follows the following events

- MN having interface-1 connected to the subnet-A and with $SS < TH1_{WLAN}$ enters the handoff area.

- The interface-2 obtains a mobile address (MAdr) from subnet-B. Interface-2 goes to standby mode. In case of MAT, the MN registers this new MAdr as priority-2 to the IMS and alerts corresponding nodes.
- When the $SS < TH2_{WLAN}$ and SNR from subnet-B plus some margin becomes more than that of subnet-A, MN registers the subnet-B address to the HA and CN (or to IMS as priority-1 address in MAT). Interface-2 becomes active and interface-1 goes to standby mode.
- MN interface-1 losses its connection with subnet-A and goes to IDLE mode.

For vertical handoff from WLAN to GPRS the procedure is as follows.

- When the active interface-1's received SS crosses the threshold $TH1_{WLAN}$, it signals the other interface to search for an available access network.
- Interface-2 tries to find another WLAN subnet. If succeeded, it follows the procedure as outlined above for horizontal handoff, else it waits for the SS to cross second threshold $TH2_{WLAN}$.
- When the SS falls below $TH2_{WLAN}$, interface-2 is attached to the GPRS and the IP address of the GPRS interface is registered with the CN and HA/IMS. Interface-1 goes to the standby mode.
- When interface-1 loses its connection to the WLAN it goes to OFF state.

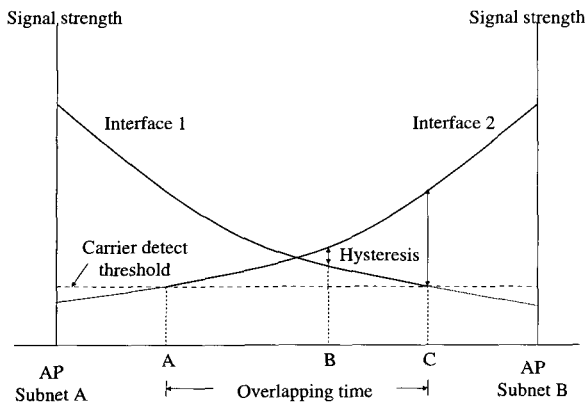


Fig. 3. Handoff thresholds.

For vertical handoff from GPRS to WLAN, the procedure is as follows.

- With one interface connected to the GPRS, the MN, through other interface, periodically searches for a WLAN. As soon as MN finds a WLAN access signal, it connects to the WLAN subnet and gets a new IP address from that subnet.
- When the SS becomes more than TH_{2WLAN} , interface-2 switches to active mode.
- Consequently GPRS interface switches to the standby mode and then to OFF mode.

IV. HANDOFF MECHANISM REALIZATION

In order to effectively realize our dual interface handoff scheme, overlapping area between subnets is required. For vertical handoff, overlapping area always exists as the different access networks are overlaid on one another but for horizontal handoff, sometimes the cells are disjoint. Moreover, within this overlapping area it is required that the two interfaces should connect to two different subnets as long as they can listen their respective APs. It means while in the overlapping area they should connect to different subnets. We propose the following solutions to these issues in order to realize dual interface handoff schemes.

A. By Controlling Link-Layer Handoff Thresholds

We can create overlapping time by having different handoff thresholds for two interfaces. For example, as depicted in Fig. 3, data link layer handoff does not take place at point B. Only the communication is switched between two interfaces at this point B. Interface-1 remains connected with the previous subnet. This could be achieved by changing the hysteresis values. As depicted in Fig. 3, we can have overlapping time equivalent to AC instead of AB by increasing the hysteresis values.

B. By Cross-Layer Interaction

We can also create an overlapping time if the upper layers can utilize the physical layer information to trigger the horizontal handoffs. In this case, the interfaces can also be forced to connect to different subnets at different times just by controlling the link layer handoff timings. In WLAN, by assigning unique service set identifiers (SSID) to the APs across the boundary of

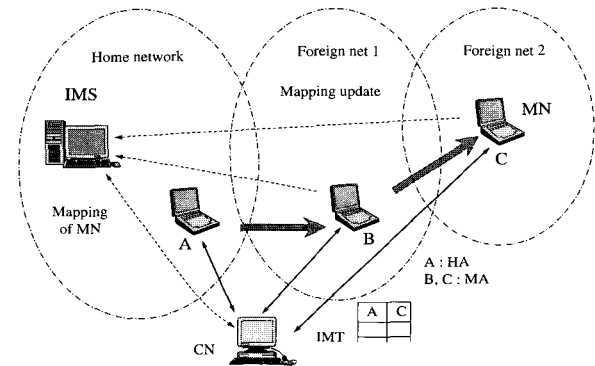


Fig. 4. Communication in MAT.

subnets, MN can also force wireless network to keep association with a specific AP. The handoff mechanism becomes more effective if the MN can distinguish the boundary cells of a wireless system from the intermediate cells.

C. By Using Software Defined Radio (SDR)

Dual interface handoff can be implemented by using SDR [15]: A radio with extensive programmability. Recently, a greater degree of programmability is being made available on devices that are nearly as inexpensive as fixed function devices, and software infrastructure to control these are emerging [1]. A number of companies are making baseband processor and chipsets that can carry out WCDMA and GSM, or CDMA2000, CDMA1, and AMPS, or 802.11b, WCDMA, and GSM. With the maturity in SDR technology, multimode interfaces will be widely available that will make our proposal realizable.

V. ADOPTING MACRO-MOBILITY PROTOCOLS IN WIRELESS IP NETWORK

As advances are being made in the different standardization areas (IETF, 3GPP, 3GPP2, and ITU) to define global mobile network architecture, it is becoming obvious that the core network of the next generation mobile network will be purely IP based. A number of IP based mobility protocols [4], [5], [17]–[19] can be adopted in wireless environments. Even though MIPv6 [5] offers a number of benefits, the signaling overhead introduced in the network load can become significant, at times, and the handoff process can be long. Hierarchical MIPv6 [17] focuses on local movement to reduce the signaling load but requires modification in the local access network. Another solution for supporting mobility in the IP networks is mobile IP with address translation (MAT) [18], [19]. In MAT, at application and transport layers, connections are established with permanent addresses of mobile nodes, but the addresses are translated into foreign addresses at IP layer at the end nodes. The end nodes request location information of correspondent nodes to special database servers called IP address mapping servers (IMS) and cache them locally. Communication in MAT is depicted in Fig. 4. The main difference in the MIPv6 and MAT is in the implementation of respective mobility agents, home agent (HA), and IP address mapping server (IMS). In contrary to HA

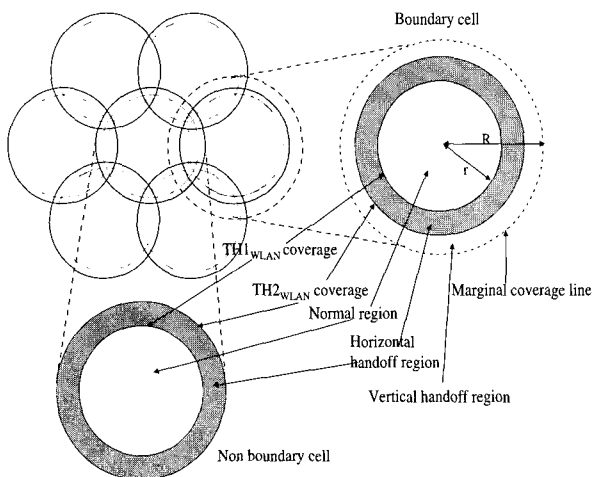


Fig. 5. WLAN coverage model.

in MIPv6, IMS in MAT can be placed at distributed locations in the Internet. Secondly, IMS stores multiple IP addresses of MN with priorities that helps to reduce service disruption [14]. MAT does not require route optimization as that of MIPv6 as the data communication is always between peer nodes. These characteristics of MAT help to reduce handoff latencies significantly.

In heterogeneous wireless environment, at the core network level, mobility protocols that are based on end-to-end abstraction are more suitable. We have chosen MIPv6 and MAT as a core network protocol because of their features that are useful for dual interface handoff. Such as in MAT, IMS stores multiple IP addresses of MN with priorities. This feature with dual interface handoff helps to reduce service disruption [14]. In GPRS [20] for mobile IP as defined by 3GPP [21], [22], the foreign agent (FA) is located in the core network in GGSN. The HA may or may not be located in the GSM network. We can also implement FA at SGSN [23]. Though this reduces the burden of GGSN but increases control traffic due to increase in binding updates to HA. This also increases number of IP handoffs as the IP address of the MN is changed when it moves between different SGSN. The FA is configured with care of address (CoA) and it maintains a list that maps IP addresses with the identities of all the visiting MNs that have registered with FA. IP packets destined for the MN are intercepted by the HA and tunneled to the FA. The FA de-tunnels the packets and forwards the packets to MN.

Adoption of MAT [18], [19] for supporting mobility between GPRS and WLAN is even easier. Being designed on end-to-end abstraction MAT does not require any modification in the core IP network. It only requires that GGSN acts as an IP router. IMS can be placed at any place in the Internet or in the GSM network. IMS is only a location database server, not a packet interceptor and packet forwarder like HA in mobile IP. Therefore, when a MN moves to a GPRS network, which to itself is a foreign network, GGSN assigns a mobile address (MAdr) to the MN. MN registers this MAdr to the corresponding IMS. The CNs get the new address from the IMS and send the packets directly to GGSN, which delivers them to the MN.

VI. PERFORMANCE EVALUATION OF DUAL INTERFACE HANDOFF

A. Coverage Model

The coverage area of 802.11 access point (AP) can be divided into four different areas. First is the cell area that has a strong SS and results in no packet loss. We refer it as 'effective' coverage area. The second is the 'marginal' coverage area. The marginal coverage area corresponds to the area outside the effective area, which has a very low packet loss (below 5%). This marginal coverage is at least equal to that of the effective radius when SS exhibits negative log function behavior [24]. The third area representing the remaining area is 'poor' coverage area. The fourth area is logical area inside the effective area referred to as 'good' area. In theory, for any overlapping set of three APs, the good coverage areas of these APs intersect at a single point. For our analysis, we assume that the threshold value $TH2_{WLAN}$ is within the effective coverage area and $TH1_{WLAN}$ is within good coverage area. As depicted in Fig. 5, the region covered by $TH1_{WLAN}$ coverage line is referred in our model as the normal region. The region between $TH1_{WLAN}$ and $TH2_{WLAN}$ is horizontal handoff region whereas the region between the marginal and $TH2_{WLAN}$ lines is referred as the vertical handoff area. As shown in Fig. 5 due to the overlapping of horizontal handoff regions in the non-boundary cells, the vertical handoff is not performed so the vertical handoff area in the non-boundary cell does not have any effect. We define the area ratio 'a' as the ratio of the total handoff area to the cell area.

We assume that all the MNs are randomly moving in the cell and the dwell times of the MN in two distinct regions (handoff and normal region) are exponentially distributed. The mean dwell time in handoff region is denoted by T_{dwell} and in normal region as T_{dwell2} . The relations between the mean dwell time in the cell, T_{cell} and the mean dwell time in each region are a function of the covered area of each region. According to study [25],

$$T_{dwell} = 2^{\log(1-a)} \cdot T_{cell}, \quad (1)$$

$$T_{dwell} = 16^{\log(a)} \cdot T_{cell}. \quad (2)$$

B. An Analytical Model for Dual Interface Handoff

We model the dual interface handoff based on queuing network and assume the network topology as depicted in Fig. 6. The following assumptions are necessary for the computational tractability reasons. All routers are modeled as simple M/M/1 queues. The exponentially distributed service time of a packet includes both the processing time and the transmission time. If service rate of the router R_i ($i = 0, 1, 2, 3, 4$, GGSN, SGSN) is denoted by μ_i and the load by ρ_i , then the response time of router R_i becomes a random variable exponentially distributed with rate $\mu_i(1 - \rho_i)$. Let L_{xy} denote the fixed propagation delay of a link between router x and router y . $dP1$ is the path delay between router R_0 and IMS/HA. $dP2$ is the path delay between router R_0 and CN. $dP3$ is the delay between CN and IMS/HA including all the processing time, propagation and transmission time of the routers on the paths. Similarly, $dP4$ and $dP5$ are the

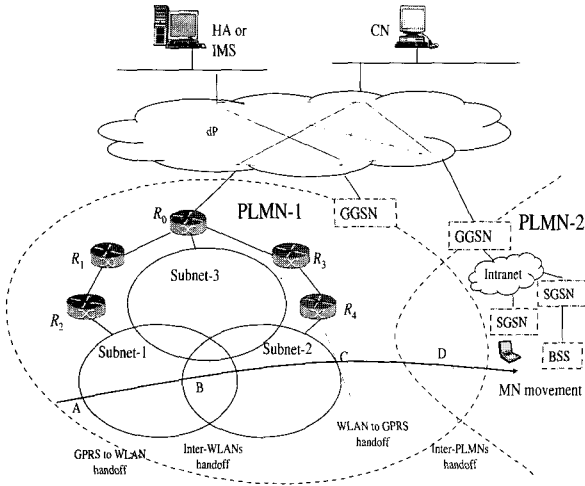


Fig. 6. Various handoffs in WLAN/GPRS network.

path delays between GGSN to IMS/HA and CN, respectively. $dL1_{ma}$ and $dL2_{ma}$ are the mobile address acquisition delays (including address authentication delays) for MN at new subnet in WLAN and GPRS, respectively. Keeping in view the coverage model, reference network topology and handoff procedure as proposed in Section III, we define the following time instants.

- t_n : Time instant when the MN leaves the normal region ($TH1_{WLAN}$ coverage line) of a cell.
- t_e : Time instant when the MN leaves the effective region and enters marginal area of a cell.
- t_h : Time instant when the SNR of subnet-B becomes greater than that of subnet-A, i.e., handoff instant.
- t_m : Time instant when the MN leaves the marginal coverage region and enters poor coverage region of a cell.
- t_{TH1} : Time instant when MN crosses the $TH1_{WLAN}$ threshold.
- t_{TH2} : Time instant when MN crosses the $TH2_{WLAN}$ threshold.
- t_1 : Time instant when address priority-change-alert/binding update reaches CN.
- t_2 : Time when mapping request reply reaches CN.
- t_3 : Time instant when binding update authentication is completed by CN.

If MN takes a linear path through the cells, the values of the above instances can be obtained easily. We denote the time duration between time instants t_x and t_y as T_{xy} .

For horizontal handoffs (between WLANs).

When $dL1_{ma} < T_{nh}$,

$$t_1 = t_h + {}^3R_4 + {}^3L_{34} + {}^3R_3 + {}^3L_{03} + {}^3R_0 + {}^2dP_1 + dP_2 \quad (3)$$

$$t_3 = t_h + {}^3R_4 + {}^3L_{34} + {}^3R_3 + {}^3L_{03} + {}^3R_0 + {}^2dP_1 + dP_2 + {}^2dP_3. \quad (4)$$

When $dL1_{ma} > T_{nh}$,

$$t_1 = t_n + dL1_{ma} + {}^3R_4 + {}^3L_{34} + {}^3R_3 + {}^3L_{03} + {}^3R_0 + {}^2dP_1 + dP_2, \quad (5)$$

$$t_3 = t_h + dL1_{ma} + {}^3R_4 + {}^3L_{34} + {}^3R_3 + {}^3L_{03} + {}^3R_0 + {}^2dP_1 + dP_2 + {}^2dP_3. \quad (6)$$

For vertical handoffs (from WLAN to GPRS),

$$t_1 = t_{TH2} + dL2_{ma} + {}^3L_{BSS} + {}^3R_{SGSN} + {}^3L_{SG} + {}^3R_{GGSN} + {}^2dP_4 + dP_5, \quad (7)$$

$$t_3 = t_{TH2} + dL2_{ma} + {}^3L_{BSS} + {}^3R_{SGSN} + {}^3L_{SG} + {}^3R_{GGSN} + {}^2dP_4 + dP_5 + {}^2dP_3. \quad (8)$$

For vertical handoffs (from GPRS to WLAN),

$$t_1 = t_e + {}^3R_4 + {}^3L_{34} + {}^3R_3 + {}^3L_{03} + {}^3R_0 + {}^2dP_1 + dP_2, \quad (9)$$

$$t_3 = t_e + {}^3R_4 + {}^3L_{34} + {}^3R_3 + {}^3L_{03} + {}^3R_0 + {}^2dP_1 + dP_2 + {}^2dP_3. \quad (10)$$

Whereas nZ_k means that Z_k is encountered by n times. If response times and link delays of routers are independent of the direction of packet's flow, then this is equal to n^*Z_k .

We divide the packets of a stream into following classes according to the time and paths that they follow.

Class-1 packets: Packets routed via the subnet-A and received by the interface-1 in effective region.

Class-2 packets: Packets routed via the subnet-A, which arrive at the AP when the MN is in marginal coverage area of a cell.

Class-3 packets: Packets routed via the subnet-A, which arrive at the AP when the MN is in poor coverage area of a cell.

Class-4 packets: Packets routed via the subnet-B and received by the interface-2.

It should be noted that classes may be empty depending upon the handoff time and coverage areas.

Now consider a constant bit rate IP stream originating from a CN destined to the MN. At every T ms, a packet is originated at the CN. We denote this packet originating time as t_{cn} . We let the originating instance t_{cn}^1 of the first packet be exponentially distributed over $[t_s, t_s + T]$. t_s is t_n for the horizontal as well as WLAN to GPRS handoffs and t_e for GPRS to WLAN handoff, then t_{cn}^k , i.e., t_{cn} for the k -th packet, will be $(k-1)T + t_{cn}^1$. For horizontal handoff, we also assume that the handoff does not affect the path which packets follow until they reach at router R_0 . We also ignore the jitter introduced in the path between CN and router R_0 .

These assumption enable us to find out the probability that a packet will be lost, expected number of lost packets as well as delay distribution and end-to-end delay of the packets taking part in the handoff. For example, to find out the expected number of lost packets due to handoff, we proceed as follows.

$$\begin{aligned} E[\text{number of lost packets}] &= \sum_{i=1, 2, 3, 4} E[\text{number of lost packets of class-}i] \\ &= \sum_{i=1, 2, 3, 4} \{E[\text{number of lost packets of class-}i] \cdot p_i\} \end{aligned} \quad (11)$$

p_i denotes the probability that a class- i packet is lost. For the above-defined IP stream, p_i is zero for class-1 and class-4 packets, one for class-3 and less than 0.05 for class-2 packets as per the coverage model. To find out the numbers of packets belonging to different classes we can proceed by finding the probability of individual packet belonging to a certain class. For example,

$$\begin{aligned} P[\text{class} - 2] &= P[\text{packet belongs to class} - 2] \\ &= P[(t_{cn}^k < t_1) \text{ AND } (t_{cn}^k < t_e - R_2 - L_{12} \\ &\quad - R_1 - L_{01} - R_0 - dP_2) \text{ AND } (t_{cn}^k < t_m \\ &\quad - R_2 - L_{12} - R_1 - L_{01} - R_0 - dP_2)] \\ &\quad \text{(for MAT)} \quad (12) \end{aligned}$$

$$\begin{aligned} &= P[(t_{cn}^k < t_3) \text{ AND } (t_{cn}^k < t_e - R_2 - L_{12} \\ &\quad - R_1 - L_{01} - R_0 - dP_2) \text{ AND } (t_{cn}^k < t_m \\ &\quad - R_2 - L_{12} - R_1 - L_{01} - R_0 - dP_2)] \\ &\quad \text{(for MIPv6)}. \quad (13) \end{aligned}$$

As for the delay distribution, we have that

$$\begin{aligned} P[\text{delay} > t] &= P[\text{packet is lost}] \\ &\quad + \sum_{\text{class}} P[\text{class AND delay} > t]. \quad (14) \end{aligned}$$

Whereas delay is a random variable that has different forms according to the class the packet belongs to, e.g., for vertical handoff from WLAN to GPRS we have

$$\begin{aligned} \text{class} - 1, 2 : \text{delay} &= dP_2 + R_0 + L_{01} + R_1 + L_{12} + R_2, \\ \text{class} - 3 : \text{delay} &= \text{infinite (as all packets are lost)}, \\ \text{class} - 4 : \text{delay} &= R_{BSS} + L_{BS} + R_{SGSN} + L_{SG} \\ &\quad + R_{GGSN} + dP_5. \end{aligned}$$

In our M/M/1 queuing model by conditioning on fixed values of t_e and dL_{ma} , the time instances and response times of routers become random variables distributed as sums of exponentially distributed random variables and constants, hence the computation of the above expressions becomes fairly straightforward.

C. Numerical Results and Discussion

In dual-interface handoff, we have exploited the overlapping coverage area between subnets. For any network, the exact value of this overlapping distance is difficult to predict. Network planners always leave some margin for this to avoid a situation of no coverage area. If the attenuation factor due to human obstruction and device orientation is assumed to be 6.4 dB and 9.0 dB, respectively, then the minimum overlapping distance for three access routers, equilaterally spaced, is more than 7 meters for coverage radius greater than 50 meters [26], when the attenuation follows a negative log function. In our model, we take the worst-case scenario with overlapping distance equals to 7 meters. Which makes overlapping time equals to 500 ms for a MN with 50 km/h speed.

Now we apply the model to investigate the handoff performance with MAT and MIPv6 being the mobility solutions in heterogeneous network. The network topology is depicted in Fig. 6. The results that will be shown are obtained with the following network characteristics. The service rate in each router

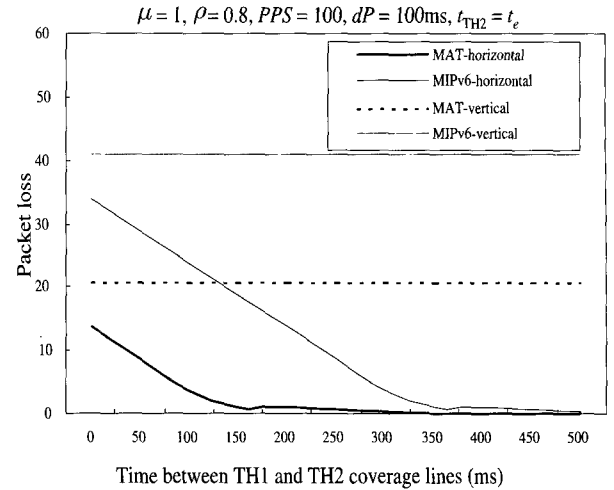


Fig. 7. The effect of TH1_{WLAN} on packet loss at various handoffs.

is set to 1000 packets per second and all routers have a load $\rho = 0.8$. The link delay among routers in the GPRS is taken as 15 ms each and among the routers in WLAN the link delay is 5 ms each [27]. As locations of CN, IMS, and MN affect the handoff significantly, so to have a better idea of this effect we place the IMS, CN, R_0 , and GGSN at the vertexes of an equilateral polygon having sides proportional to the path delays among IMS, CN, R_0 , and GGSN. This makes path delays (dP) equal. We take dP as 100 ms [28]. We also assume that IMS is at the same location as that of HA. We take the address acquisition delays $dL1_{ma}$ and $dL2_{ma}$ as 50 ms and 100 ms, respectively. Now we show some of the results depicting the effects of various handoffs on packet loss and power consumption in MAT and MIPv6 IP network environment. Packet originating rate at CN is assumed as 100 packets per second (PPS). We will analyze horizontal handoff in WLAN environment only. Moreover, vertical handoffs represent an upward handoff from WLAN to GPRS network. For the downward vertical handoff, there is obviously no packet loss as the GPRS network is available as an umbrella to the WLAN. In figures, results in MAT environment are shown by thick lines and of MIPv6 as thin lines. Horizontal handoff is represented by solid line and vertical as dashed line.

Fig. 7 (from (11)) depicts the effect of TH1_{WLAN} threshold on packet loss at various handoffs. TH2_{WLAN} coverage line is assumed to be coinciding with the effective coverage line and TH1_{WLAN} is varied in the handoff region. It is obvious that as the TH1_{WLAN} threshold increases the packet loss decreases significantly for horizontal handoff whereas the loss is constant for vertical handoff because vertical handoff does not depend upon TH1_{WLAN} . The effect of TH2_{WLAN} on packet loss is depicted in Fig. 8. For this result, we fix the TH1_{WLAN} and let TH2_{WLAN} vary. As can be seen the loss during horizontal as well as vertical handoff decreases as the threshold is increased. The significant packet loss at smaller values of thresholds is due to the long time the mapping registration requires. Packet loss becomes zero after specific thresholds. The threshold values should be sufficient enough to let the interface-2 registers its MAdr with IMS/HA and alerts CN.

In fact, the quality of handoff depends upon the rate with

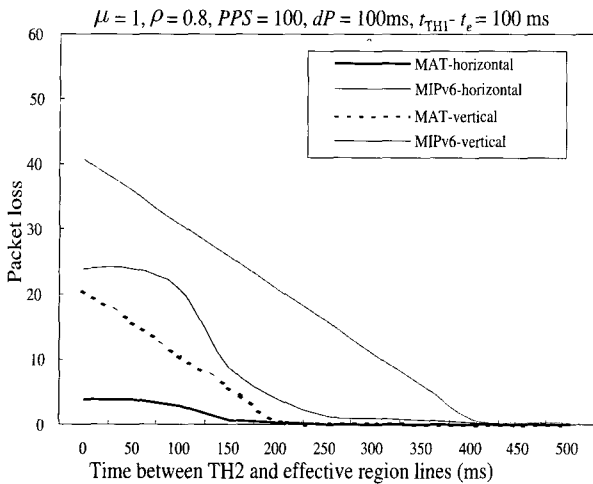


Fig. 8. The effect of $TH2_{WLAN}$ on packet loss at various handoffs.

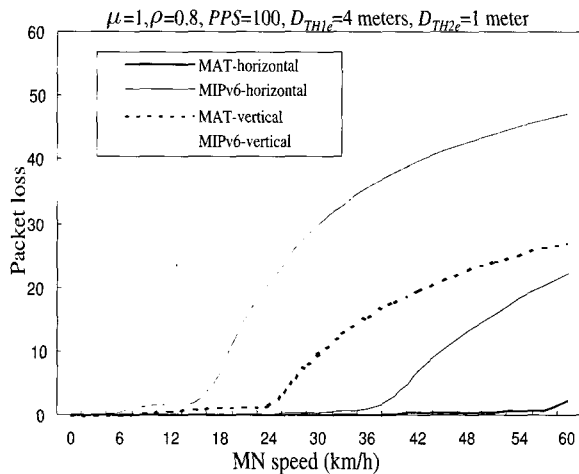


Fig. 9. Packet loss vs. MN speed (km/h) for different handoffs.

which the time instants of thresholds approach. This is depicted in Fig. 9. As can be seen, for a specific $TH1_{WLAN}$ and $TH2_{WLAN}$ the packet loss can be decreased to zero by reducing the speed of MN. It can be seen in all figures that MAT offers a considerable gain over MIPv6.

Fig. 10 (from (14)) depicts the delay distribution of individual packets during WLAN to GPRS handoff. For some of the packets the probability of delay not converging to zero is due to the lost probability of the packets. However, it is clear that handoff per se does not introduce any additional delays. Fig. 11 depicts the effect on power consumption due to using dual interfaces in mobile devices at various handoff-ring width to cell radius ratios (i.e., $(R-r)/R$, see Fig. 5). For this, we assume that the MN is moving randomly in the cell with the dwell times in handoff and normal regions exponentially distributed. Fig. 11, plotted with the help of coverage model in Fig. 6 and (1) and (2), shows the power consumed relative to a single interface for various ratios of standby state to active state power consumption. The solid line shows the relative power consumed for dual interface when the power consumed in standby state and active state is the same. As can be seen for area ratio of 10% the extra power

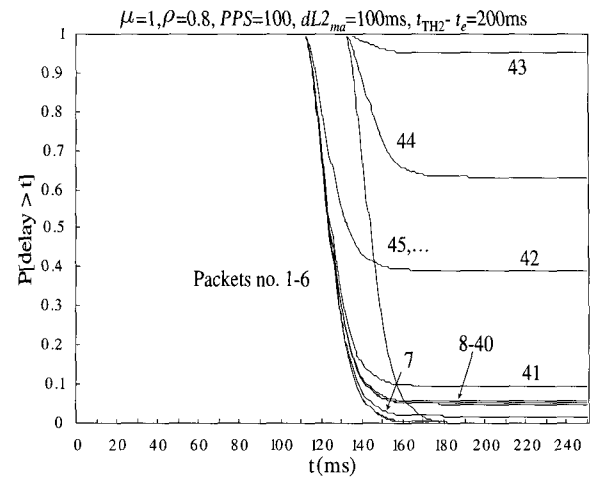


Fig. 10. Delay distribution of IP packets during vertical handoff.

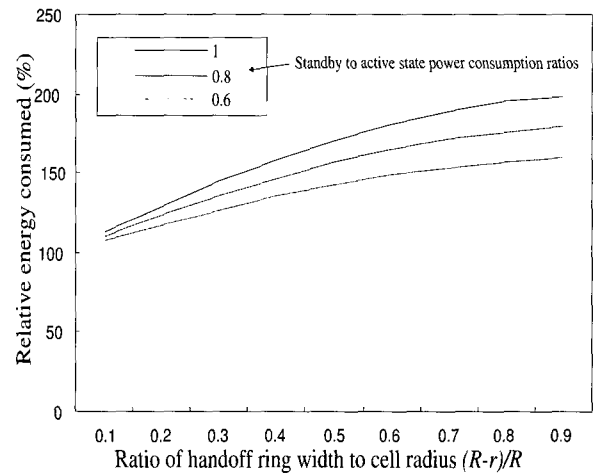


Fig. 11. Relative energy consumed for dual interface for various handoff-ring width to cell radius ratios.

consumed by the dual interface is about 13% of that of single interface. We see that the advantages we achieved by using dual interface in respect of handoff performance outweigh the extra power consumption.

D. Considerations

Proper selection of threshold values, $TH1_{WLAN}$, $TH2_{WLAN}$, and TH_{GPRS} , has significant impact on the power consumption of the MN as well as on packet loss. Especially, the threshold, $TH2_{WLAN}$ that is the switching value from WLAN to the GPRS. If we keep $TH2_{WLAN}$ very low, then MN may lose its connection with the WLAN without attaching, the interface-2 to the GPRS, thus causing a service disruption. If $TH2_{WLAN}$ is very high, then in addition to more power consumption, the communication is switched to low bandwidth network GPRS and thus lowering the throughput. In fact, these values are very network layout specific. Speed of MN also affects their selection significantly. These threshold values should have to be dynamically selected depending upon the network layout and speed of MN. We believe that in future mobile network, mechanisms by

which mobile node can know its speed have to be devised. GPS system can be used for this purpose. Measuring the time between last handoffs and relating them to the speed of MN can also be a solution. The quality of our scheme depends upon the proper selection of these thresholds.

With dual interfaces, though at present we cannot utilize all of the networks available in overlay systems but even cellular and WLAN are quite widely available and accommodate a number of applications. In fact, this is not the limitation of our handoff scheme rather the limitation of multiple access technology. We perceive in near future reconfigurable transceivers using software radio will be available which will greatly suit the proposed handoff scheme.

IP macro mobility protocols that can utilize multiple addresses of MN effectively can also improve the handoff quality. The proposed handoff design is based on end-to-end abstraction without requiring any modification in the access or core network. Therefore, the proposed scheme can complement other micro-mobility and vertical handoff solutions rather than having a conflict with them. Fully mobile-controlled procedures though require more intelligence but are more suitable for heterogeneous networks.

The paper [8] has also proposed a vertical handoff system that allows users to roam between cells in wireless overlay networks. Though the goal of our proposal and proposal [8] is same as to provide a user with the best possible connectivity for as long as possible with a minimum of service disruption during handoff, our approach is quite different than that of their approach [8]. We propose to use a fully mobile controlled handoff system that does not require any kind of modification in the access technology whereas they suggested a number of modifications in the network, which can be problematic. Furthermore, our mechanism is applicable to both vertical and horizontal handoffs.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a mobile-controlled mechanism for reducing the handoff latencies during horizontal and vertical handoffs in heterogeneous overlay networks. IP has been assumed as the core network protocol. We have devised the interface switching criterion and handoff algorithms keeping in mind the power consumption and quality parameters of the handoff. The proposed handoff procedure is then applied to the WLAN/GPRS network by using dual-mode multi-interfaced devices. The handoff requires that MN can control the switching of communication between two dual mode interfaces. By dual mode, we mean that interface is capable of connecting to GPRS and WLAN networks. The handoff utilizes the overlapping area between similar subnets for horizontal handoff and between different networks for vertical handoff. Handoff mechanism is analyzed with MIPv6 and MAT as macro-mobility protocols. The simultaneous connection capability with dual interface provides a sort of IP diversity that helps to reduce service disruption time. In the course of developing and analyzing dual interface handoff following conclusions can be drawn.

- The evaluation of dual interface handoff shows that it is possible to reduce packet loss and handoff latency by exploiting the overlay infrastructure and speed of the mobile hosts.

With dual interfaces starting the IP address acquisition and authentication procedures before the actual handoff can reduce the handoff latency significantly.

- Results also show that implementation of the mobility or location database agents affects the quality of handoff. Handoff latency in MAT is reduced at least by 33% as that of MIPv6 as the latter, for route optimization, requires return routability method to authenticate the binding updates [5]. The capability of handling multiple IP address by the IMS, location database server in MAT, also reduces the handoff latency.
- The proposed handoff scheme does not require any special service from the access networks and thus is scalable and works fine in different hybrid networks. In heterogeneous environment due to the high pace with which the new media access standards are proliferating, the client-based handoff mechanisms like dual interface handoff is very suitable.
- Results show that dual interface consumes 13% more power than that of single interface. In many cases, this bad impact is relatively negligible as compared to the improvement in the handoff quality. In the future, we are planning to evaluate the dual interface performance in real network. We will also devise a mechanism to dynamically select the communication switching threshold values depending upon the speed of mobile node and network layout.

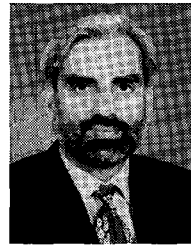
ACKNOWLEDGMENTS

The authors would like to thank Prof. Kaori Maeda of Hiroshima City University and Mr. Takahiro Fujita for their invaluable input to this research. The authors also express their deepest appreciation to the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

- [1] R. Berezdivin, R. Breinig, and R. Topp, "Next generation wireless communications concepts and technologies," *IEEE Commun. Mag.* vol. 40, no. 3, pp. 108–116, Mar. 2002.
- [2] E. Brewer *et al.*, "A network architecture for heterogeneous mobile computing," *IEEE Pers. Commun.*, Oct. 1998.
- [3] K. Pahlavan *et al.*, "Handoffs in hybrid mobile data networks," *IEEE Pers. Commun.*, Apr. 2000.
- [4] C. Perkins, "IP mobility support for IPv4," *RFC 3344*, IETF, 2002.
- [5] D. B. Johnson, C. Perkins, and J. Arkko, "Mobility support in IPv6," *RFC 3775*, IETF, 2004.
- [6] A. Sanmatau *et al.*, "Using mobile IP for provision of seamless handoff between heterogeneous access networks, or how a network can support the Always-On concept," EURESCOM Summit, 2001.
- [7] X. Zhao *et al.*, "Flexible network support for mobile hosts," in *Proc. ACM/IEEE, MobiCom'98*, USA.
- [8] M. Stemm and R. Katz, "Vertical handoffs in wireless overlay networks," *ACM MONET*, vol. 3, no. 4, pp. 335–350, Apr. 1998.
- [9] D. Wong and T. J. Lim, "Soft handoff in CDMA mobile systems," *IEEE Commun. Mag.* vol. 34, no. 12, pp. 6–17, Dec. 1997.
- [10] P. Hyosoon *et al.*, "Vertical handoff procedure and algorithm between IEEE802.11 WLAN and CDMA cellular network," *Lecture Notes in Computer Science (LNCS)*, no. 2524, 2003.
- [11] M. Ye *et al.*, "The mobile IP handoff between hybrid networks," in *Proc. PIMRC 2002*, Portugal, Sept. 2002.
- [12] M. Ylianttila *et al.*, "Optimization scheme for mobile users performing vertical handoffs between IEEE 802.11 and GPRS/EDGE networks," in *Proc. GlobeCom 2001*, Texas, Sept. 2001.
- [13] M. Stemm and R. Katz, "Reducing power consumption of network interfaces in handheld devices," in *Proc. Mobile Multimedia Commun. (MoMuc-3)*, Dec. 1996.

- [14] R. Inayat *et al.*, "Realizing high mobility through an end-to-end network architecture with IP diversity support for real time internet applications," in *Proc. SAINT 2004*, Tokyo, Jan. 2004.
- [15] J. Mitola, "The software radio architecture," *IEEE Pers. Commun.*, May 1995.
- [16] P. M. L. Chan *et al.*, "Mobility management incorporating fuzzy logic for a heterogeneous IP environment," *IEEE Commun. Mag.*, vol. 39, no. 12, pp. 42–51, Dec. 2001.
- [17] H. Soliman *et al.*, "Hierarchical MIPv6 mobility management (HMIPv6)," Internet-draft, IETF 2004, (work in progress).
- [18] R. Inayat *et al.*, "An end-to-end network architecture for supporting mobility in wide area wireless networks," *IEICE Trans. Commun.*, vol. E87-B, no. 6, June 2004.
- [19] R. Inayat *et al.*, "MAT: An end-to-end mobile communication architecture with seamless IP handoff support for the next generation Internet," in *Proc. second Human.Society@internet Conference*, Seoul, June 2003.
- [20] B. Sarikaya, "Packet mode in wireless networks: Overview of transition to third generation," *IEEE Pers. Commun.*, Nov. 2000.
- [21] 3GPP TS 23.060: General Packet Radio Service (GPRS); Service Description; Stage 2, 2002.
- [22] 3GPP TS 29.060: General Packet Radio Service (GPRS); GPRS Tunneling Protocol across the Gn and Gp Interface, 2002.
- [23] 3GPP TS 29.061: Packet Domain; Internetworking between the Public Land Mobile Network (PLMN) Supporting Packet Based Services and Packet Data Network, 2003.
- [24] R. Hsieh, Z. G. Zhou, and A. Seneviratne, "S-MIP: A seamless handoff architecture for mobile IP," in *Proc. INFOCOM 2003*, San Francisco, USA, 2003.
- [25] S. L. Su *et al.*, "Performance analysis of soft handoff in CDMA cellular networks," *IEEE J. Select. Areas Commun.*, vol. 14, no. 9, pp. 1762–1769, Dec. 1996.
- [26] Z. G. Zhou *et al.*, "A software nased indoor relative location management system," in *Proc. Wireless and Optical Commun. Canada*, 2002.
- [27] A. Corlett *et al.*, "Statistics of one-way internet packet delay," in *Proc. IETF*, Minneapolis, Mar. 2002.
- [28] C. Bovy *et al.*, "Analysis of end-to-end delay measurement in the internet," in *Proc. RIPE41*, Amsterdam, Jan. 2002.



Riaz Inayat received his B.Sc. in Electrical Engineering from University of Engineering and Technology (UET) Lahore, Pakistan and M.Eng. in Telecommunication from Asian Institute of Technology (AIT), Thailand in 1991 and 1997, respectively. In 2004, he completed his Ph.D. from Graduate School of Engineering, Hiroshima University. Since 1993, he has been working in Pakistan Telecommunication Company Limited under Ministry of IT and Telecommunications of Pakistan. His research interests include data communication, mobile networks, and multimedia communication. He is a member of IEEE Communication Society.



Reiji Aibara received the B.E., M.E., and D.E. degrees from Hiroshima University Japan in 1981, 1983, and 1986, respectively. He served at Hiroshima University as a research associate from 1986 to 1989 and as an associate professor of the research center for Integrated Systems from 1989 to 1994. He is presently a professor at Hiroshima University. His research interests include computer architecture, real-time computing, local/wide area networks, and high-speed computer networks. He is a member of IEEE Computer Society, IEEE ComSoc, IEICE, and IPSJ.



Kouji Nishimura received the B.S., M.S., and Ph.D. degrees in information engineering from Hiroshima University, Japan, in 1989, 1991, and 2002, respectively. He worked at ANA System Planning, Co. from 1991 to 1994. He is currently a research associate at Information Media Center of Hiroshima University. His research interests include real-time remote control protocols for multimedia devices and management of computer networks. He is a member of IEICE, IPSJ and ISOC.