Abstract

With the emergence of a variety of mobile data services with variable coverage, bandwidth, and handoff strategies, and the need for mobile terminals to roam among these networks, handoff in hybrid data networks has attracted tremendous attention. This article presents an overview of issues related to handoff with particular emphasis on hybrid mobile data networks. Issues are logically divided into architectural and handoff decision time algorithms. The handoff architectures in high-speed local coverage IEEE 802.11 wireless LANs, and low-speed wide area coverage CDPD and GPRS mobile data networks are described and compared. A survey of traditional algorithms and an example of an advanced algorithm using neural networks for HO decision time in homogeneous networks are presented. The HO architectural issues related to hybrid networks are discussed through an example of a hybrid network that employs GPRS and IEEE 802.11. Five architectures for the example hybrid network, based on emulation of GPRS entities within the WLAN, mobile IP, a virtual access point, and a mobility gateway (proxy), are described and compared. The mobility gateway and mobile IP approaches are selected for more detailed discussion. The differences in applying a complex algorithm for HO decision time in a homogeneous and a hybrid network are shown through an example.

Handoff in Hybrid Mobile Data Networks

KAVEH PAHLAVAN, PRASHANT KRISHNAMURTHY, AHMAD HATAMI,
WORCESTER POLYTECHNIC INSTITUTE
MIKA YLIANTTILA, JUHA-PEKKA MAKELA, ROMAN PICHNA, JARI VALLSTRÖM,
UNIVERSITY OF OULU

Handoff (HO) [1] is extremely important in any mobile network because of the default cellular architecture employed to maximize spectrum utilization. When a mobile terminal moves away from a base station, the signal level degrades and there is a need to switch communications to another base station. Handoff is the mechanism by which an ongoing connection between a mobile terminal or host (MH) and a correspondent terminal or host (CH) is transferred from one point of access to the fixed network to another. In cellular voice telephony and mobile data networks, such points of attachment are referred to as *base stations* (BSs) and in wireless LANs (WLANs), they are called *access points* (APs). In either case, such a point of attachment serves a coverage area called a *cell*. Handoff, in the case of cellular telephony, involves the transfer of a voice call from one BS to another. In the case of WLANs, it involves transferring the connection from one AP to another. In hybrid networks it will involve the transfer of a connection from one BS to another, from an AP to another, between a BS and an AP, or vice versa.

For a voice user, HO results in an audible click interrupting the conversation for each HO [1]; and because of HO, data users may lose packets and unnecessary congestion control measures may come into play [2]. Degradation of the signal level, however, is a random process, and simple decision mechanisms such as those based on signal strength measurements result in the *ping-pong effect*. The ping-pong effect refers to several HOs that occur back and forth between two BSs. This takes a severe toll on both the user's quality perception and the network load. One way of eliminating the ping-pong effect is to persist with a BS for as long as possible. However, if HO is delayed, weak signal reception persists unnecessarily, resulting in lower voice quality, increasing the probability of call drops and/or degradation of quality of service (QoS). Consequently, more complex algorithms are needed to decide on the optimal time for HO. Handoff also involves a sequence of events in the backbone network, including rerouting the connection and reregistering with the new AP, which are additional loads on network traffic. Handoff has an impact on traffic matching and traffic density for individual BSs (since the load on the air interface is transferred from one BS to another). In the case of random access techniques employed to access the air interface, or in code-division multiple access (CDMA), moving from one cell to another impacts QoS in both cells since throughput and interference depend on the number of terminals competing for the available bandwidth. In hybrid data networks, a decision on HO has an impact on the throughput of the system.

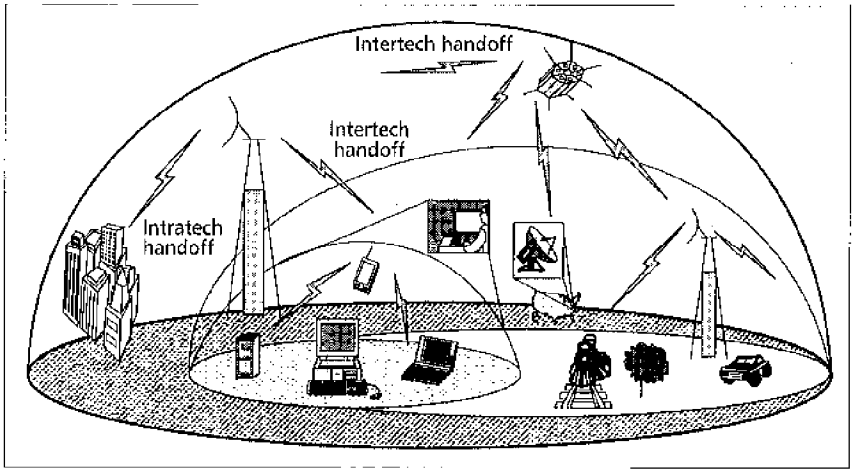
While significant work has been done on HO mechanisms in circuit-switched mobile networks [1, 3], there is not much literature available on packet-switched mobile networks. In this study we are mainly interested in non-real-time applications in wireless networks. Performance measures such as call blocking and call dropping are applicable only to real-time traffic and may not be suitable for the bursty traffic that exists in client-server applications. When a voice call is in progress, allowed latency is very limited, resource allocation has to be guaranteed, and, while occasionally some packets may be dropped and moderate error rates are permissible, retransmissions are not possible, and connectivity has to be maintained continuously. On the other hand, bursty data traffic by definition needs only intermittent connectivity, and can tolerate greater latencies and employ retransmission of lost packets. In such networks HO is warranted only when the terminal moves out of coverage of the current point of attachment or the traffic load is so high that an HO may result in greater throughput and utilization.

Wireless data services are becoming increasingly popular, but are not ubiquitous. Consequently, the natural trend has

P. Krishnamurthy is currently with the University of Pittsburgh.
A. Hatami is currently with Lucent Technologies.
R. Pichna and J. Vallström are currently with Nokia.

1070-9916/00/$10.00 © 2000 IEEE

been toward utilizing small-coverage high-bandwidth data networks such as IEEE 802.11 whenever they are available and switching to an overlay service such as the General Packet Radio Service (GPRS) network with low bandwidth when the coverage of a wireless local area network (WLAN) is not available (Fig. 1). We refer to such a procedure as *intertech roaming* or *HO in hybrid networks*. When we consider HOs between hybrid packet-switched networks, an HO from a WLAN AP to a GPRS BS, for example, should be done only with very low priority, while an HO from a GPRS BS to a WLAN AP should be done whenever it is possible in light of the orders of magnitude of difference in available bandwidth between the two systems.



■ **Figure 1.** *Hybrid or nonhomogeneous mobile data networks.*

We present an overview of the issues related to HO. Issues are logically classified into architectural issues and HO decision time algorithms. A brief survey of traditional HO algorithms based on received signal strength is provided. The HO architectures in low-speed wide-area-coverage Cellular Digital Packet Data (CDPD) and GPRS mobile data networks and high-speed local-coverage IEEE 802.11 WLANs are described and compared. An example of an advanced algorithm using neural networks for HO decision time in homogeneous networks is presented. The HO architectural issues related to hybrid networks are discussed through an example of a hybrid network that employs GPRS and IEEE 802.11. Five architectures for the example hybrid network, based on emulation of GPRS entities within the WLAN, mobile IP, a virtual access point, and a mobility gateway (proxy), are described and compared. The mobility gateway and mobile IP approaches are selected for more detailed discussion. The differences in applying a complex algorithm for HO decision time in a homogeneous and a hybrid network are shown through an example.

Issues in Handoff

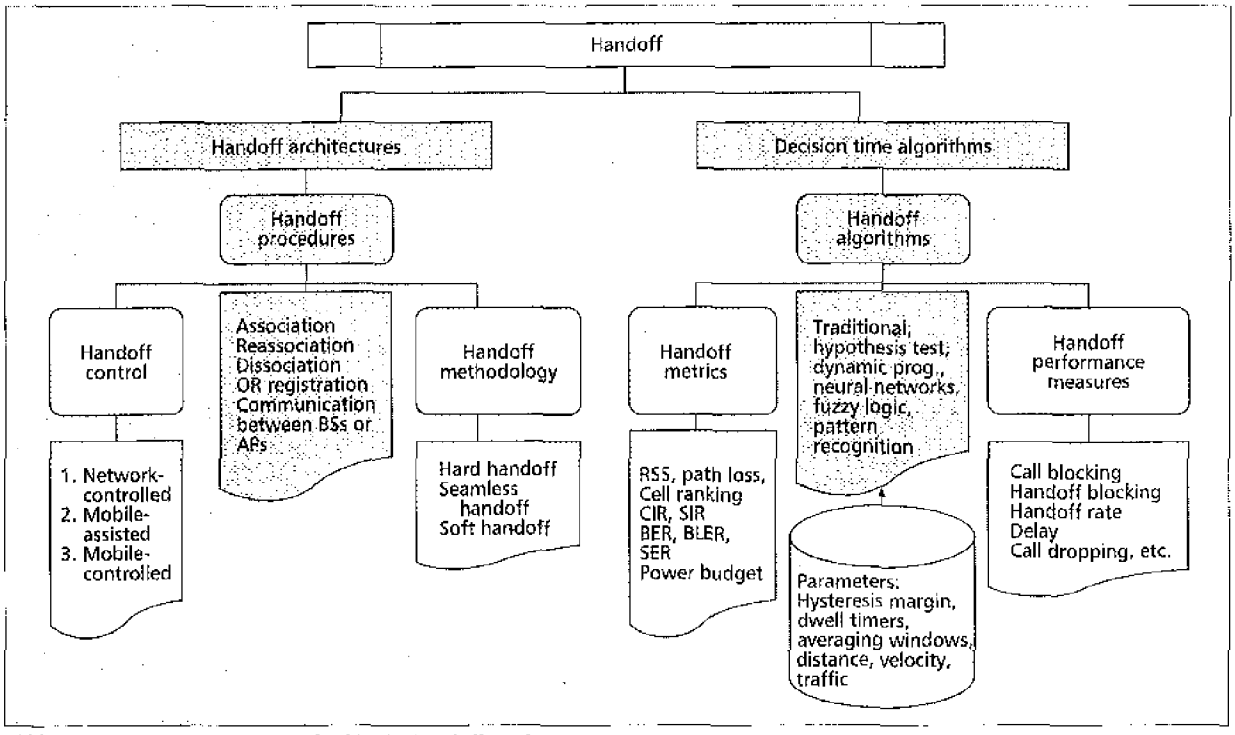
There are a variety of issues related to HO. As shown in Fig. 2, these issues are divided into two categories: architectural issues and HO decision time algorithms. Architectural issues are those related to the methodology, control, and software/hardware elements involved in rerouting the connection. Issues related to the decision time algorithms are the types of algorithms, metrics used by the algorithms, and performance evaluation methodologies.

Architectural Issues

Handoff procedures involve a set of protocols to notify all the related entities of a particular connection that an HO has been executed and that the connection has to be redefined. In data networks, the MH is usually registered with a particular point of attachment. In voice networks, an idle MH would have selected a particular BS that is serving the cell in which it is located. This is for the purpose of routing incoming data packets or voice calls appropriately. When the MH moves and executes an HO from one point of attachment to another, the old serving point of attachment has to be informed about the change. This is usually called *dissociation*. The MH will also have to reassociate itself with the new point of access to the fixed network. Other network entities involved in routing data packets to the MH or switching voice calls have to be aware of the HO in order to seamlessly continue the ongoing connection or call. Depending on whether a new connection is created before breaking the old one or not, HOs are classified into hard and seamless HOs. In CDMA, the existence of two simultaneous connections during HO results in soft HO [4]. The decision mechanism or *handoff control* may be located in a network entity (as in cellular voice) or in the MH itself (as in mobile data and WLANs). These cases are called *network-controlled handoff* (NCHO) and *mobile-controlled handoff* (MCHO), respectively. In GPRS, information sent by the MH can be employed by the network entity in making the handoff decision. This is called *mobile-assisted handoff* (MAHO). In any case, the entity that decides on the HO uses some metrics, algorithms, and performance measures in making the decision. These are discussed below.

Decision Time Algorithms

Several algorithms are being employed or investigated to make the correct decision to hand off [1, 3]. Traditional algorithms employ thresholds to compare the values of *metrics* from different points of attachment and then decide on when to make the HO. A variety of metrics have been employed in mobile voice and data networks to decide on an HO. Primarily, the received signal strength (RSS) measurements from the serving point of attachment and neighboring points of attachment are used in most of these networks. Alternatively or in conjunction, the path loss, carrier-to-interference ratio (CIR), signal-to-interference ratio (SIR), bit error rate (BER), block error rate (BLER), symbol error rate (SER), power budgets, and cell ranking have been employed as metrics in certain mobile voice and data networks. In order to avoid the ping-pong effect, additional parameters are employed by the algorithms such as hysteresis margin, dwell timers, and averaging windows. Additional parameters (when available) may be employed to make more intelligent decisions. Some of these parameters also include the distance between the MH and the point of attachment, the velocity of the MH, and traffic characteristics in the serving cell. The performance of HO algorithms is determined by their effect on certain performance measures. Most of the performance measures that have been considered, such as call blocking probability, HO blocking probability, delay between HO request and execution, and call dropping probability, are related to voice connections. Handoff rate (number of HOs per unit of time) is related to the ping-pong effect, and algorithms are usually designed to minimize the number of unnecessary HOs. While minimizing the HO rate is important in mobile data networks, other issues include



■ **Figure 2.** *Important issues involved in the handoff mechanism.*

throughput maximization and maintaining QoS guarantees during and after HO. However, these issues have not received sufficient attention in the literature.

Traditional HO algorithms are all based on the received signal strength (RSS) or received power P. Some of the traditional algorithms [1] are as follows:

- *RSS*: The BS whose signal is being received with the largest strength is selected (choose BS B_{new} if $P_{new} > P_{old}$).
- *RSS plus Threshold*: An HO is made if the RSS of a new BS exceeds that of the current one and the signal strength of the current BS is below a threshold T (choose B_{new} if $P_{new} > P_{old}$ and $P_{old} < T$).
- *RSS plus Hysteresis*: An HO is made if the RSS of a new BS is greater than that of the old BS by a hysteresis margin H (choose B_{new} if $P_{new} > P_{old} + H$).
- *RSS, Hysteresis, and Threshold*: An HO is made if the RSS of a new BS exceeds that of the current BSby a hysteresis margin H and the signal strength of the current BS is below a threshold T (choose B_{new} if $P_{new} > P_{old} + H$ and $P_{old} < T$).
- *Algorithm plus Dwell Timer*: Sometimes a dwell timer is used with the above algorithms. A timer is started the instant the condition in the algorithm is true. If the condition continues to be true until the timer expires, an HO is performed.

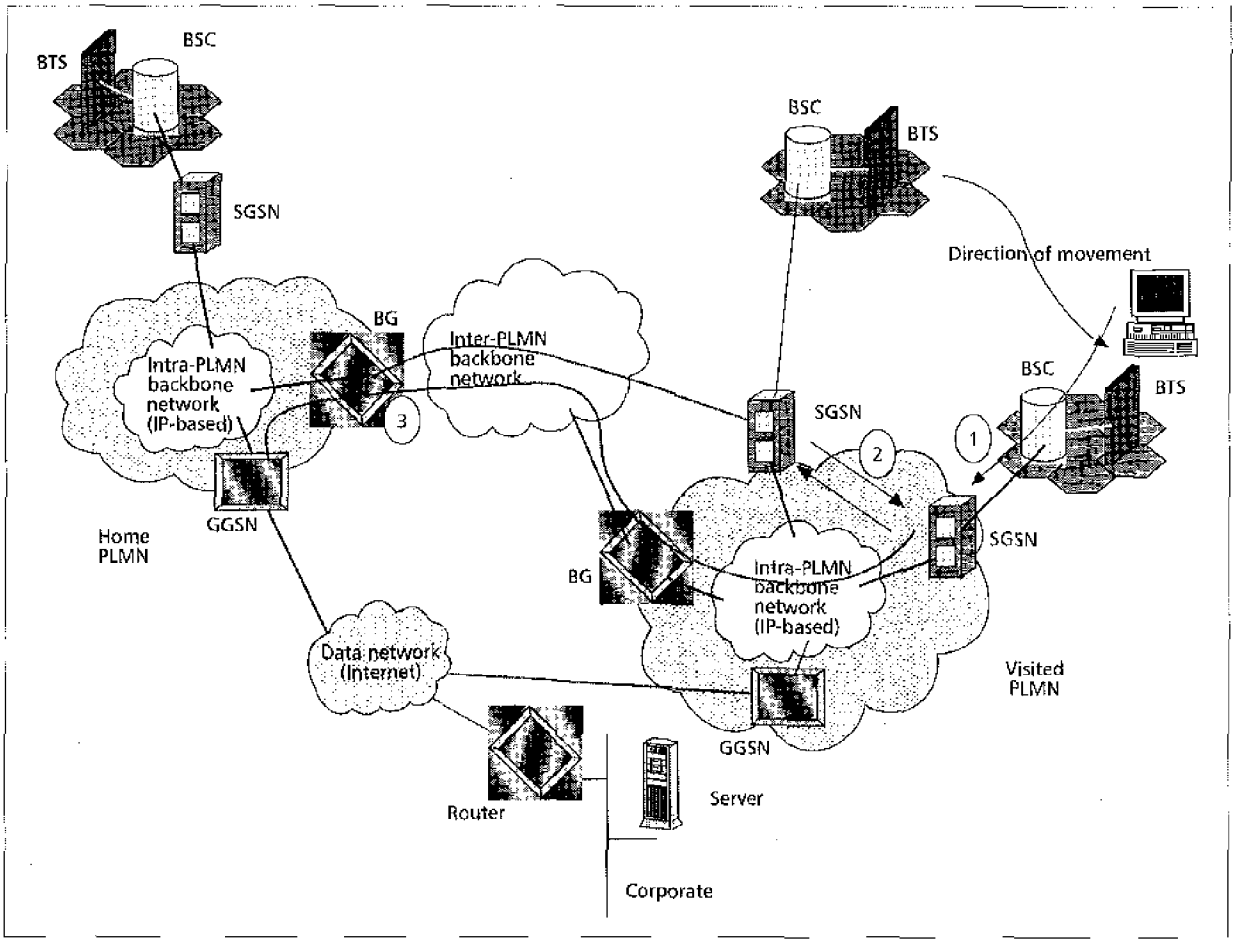
Recently, other techniques are emerging such as hypothesis testing [5], dynamic programming [6], and pattern recognition techniques based on neural networks or fuzzy logic systems [7] (for an excellent survey of various algorithms, see [1, 7]). These complicated algorithms are necessitated by the complexity of the HO problem, especially in hybrid data or voice networks. The mobile terminal has to monitor the air for wireless data services that may be available for attachment. As an example, consider an MH that could connect to either an 802.11 WLAN AP connected to a LAN or a GPRS BSS connected to a backbone GPRS network. There must be a mechanism or algorithm within the MH that will enable it to choose the best available service and switch to this service as soon as it is available. For example, the MH must be able to switch from the GPRS service to the WLAN AP as soon as it detects the availability of a connection to an AP. Most of the emergent algorithms are in their nascent stages, and have been analyzed or simulated only for voice networks and only in extremely simple scenarios.

Handoff Architectures and Algorithms in Homogeneous Mobile Data Networks

Handoff in various technologies is different. To illustrate the similarities and differences, we consider GPRS, CDPD and IEEE 802.11 as examples below. Architectures are open and standardized but very often the HO decision time algorithms are proprietary. We consider an example of a neural network (NN) algorithm in the last subsection which indicates the difference in performance between traditional and advanced algorithms with the HO rate and delay as performance measures.

Handoff in General Packet Radio Service

GPRS [8–10] is an enhancement of the Global System for Mobile Communications (GSM). It uses exactly the same physical radio channels as GSM, and only new logical GPRS radio channels are defined. Allocation of these channels is flexible: from one to eight radio interface time slots can be allocated per time-division multiple access (TDMA) frame. Time slots are shared by active users, and the uplink and downlink are allocated separately. Physical channels are taken from the common pool of available channels in the cell. Allocation to circuit-switched services and GPRS is done dynamically according to a *capacity on demand* principle. This means that the capacity allocation for GPRS is based on the actual need for packet transfers. GPRS does not require permanently allocated physical channels. Logical network nodes called *GPRS support nodes* (GSNs) are used for packet routing in



■ **Figure 3.** *The GPRS routing update procedure.*

the backbone. The *gateway GSN* (GGSN) acts as the interface to public data networks such as the Internet and contains the routing information to be used to tunnel packets to the MH through a *serving GSN* (SGSN). The SGSN is responsible for location management and delivery of packets.

The GGSN and SGSN can be considered the mobile IP equivalents of the home agent (HA) and foreign agent (FA). Packets originating from an MH are routed by an SGSN to its destination as in any packet-switched network. Packets intended for an MH reach the GGSN associated with its *home* network. The GGSN determines which SGSN is serving the MH, encapsulates the packet, and forwards (tunnels) it to the SGSN. Information related to the MH is stored in a *GPRS register* (GR), which is part of the *home location register* (HLR) of GSM.

A GPRS MH can be in one of three states: idle (unreachable), ready (where it is registered with an SGSN), and standby (inactive for a long time). In order to communicate, an MH does a GPRS attach and enters a ready state. The MH is responsible for cell reselection independently, and this is done in the same way as in GSM. The MH measures the RSS of the current *broadcast control channel* (BCCH), compares it to the RSS of the BCCH of adjacent cells, and decides to which cell to attach. There is, however, an option available to operators to make the BSS ask for reports from the MH (as in GSM), and then the HO is done as in GSM (MAHO). Plain GPRS-specific information can be sent in a *packet BCCH* (PBCCH), but the RSS is always measured from the BCCH. There are also other principles which may be considered in HO decision (path loss, cell ranking, etc.).

The location is updated with a routing update procedure, as shown in Fig. 3. A routing area (RA) corresponds to a group of cells. When an MH changes its RA, it sends an RA update request containing the cell identity and the identity of the previous RA to the new SGSN (1). Note that an intra-SGSN routing area update is also possible when the same SGSN serves the new RA. The new SGSN asks the old SGSN to provide the routing context (GGSN address and tunneling information) of the MH (2). The new SGSN then updates the GGSN of the home network with the new SGSN address and new tunneling information (3). The new SGSN also updates the HLR. The HLR cancels the MH information context in the old SGSN and loads the subscriber data to the new SGSN. The new SGSN acknowledges the MH. The previous SGSN is requested to transmit undelivered data to the new SGSN.

Handoff in Cellular Digital Packet Data

The CDPD network [11–13] operates as a connectionless network that is a wireless extension to existing wired connectionless networks. It shares the existing infrastructure as well as spectrum of the Advanced Mobile Phone Service (AMPS) analog cellular telephone network in the United States. Consequently, it employs different physical channels or frequency bands for uplink and downlink transmissions. Handoff occurs when an MH moves from one cell to another, the CDPD channel quality deteriorates, the current CDPD channel is

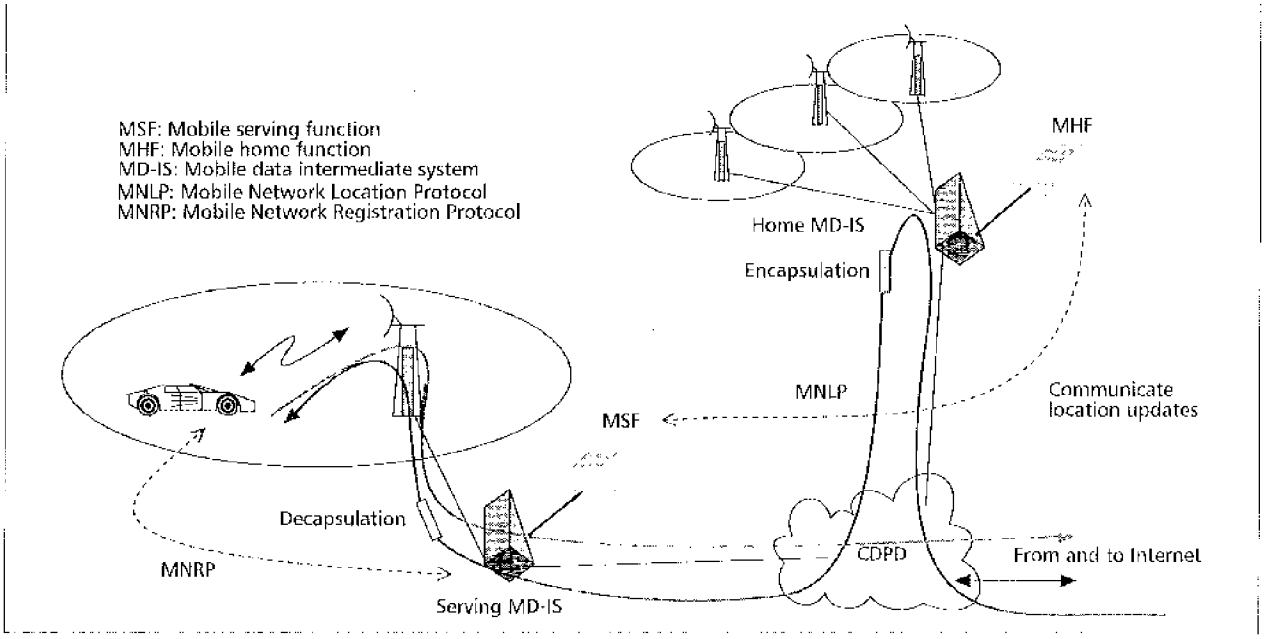


Figure 4. *Mobility management in CDPD.*

requested by an AMPS voice call (forced hop), or the load on CDPD channels in the current cell is much more than that on the channels in an overlapping cell.

Before going into the details of the HO procedure in CDPD, we briefly consider some of the salient features of CDPD. Mobile hosts are usually full-duplex (they can transmit and listen at the same time), although low-cost devices may be half-duplex. The physical layer of CDPD provides the ability to tune to a specific RF channel, the ability to measure the RSS indication (RSSI) of the received signal, the ability to set the power of the MIT transmitted signal to a specified level, and the ability to suspend and resume monitoring of RF channels in the MIT. Both uplink and downlink channels are slotted. There is no contention on the downlink, and the BS will transmit link layer frames sequentially. On the uplink, a digital sense multiple-access with collision detection (DSMA/CD)

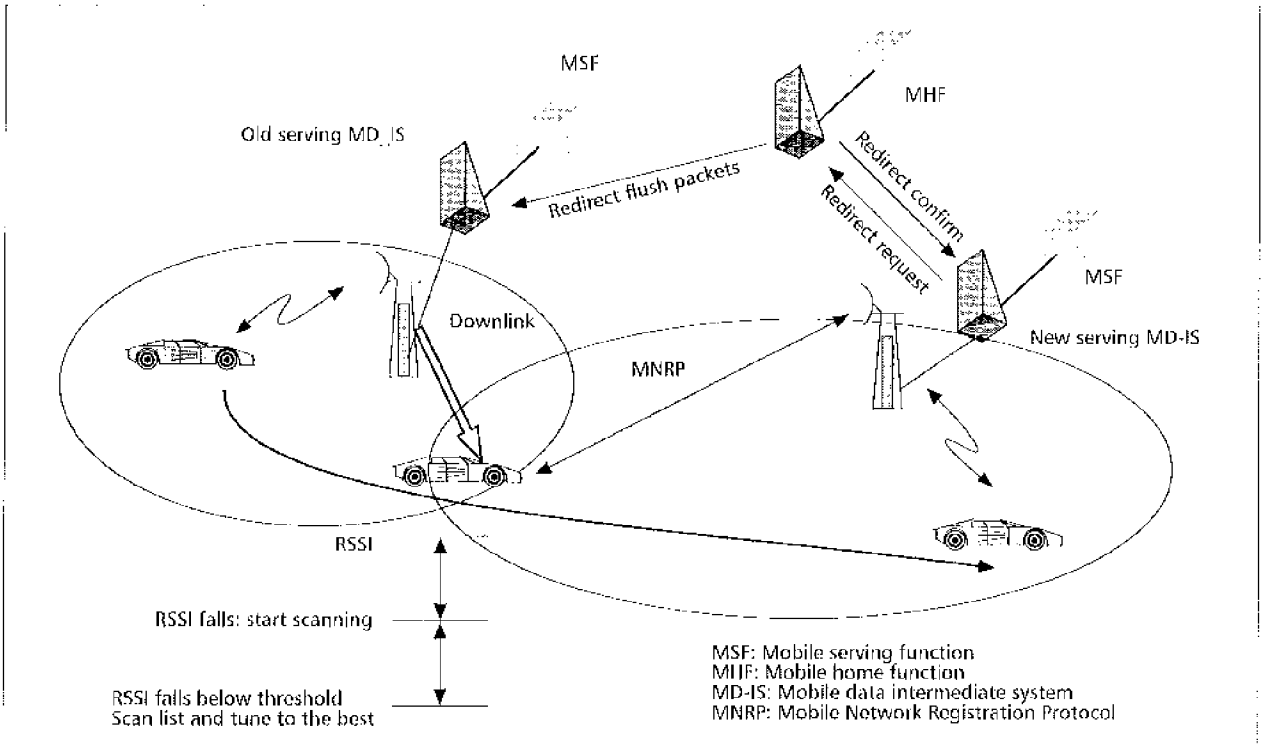


Figure 5. *The handoff procedure in CDPD.*

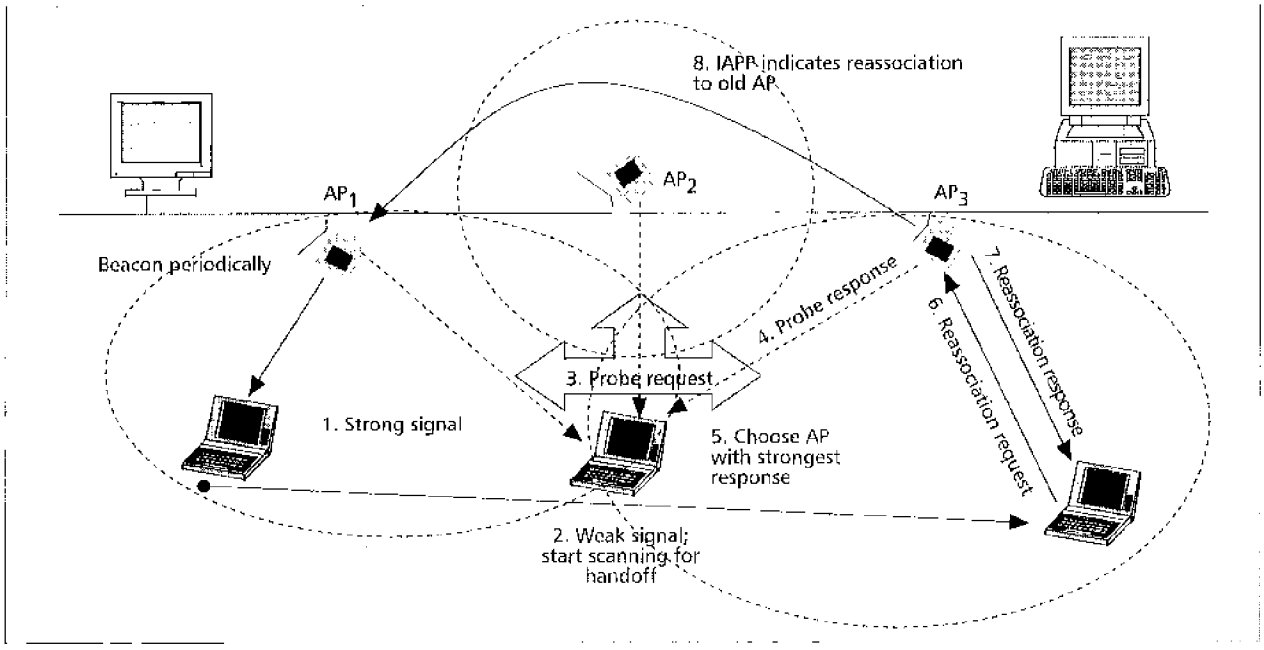


Figure 6. *Handoff procedures in an IEEE 802.11 WLAN.*

protocol is employed. Collision detection is at the BS and informed to the MHs on the downlink. On the downlink, multiple *cell configuration messages* are broadcast which include, for the given cell and its neighbors, the cell identifier, a reference channel for the cell, a value that provides the difference in power between the reference channel and the actual CDPD data channel, an RSS bias to compare the RSS of the reference channels of the given cell and adjacent cells, and a list of channels allocated to CDPD within the given cell. RSS measurements are always done on the reference channel since the CDPD channel list may keep changing [12].

Upon powering on, the MH scans the air and locks onto the strongest "acceptable" CDPD channel stream it can find and registers with the mobile data intermediate system (MD-IS) that serves the base station. This is done via the Mobile Network Registration Protocol (MNRP) whereby the MH announces its presence and also authenticates itself [12]. Registration protects against fraud and enables the CDPD network to know the mobile location and update its mobility databases. The MH continues to listen to the CDPD channel unless it (or the CDPD network) initiates an HO.

CDPD mobility management [11] is based on principles similar to mobile IP. The details are shown in Fig. 4. The MD-IS is the central element in the process. An MD-IS is logically separated into a *home MD-IS* and a *serving MD-IS*. A home MD-IS contains a subscription database for its geographical area. Each subscriber is registered in his/her home MD-IS associated with his home area. The IP address of a subscriber points to his/her home MD-IS. At the home MD-IS, a *mobile home function* (MHF) maintains information about the current location of MHs associated with (homed at) that home MD-IS. The MHF also encapsulates any packet addressed to an MH homed with it, directing it to a *mobile serving function* (MSF) associated with the serving MD-IS whose serving area the MH is currently visiting. A serving MD-IS manages one serving area. Mobile data BSs that provide coverage in this area are connected to the serving MD-IS, whose MSF contains information about all subscribers currently visiting the area and registered with it. The MSF employs the Mobile Network Location Protocol (MNLP) to notify the MHF about the presence of the MH in its service area. The channel stream in which a subscriber is active is also indicated. The MSF decapsulates forwarded packets and routes them to the correct channel stream in the cell.

The HO procedure in CDPD is shown in Fig. 5. The HO initiation and decision in CDPD are as follows. The HO is mobile-controlled. The MH always measures the signal strength of the reference channel [12]. An MH scans for alternative channels when its signal deteriorates. Since certain cells may have large shadowing effects within them, the operator can set an RSSI scan value to determine when an MH should start scanning for alternative channels. An MH will ignore a drop in signal level if the RSSI scan value is large enough or start scanning for alternative channels if it is small. This value is also useful (and should be made small) when the signal strength does not drop even when the MH has moved well into a neighboring cell. When additional thresholds for RSSI hysteresis, BLER, and/or symbol error rate (SER) are reached, the MH will go through a list of channels of adjacent cells that the current BS is broadcasting and tune in to the one with the best signal strength. The MH informs the new BS that it has entered its cell. The MSF of the new MD-IS uses a redirect-request and redirect-confirm procedure with the MHF of the MH. The MHF also informs the old serving MD-IS about the HO and directs it through its MSF to redirect packets it may have received for the MH to the new serving MD-IS or flush them. Depending on the nature of HO (interoperator or intra-operator), the delay of registration and traffic redirection will vary.

Handoff in IEEE 802.11 Wireless Local Area Networks

The IEEE 802.11 WLAN standard [14–16] defines the coverage area of a single AP as a *basic service set* (BSS); to extend this, multiple BSSs are to be connected through a *distribution system* (usually the wired network) to form an *extended service set* (ESS). The 802.11 standard defines only the over-the-air interactions (communication between MHs and the AP). The internals of how the ESS should be formed are left to the AP management entity and are not defined by the 802.11 standard. Recently, a draft *inter-access-point protocol* (IAPP) has been specified to standardize the communication between APs over the wired interface [17].

[illegible]	[illegible]	[illegible]
Beacon is on the same physical channel as data	Beacon (BCCH) is on a separate physical channel from data traffic	Beacon (reference) is usually on a separate physical channel from data traffic
Decision on handoff is made at the mobile terminal	Decision on handoff is made at the base station controller or mobile host	Decision on handoff is made at the mobile terminal
IAPP protocol to inform old AP about handoff	SGSN updates GGSN in GPRS	Flush/redirect packet message from MHF to MSF of the old BS
The multiple access is CSMA, and the channel is monitored all the time before packet transmission	The multiple access is TDMA-based, and channel monitoring is done during times when the MH does not transmit or receive	The multiple access is DSMA/CD, and the channel is monitored all the time before packet transmission

Table 1. *Comparison of handoff procedures in voice and data mobile networks.*

The HO procedures in a WLAN are as shown in Fig. 6. The AP broadcasts a beacon signal periodically (typically the period is around 100 ms). An MH that powers on scans the beacon signal and *associates* itself with the AP with the strongest beacon. The beacon contains information corresponding to the AP such as a time stamp, beacon interval, capabilities, ESS ID, and traffic indication map (TIM). The MH uses the information in the beacon to distinguish between different APs.

The MH keeps track of the RSS of the beacon of the AP with which it is associated; when the RSS becomes weak, it starts to scan for stronger beacons from neighboring APs. The scanning process can be either active or passive. In passive scanning, the MH simply listens to available beacons. In active scanning, the MH sends a probe request to a targeted set of APs that are capable of receiving its probe. Each AP that receives the probe responds with a probe response that contains the same information available in a regular beacon with the exception of the TIM. The probe response thus serves as a *solicited beacon*. The mobile chooses the AP with the strongest beacon or probe response and sends a *reassociation request* to the new AP. The reassociation request contains information about the MH as well as the old AP. In response, the new AP sends a *reassociation response* that has information about the supported bit rates, station ID, and so on needed to resume communication. The old AP is not informed by the MH about the change of location. So far, each WLAN vendor had some form of proprietary implementation of the emerging IAPP standard for completing the last stage of the HO procedure (intimating the old AP about the MH's change of location). The IAPP protocol employs two protocol data units (PDUs) to indicate that an HO has taken place. These PDUs are transferred over the wired network from the new AP to the old AP using UDP/IP. If the AP does not have an IP address, an 802.11 subnetwork access protocol (SNAP) is employed for transferring the PDUs.

Comparison of Handoff Procedures in IEEE 802.11, GPRS, and CDPD

Even though the functionalities of IEEE 802.11, GPRS, and CDPD networks are different, the HO procedures have several similarities (Table 1). All the networks use a separate signal (beacon, BCCH, or reference channel) with a constant transmit power in order to enable RSS measurements for HO decisions. While the beacon in 802.11 is on the same channel as data, it is on different physical channels in both GPRS and CDPD (since the reference channel may not be the data channel and the uplink data channels are physically apart). The primary difference is the fact that circuit-switched voice net-

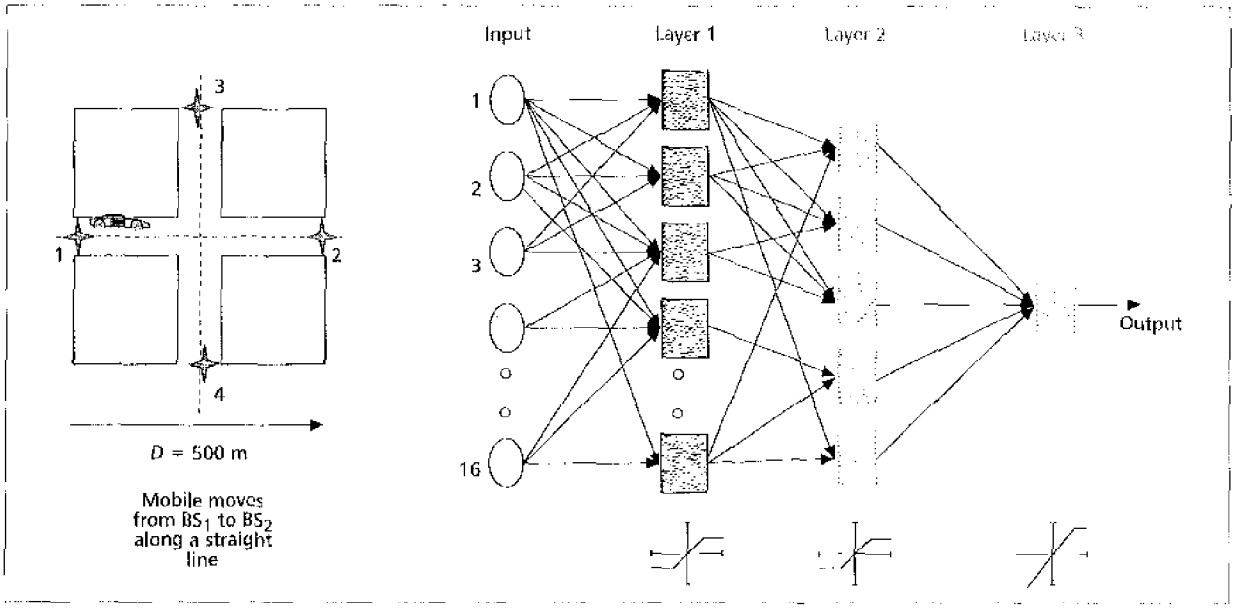


Figure 7. *The microcellular scenario and neural network architecture.*

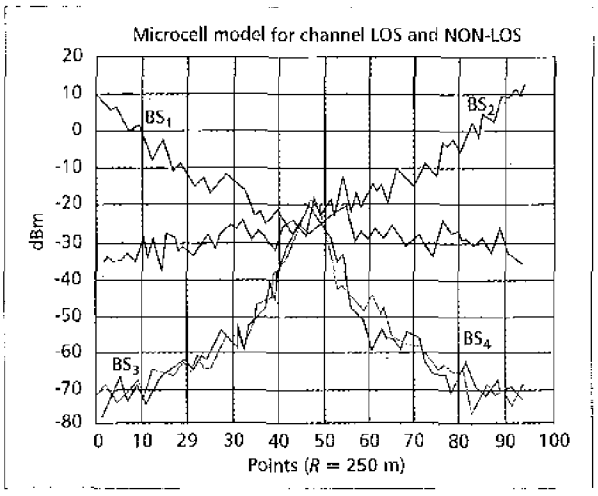
works have NCHO and data networks prefer MCHO. In both cases, channel monitoring is always performed at the terminal. When the HO control is with the network, the mobile has to transmit the measured information to the decision entity.

The Effects of Handoff Algorithms on Homogeneous Mobile Data Networks

HO decision time algorithms are usually proprietary and scenario-dependent. They are, however, independent of the HO architecture. In CDPD and GPRS, the algorithms are similar to the AMPS and GSM algorithms simply because it is desirable for a data MH to make an HO at around the same location at which a voice terminal would. This keeps the traffic loads balanced in each cell. As mentioned earlier, there are several emerging algorithms that employ complicated techniques to arrive at the correct time for HO. Pattern recognition HO algorithms [1] train a system using available metrics (e.g., RSS) and the locations where HOs should be made so that the system acquires knowledge of the RSS patterns at such locations. Neural networks (NNs) can be used for such pattern classification [18, 19]. The basic idea of an NN is to design a system that takes a few inputs that appear to be random but have some pattern associated with them and, regardless of the nature of the inputs, adjusts the parameters of the system in order to get some desired outputs through a learning process. By adjusting these system parameters the process of learning is completed, and the system can be considered as a black box that is capable of producing the desired output.

As an example, we consider a scenario of four identical BSs in a microcellular environment and an MH which is moving from the neighborhood of BS_1 toward BS_2 along a direct path, as shown in Fig. 7. It is assumed that all the BSs can provide the same service to the MH. The objective is to develop an NN system which takes a number of power samples from all the BSs and, using a pattern recognition technique, selects the BS which is most suitable, while minimizing the HO delay and ping-pong effect. The distance of each block is assumed to be $R = 250$ m. For the input of the NN we use the RSS from each BS. The output is a control signal that is zero as long as the MH is closer to BS_1 and one whenever the MH is closer to BS_2. The RSS level depends on the channel between the MH and each BS. For the purpose of this simulation we have used a microcellular path loss model with lognormal fading. Further details about this model can be found in [7, 20]. Some preprocessing of the input vectors also improves the performance of the neural networks. Figure 8 shows a sample of the RSS from four BSs as a function of location of the MH along the straight line connecting BS_1 and BS_2. The MH moves from a distance of 12.24 m from BS_1 to 487.76 m from BS_1. This distance is split into roughly 100 equally spaced points, and RSS samples are obtained every 4.76 m.

A three-layer back-propagation NN has been designed and is shown in Fig. 7. The inputs to this system are samples of RSS from each of the four BSs. These samples are taken using a sliding window of 4 samples/window. The NN is trained with the previously mentioned RSS and desired outputs. Once the neural network "black box" is designed, we feed as inputs a freshly generated input vector of RSS from the four BSs as the mobile moves from BS_1 to BS_2. We use the same inputs to test two traditional algorithms that compare only the RSS and use the RSS with a hysteresis margin. As our performance measures, we use the number of HOs and the delay in executing the HO. Figure 9 demonstrates some of the results of our simulations. In Fig. 9a the *ideal* control signal that switches the mobile from BS_1 to BS_2 at the midpoint between the two base stations is shown. If the simple RSS criterion is used (the mobile is switched to the BS providing the largest RSS), several unnecessary HOs occur. In Fig. 9b, a sample case is shown where seven HOs take place, six of them unnecessary. If a hysteresis margin of 10 dB is employed to reduce this ping-pong effect, only one HO takes place. However, the delay in HO execution is large. In Fig. 9c, the effect of adding a hysteresis is shown. The HO takes place at around 267 m from BS_1 that is about 17 m away from the correct location where the HO should have been executed. Figure 9d shows the performance of an NN with some preprocessed input. There is exactly one HO executed at approximately 255 m from BS1. The penalty for this is in the increased algorithm complexity and training of the NN that has to be done beforehand.

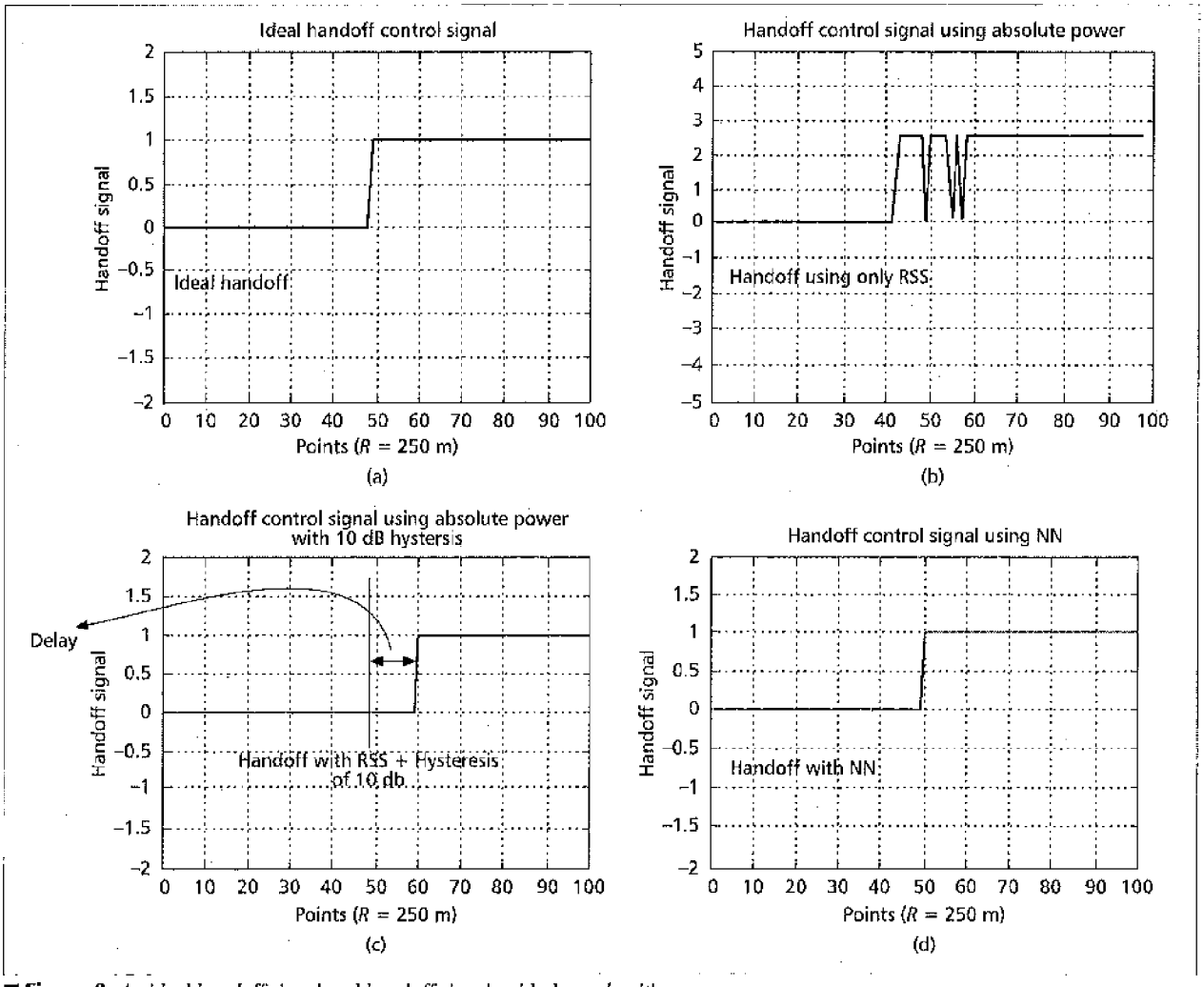


■ **Figure 8.** *RSS from four BSs in the microcellular scenario.*

Handoff Architecture in Hybrid Networks (Intertech or Vertical Roaming)

The motivation for hybrid networks arises from the fact that no one technology or service can provide ubiquitous coverage, and it will be necessary for a mobile terminal to employ various points of attachment to maintain connectivity to the network at all times. The natural trend has been toward utilizing local-coverage high-bandwidth data networks such as IEEE 802.11 whenever available and to switch to an overlay service such as a GPRS network with low bandwidth when the coverage of a WLAN is not available. For example, the Bay Area Research Wireless Access Network (BARWAN) [21] is implementing roaming between WaveLAN wireless LANs and Metricom packet data services. They differentiate between *horizontal* and *vertical* HOs [22]. Horizontal HOs refer to HO between BSs using the same kind of network interface. Vertical HOs refer to HOs between BSs employing different wireless technologies. An *upward vertical HO* occurs from a BS of a service with a smaller-size cell to a BS of a service with wider coverage. A *downward vertical handoff* takes place in the reverse direction. While the upward vertical HO occurs when the MH moves out of coverage of a service, the downward vertical HO has to take place when coverage of a service with smaller coverage becomes available when the user still has connection to the service with wider coverage.

In the following sections we discuss example architectures for a hybrid network consisting of GPRS and IEEE 802.11 that were considered in [23, 24]. Five architectures for the



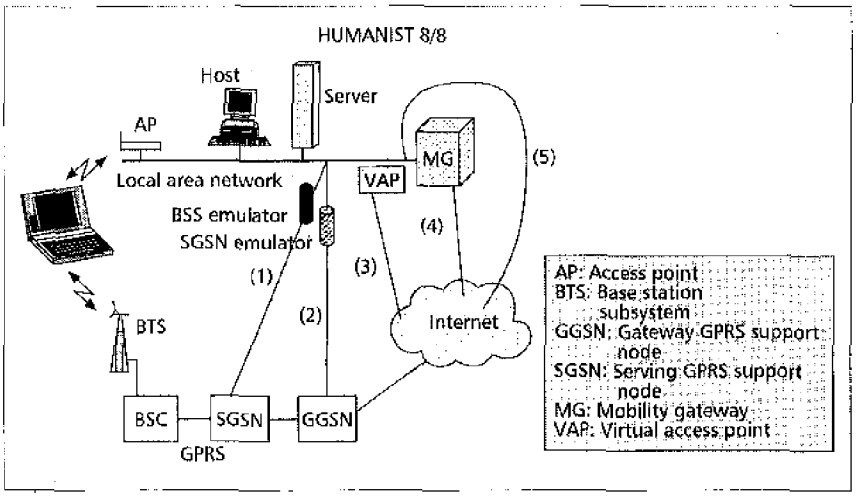
■ **Figure 9.** *An ideal handoff signal and handoff signals with three algorithms.*

example hybrid network, based on emulation of GPRS entities within the WLAN, mobile IP, a virtual AP, and a mobility gateway (proxy) are described and compared. The mobility gateway and mobile IP approaches are selected for more detailed discussion.

An Example of Hybrid Mobile Data Networks: GPRS and IEEE 802.11

Figure 10 shows five different architectures [23, 24] for implementing HO between GPRS and IEEE 802.11 networks. The objective here is to reduce, as far as possible, major changes to existing networks and technologies, especially at the lower layers such as MAC and physical layers. This will ensure that existing networks will continue to function as before without requiring current users to change to the new approach. The implementation involves incorporating new entities or protocols that operate at the network or higher layers to enable intertech roaming that will be transparent to the mobile user to the extent possible.

The first two architectures involve connecting the WLAN to the GPRS network through GPRS entities such as the SGSN and GGSN. In these cases, the WLAN will appear to be a GPRS cell or RA, respectively. GPRS will be the *master network* and the WLAN will be the *slave network*. This means that mobility will be handled by GPRS, considering the WLAN one of its cells or RAs. This may require dual-mode PCMCIA cards to access two different physical layers. In addition, all traffic will first reach the GPRS SGSN or GGSN before reaching its final destination even if the final destination is in the WLAN/LAN itself. This will potentially cause bottlenecks in the GPRS network. The virtual AP (3) reverses the roles played by the GPRS and WLAN in the first two architectures. Here, the WLAN is a master network and the GPRS is the slave network. Mobility is managed according to the IEEE 802.11 and IAPP specifications by the WLAN. The fourth approach introduces a *mobility gateway* (MG) between the GPRS and WLAN networks. The MG is a proxy implemented on either the GPRS or the WLAN sides, and will handle the mobility and routing issues. The last architecture employs mobile IP to handle the issue of mobility management. Here, GPRS and WLAN are peer networks. Certain changes will be needed to support intertech roaming on both the terminal and network sides. We consider only the last two approaches since the first three are inefficient and render one of the two networks as a slave network. We do not address the issue of location management. It is assumed that when an MH is attached to one particular network, the location management functionality of that network is used. Either the proxy tracks the network to which the MH is attached, or mobile IP location management features are used to determine to which network the MH is connected at a particular time.



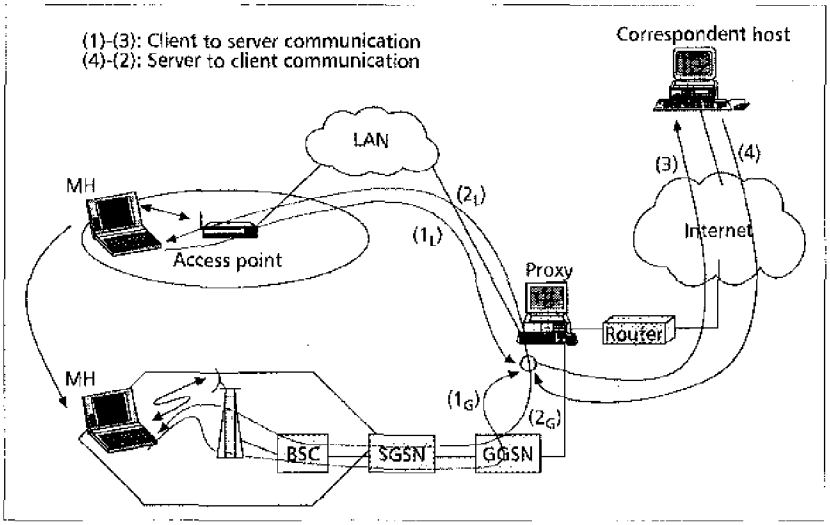
■ **Figure 10.** *GPRS–WLAN interconnection architectures.*

Mobility Gateway/Proxy-Based Architecture

In Fig. 11, the general architecture for a proxy- or mobility-gateway-based approach to intertech roaming is shown. An intermediate server is placed in the network so that any traffic to and from the MH is forced to pass through it. Consequently, it is possible for this entity to perform certain functionalities that may be required on behalf of the MH in a manner that is transparent to the rest of the network. When the MH is attached to an AP (and thus connected to a WLAN), the communication path between the MH and a correspondent host (CH) on the Internet will be (1_L)–(3). The CH–MH communication path will be (4)–(2_L). When the MH is on the GPRS network, these paths will be (1_G)–(3) and (4)–(2_G), respectively. It should be observed that segments (3) and (4) in any of the paths *do not change* regardless of where the MH is located. Only links (1) and (2) will be continually changing depending on the movement of the MH. Clearly, then, only the communication links between the MH and the proxy server (PS) are subject to changes, whatever they may be. Allowing the proxy–mobile connection to change while maintaining the proxy–CH connection unchanged supports mobility. The changes needed are with respect to the communication protocols between the MH and the PS. Both the MH and PS are presumably under common ownership [25], and hence it will be quite easy to tune the required characteristics to the specific needs of the users of that system.

There are several advantages to employing a proxy architecture for intertech roaming. There is the possibility of further minimizing the encapsulation and routing inefficiencies associated with mobile IP. However, the real reduction in overhead may not be very significant due to the need for additional control protocols. If the proxy is under the control of the same organization that owns the MHs, it is possible to configure the proxy to support the peculiar needs of its population of MHs. An optimized protocol may be run between the mobile and proxy depending on the link being employed. The proxy can manage the limited resources of certain connections more efficiently. For example, when the mobile is connected to GPRS, the proxy can drop structured data such as e-mail headers, frames in an MPEG stream, and so on selectively or drop unstructured data using some heuristics, as in the case of quoted text in e-mails [26]. The proxy can compress data or delay data by spooling, resegment packets, and respond to ICMP messages on behalf of an MH to improve the performance of the wireless link. Proxies are already in place in many organizations as firewalls or Web caching servers. These may be reused for mobility management and intertech roaming. Proxies can be used for logging the characteristics of a connection, details of which may be usefully employed in various applications including accounting and fraud management.

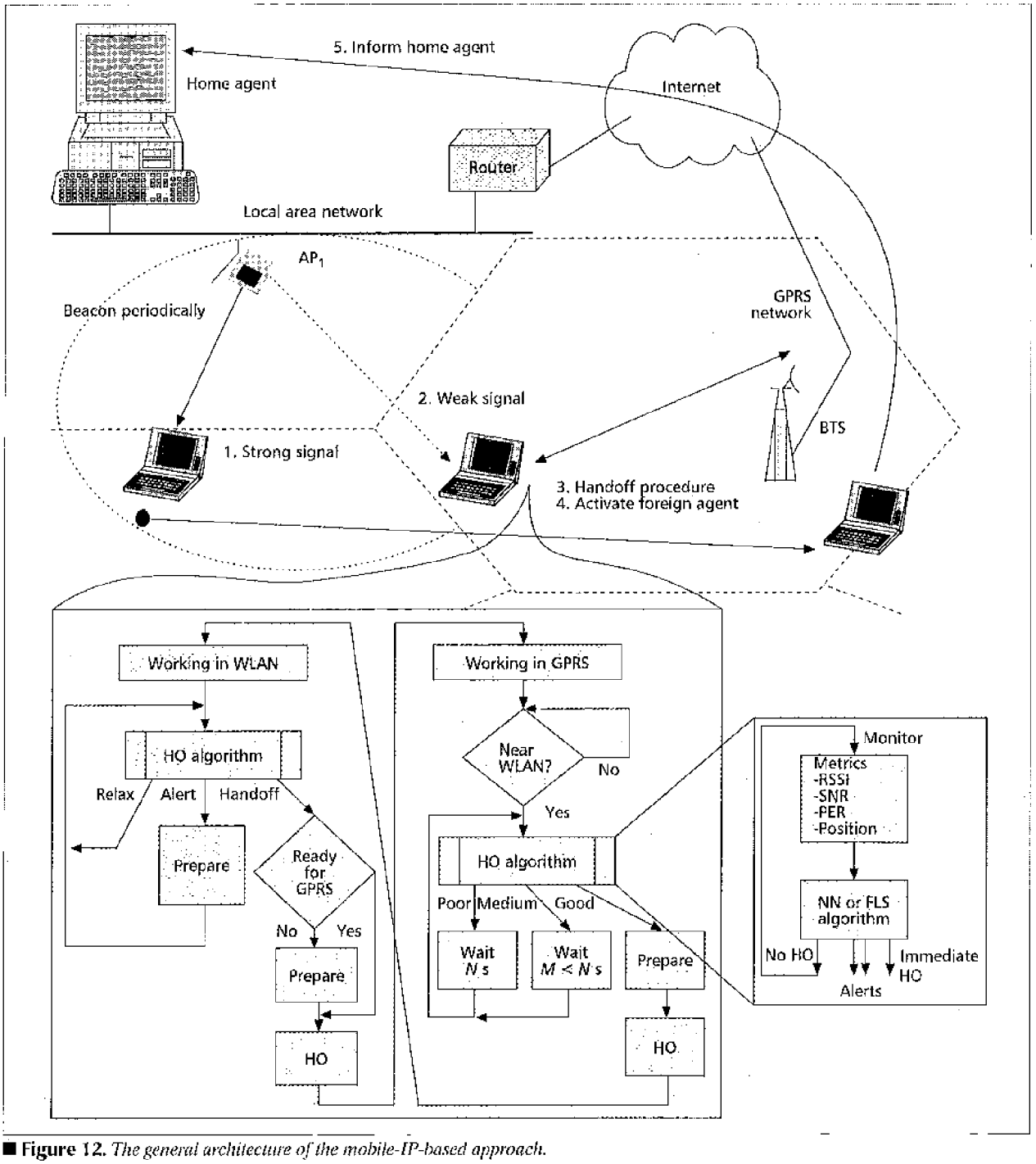
The main disadvantages of the proxy architecture are as follows. The architecture is not standardized, requiring proprietary protocols for intertech roaming. The performance of a proxy is poor since significant latency is added to the client-server communication path. Potentially, the end-to-end semantics of the transport protocol may also be violated. If a single proxy is employed and it fails, it may result in the failure of the entire network, and there is a need to have some fault tolerance. In addition to the significant issue of developing protocols for mobility management with the proxy architecture, there are some more open issues. The placement and number of proxies to be employed may depend on the situation. It is preferable to have the proxy connected to the last links of each service the MH may use so that it can gather information about the quality of each last link. However, the ownership of such a proxy will be contentious. The number of proxies that have to be placed for optimum performance is also subject to network conditions, and an easy answer is not possible.



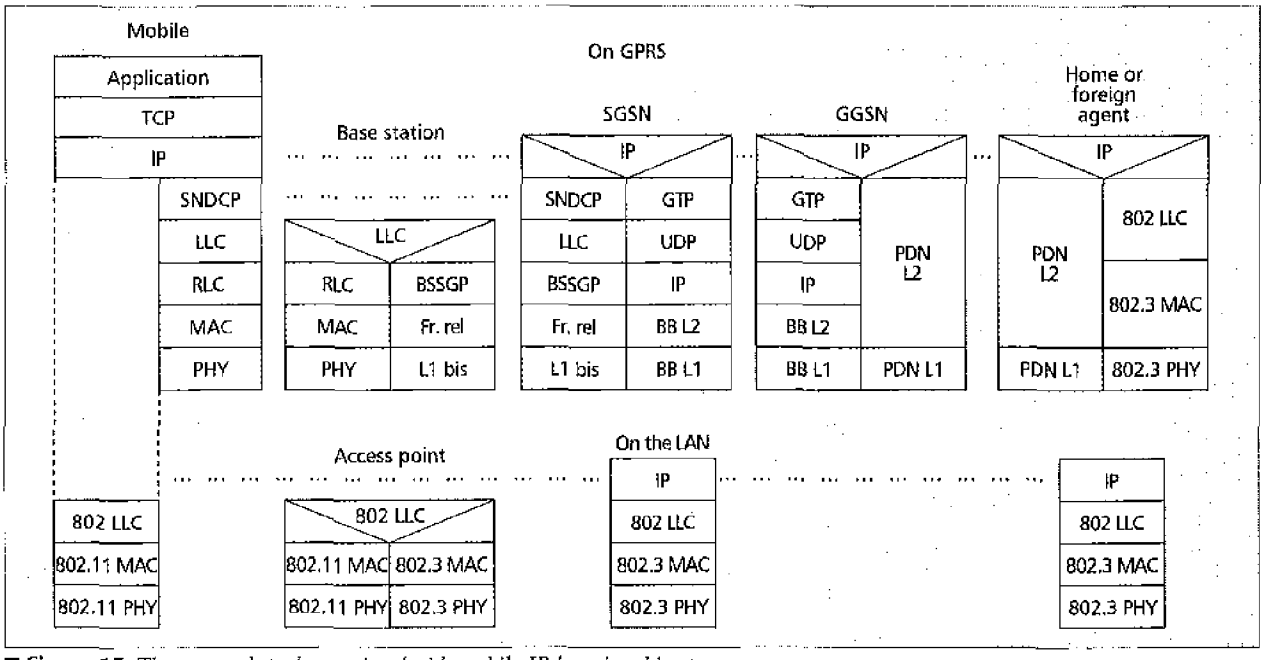
■ **Figure 11.** *The general architecture of the proxy-/mobile-gateway-based approach to intertech roaming.*

Mobile-IP-Based Architecture

This approach employs mobile IP [27] to restructure connections when an MH roams from one data network to another. Outside of its home network the mobile host is identified by a *care-of address* associated with its point of attachment and a collocated *foreign agent* (FA) that manages decapsulation and delivery of packets. The mobile host registers its care-of address with a *home agent* (HA). The HA resides in the home network of the MH and is responsible for intercepting datagrams addressed to the MH's home address as well as encapsulating and tunneling them to the associated care-of address. Datagrams to an MH are always routed through the HA. Datagrams from the MH are relayed along an optimal path by the Internet routing system, although it is possible to employ reverse tunneling through the HA. Figure 12 shows the general architecture of the mobile-IP-based approach. As shown, there it is assumed that the WLAN is the home network (with



■ **Figure 12.** *The general architecture of the mobile-IP-based approach.*



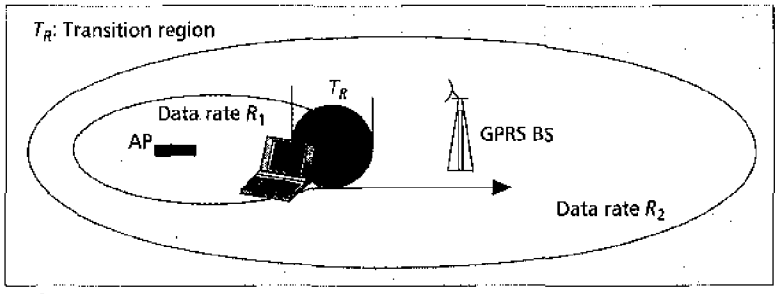
■ **Figure 13.** *The protocol stack associated with mobile-IP-based architecture.*

the home agent residing on the home LAN), and the GPRS network is the foreign or visited network. Figure 13 shows the user plane protocol stack associated with this implementation. Clearly, both the GPRS [9] and WLAN networks are peer networks. The protocol stack, however, does not show the functionality of the HA and FAs, which exist at the IP layer in each network.

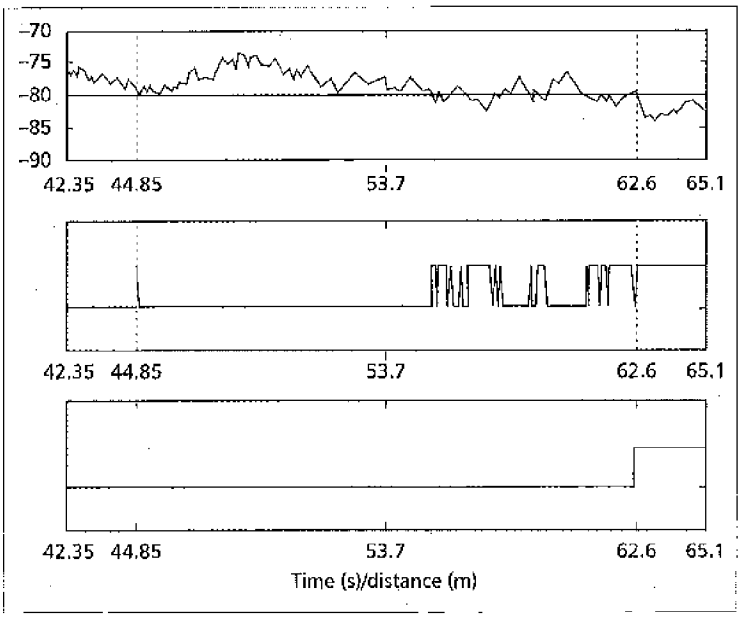
The general architecture describes the functions of elements of the network when we have an HO from WLAN to GPRS or the other way around. First, we describe the step-by-step procedure for the MH handing off from the WLAN connection to the GPRS connection. The following steps occur while the mobile moves away from the coverage of the WLAN within the GPRS coverage. The signal received from the AP in the WLAN is initially strong. The signal from the AP becomes weaker as the MH moves away. The HO algorithm in the MH decides to dissociate from the WLAN and associate with GPRS. The FA in the MH is activated, and the MH uses the new visiting IP address. The HA in the WLAN is informed about the new IP address. In the reverse situation, when the MH is connected to GPRS and realizes that a WLAN is available, the following steps will occur. The signal from any WLAN is initially not detected. The MH then detects a signal from the AP of a WLAN. The HO algorithm decides on making an HO from GPRS to the WLAN. The FA in the MH is deactivated, and the home IP address is used. The HA in the WLAN is instructed by the MH to no longer do a proxy address resolution on its behalf. Here we are assuming movement only between the home WLAN and GPRS.

The HO problem is twofold: from WLAN to GPRS and from GPRS to WLAN. The difference between these is that a user operating in GPRS does not have to worry about losing the connection. Therefore, the user attached to the overlay network just occasionally checks for the availability of the underlay network. The WLAN-to-GPRS HO triggering algorithm is more crucial for reliable operation of the system, since an MH moving away from the underlay network coverage may suddenly experience severe degradation of service and will have to HO very fast to maintain the higher-layer connection. For these reasons the triggering algorithm is decomposed into two parts, underlay (WLAN) and overlay (GPRS). To enable reuse of the code, both parts use the same NN; only the frequency of its invocation and the action on its output are different. The larger block in Fig. 13 shows the block diagram of the algorithm. Simulating the results for a complex system of several BSs and APs is complicated, and a simple system as described below provides preliminary insights into what might be appropriate performance measures for initiating HO.

Figure 14 shows a simple "moving away scenario" where the MH moves away from a WLAN AP. One can infer in this scenario (based on [28]) that an efficient algorithm will try to use the services of the AP as long as possible and do the HO to the BS as the last alternative. This is unlike the microcellular scenario in the earlier section, where the best possible time to HO was when the MH was midway between two identical BSs. In this case the AP has much higher priority than the BS. The reason for this difference in HO strategy is that it is almost always preferable *not* to make an HO when the wide-area service provides a data rate two orders of magnitude smaller than the local-area service, since transmitting at 2 Mb/s for 1 s is preferable to transmitting at 19.2 kb/s for 100 s. One possible implementation of such a scheme will be to employ time hysteresis. Here, the MH will take the samples of the RSS from the AP and compare it with a predefined



■ **Figure 14.** *A moving away scenario.*



■ **Figure 15.** *Simple RSS-based and NN-based algorithms for HO from WLAN to GPRS.*

threshold (e.g., $\chi = -80$ dBm); if a predefined number of consecutive samples are below the threshold, the MH initiates the HO; otherwise, it will persist with the AP. This is also called a *dwell timer* (time for which the MH persists with a point of attachment even if the signal strength is low) [1]. The region where the RSS first falls below χ and is last above χ is called the *transition region* T_R. An HO should thus be made only once: at the edge of T_R.

The first plot of Fig. 15 shows the RSS profile within T_R and a threshold of –80 dBm. A time hysteresis algorithm does the HO at the last crossing of the threshold χ (in this case at t = 62 s). But since this time is random, an NN is again preferable. Figure 16 shows a suggested architecture for such a network. This network consists of three layers. The input to the system consists of samples of the RSS from the AP (in a sliding window of five samples). The output of the system is a binary signal, zero meaning the MH should continue communicating with the AP and one implying that the MH should make the HO and communicate with the BS. The second plot in Fig. 15 shows an HO control signal using only RSS, and the third plot, an HO control signal that is the output of the NN architecture of Fig. 16. Clearly the NN architecture is preferable since it eliminates the ping-pong effect. There is a need for further work in this area for more complicated scenarios and performance measures.

Conclusions

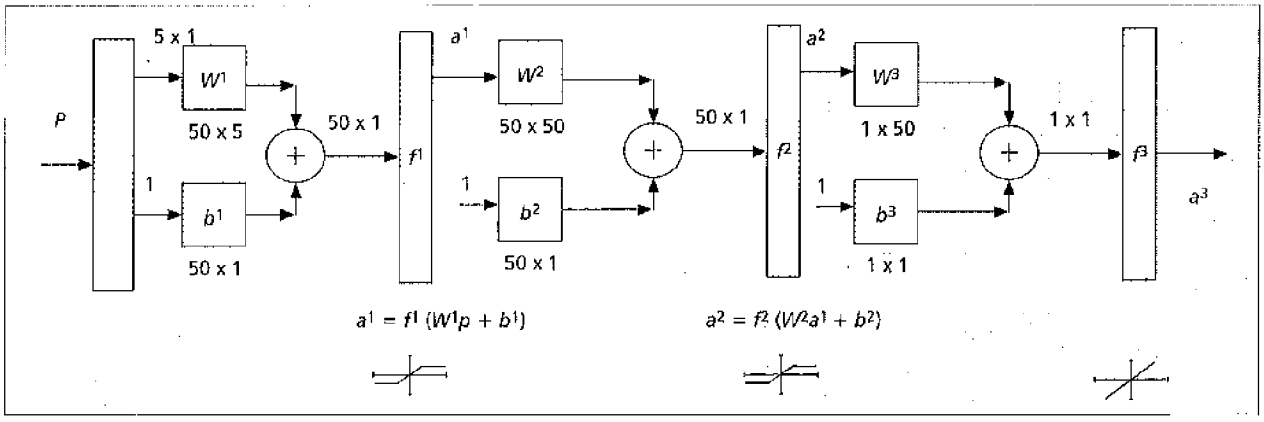
This article presents an overview of the issues related to handoff with particular emphasis on hybrid mobile data networks. Handoff issues can be classified into two independent parts: architectural issues and HO decision time algorithms. The former are open and standardized, but the latter are usually proprietary. Handoff architectures in mobile data networks and high-speed WLANs have several similarities. Traditional HO algorithms have poor performance and are being replaced by emerging advanced HO algorithms. An example neural network algorithm not only reduces the number of unnecessary HOs, but also minimizes the HO delay in a microcellular scenario. No one technology or service can provide ubiquitous coverage, and it will be necessary for a mobile terminal to employ various points of attachment to maintain connectivity to the network at all times, resulting in the need for HO in a hybrid network of mobile data and WLAN. The HO architectural issues related to hybrid networks are discussed through an example of a hybrid network that employs GPRS and IEEE 802.11. Mobile IP is the most suitable architecture for HO in a hybrid network. Again, neural-network-based algorithms perform better than traditional algorithms for handoff time in hybrid networks.

Acknowledgments

The authors would like to thank TEKES, Nokia, Sonera Ltd., and Finnish Airforce for supporting this project. We would like to thank Dr. Jaakko Talvitie, Dr. Ali Zahedi and Yan Xu for participation in this project.

References

[1] G. P. Pollini, "Trends in Handover Design," *IEEE Commun. Mag.*, Mar. 1996.
[2] R. Cacares and L. Iftode, "Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments," *IEEE JSAC*, June 1995, pp. 850–57.
[3] N. D. Tripathi, J. H. Reed, and H. F. VanLandingham, "Handoff in Cellular Systems," *IEEE Pers. Commun.*, Dec. 1998.
[4] S. Tekinay and B. Jabbari, "Handover Policies and Channel Assignment Strategies in Mobile Cellular Networks," *IEEE Commun. Mag.*, vol. 29, no. 11, Nov. 1991.



■ **Figure 16.** *A neural network architecture for handoff from WLAN to GPRS.*

[5] G. Liodakis and P. Stravroulakis, "A Novel Approach in Handover Initiation for Microcellular Systems," *Proc. VTC '94*, Stockholm, Sweden, 1994.
[6] R. Rezaiifar, A. M. Makowski, and S. Kumar, "Optimal Control of Handoffs in Wireless Networks," *Proc. IEEE VTC '95*, 1995, pp. 887–91.
[7] N. Tripathi, "Generic Adaptive Handoff Algorithms using Fuzzy Logic and Neural Networks," Ph.D. Thesis, VA Polytechnic Inst. and State Univ., Aug. 1997.
[8] J. Vallström, "GPRS," WiLU tech. rep., Oct. 1997.
[9] C. Bettstetter, H.-J. Vogel, and J. Eberspacher, "GSM Phase 2+ General Packet Radio Service GPRS: Architecture, Protocols, and Air Interface," *IEEE Commun. Surveys*, 3rd qtr. 1999.
[10] J. Cai and D. J. Goodman, "General Packet Radio Service in GSM," *IEEE Commun. Mag.*, Oct. 1997.
[11] "CDPD System Specification," Rel. 1.1, CDPD Forum Inc, Jan. 19, 1995.
[12] M.S. Taylor, W. Waung, and M. Banan, *Internetwork Mobility: The CDPD Approach*, Prentice Hall, 1997.
[13] K. C. Budka, H. Jiang, and S. E. Sommers, "Cellular Digital Packet Data Networks," *Bell Labs Tech. J.*, Summer 1997.
[14] The IEEE 802.11 Standard: http://grouper.ieee.org/groups/802/11/index.html
[15] K. Pahlavan, A. Zahedi, and P. Krishnamurthy, "Wideband Local Access: WLAN and WATM," *IEEE Commun. Mag.*, Nov. 1997.
[16] B. P. Crow *et al.*, "IEEE 802.11 Wireless Local Area Networks," *IEEE Commun. Mag.*, Sept. 1997.
[17] H. Moelard, "Inter Access Point Protocol: Wireless Networking Distribution System Communication," Stds. track Internet draft, Mar. 1998.
[18] S. Haykin, "Neural Networks Expand SP's Horizons," *IEEE Signal Proc.*, Mar. 1996.
[19] D. Cox, "Pattern Recognition Handoff Algorithms for Cellular Communication Systems," project description submitted to Stanford Ctr. for Telecommun., Sept. 1996.
[20] J.-E. Berg, R. Bownds, and F. Lotse, "Path Loss and Fading Models for Microcells at 900 MHz.," *42nd IEEE VTC*, Denver, CO, May 1992, pp. 666–71.
[21] E.A. Brewer *et al.*, "A Network Architecture for Heterogeneous Mobile Computing," *IEEE Pers. Commun.*, Oct. 1998.
[22] M. Stemm and R. Katz, "Vertical Handoffs in Wireless Overlay Networks," *ACM MONET*, Summer 1998, pp. 335–50.
[23] P. Krishnamurthy *et al.*, "Handoff In 3G Non-Homogeneous Mobile Data Networks," *Euro. Microwave Week*, Netherlands, Oct. 1998.
[24] P. Krishnamurthy et al., "Scenarios for Inter-Tech Mobility," WiLU tech. rep., Jan. 1998.
[25] D. A. Maltz and P. Bhagwat, "MSOCKS: An Architecture for Transport Layer Mobility," *IEEE INFOCOM*, 1998.
[26] B. Zenel and D. Duchamp, "General Purpose Proxies: Solved and Unsolved Problems," *Proc. HOT-OS VI*, May 1997.
[27] C. E. Perkins, *Mobile IP: Design Principles and Practices*, Addison Wesley Communications Series, 1997.
[28] A. Hatami *et al.*, "Analytical Framework for Handoff in Non-Homogeneous Mobile Data Networks," *PIMRC '99*, Osaka, Japan, Sept. 1999.

Biographies

KAVEH PAHLAVAN [F] (kaveh@ece.wpi.edu) is a professor of ECE and director of the Center for Wireless Information Network Studies, Worcester Polytechnic Institute, Massachusetts. His area of research is broadband wireless local networks. His previous research background is on modulation, coding, and adaptive signal processing for digital communications. He has contributed to more than 200 technical papers and presentations in various countries. He is the principal author of *Wireless Information Networks* (Wiley, 1995. He has been a consultant to many companies in the United States, Finland, and Japan. Before joining WPI, he was director of advanced development at Infinite Inc., Andover, Massachusetts,. working on data communications. He started his career as an assistant professor at Northeastern University, Boston, Massachusetts. He is Editor-in-Chief of the *International Journal on Wireless Information Networks*. He was program chair and organizer of the IEEE Wireless LAN Workshop, Worcester, in 1991 and 1996 and the organizer and the technical program chairman of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications, Boston, MA, 1992 and 1998. For his contributions to wireless networks he was the Westin Hadden Professor of Electrical and Computer Engineering at WPI during 1993-1996, and became a fellow of Nokia in 1999.

PRASHANT KRISHNAMURTHY [M] (prashant@tele.pitt.edu) is an assistant professor in the Department of Information Science and Telecommunications at the University of Pittsburgh. Earlier, he was a research assistant at the Center for Wireless Information Network Studies (CWINS) at Worcester Polytechnic Institute working on radio propagation modeling for geolocation applications and mobility management in mobile data networks, where he also got his Ph.D. in 1999. His interests are in the areas of radio propagation, wireless data networks, and wireless network security.

AHMAD HATAMI (ahatami@lucent.com) is a software engineer in Lucent Technologies Internetworking systems division. He joined Lucent in 1999 and has worked on layer 2 and 3 protocols for multiservices core switches. His interests are in the areas of communication and data networking. Prior to joining Lucent he worked with 3Com developing variable bit rate sources. Prior to that he was a research assistant at the Center for Wireless Information Network Studies working on Mobile IP and Handover in mobile data networks. He received his M.Sc. in communication and computer networking from Worcester Polytechnic Institute, and his B.Sc. in electrical engineering from the University of Tehran.

MIKA YLIANTTILA (mika.ylianttila@oulu.fi) received his M.Sc. degree in electrical engineering from Oulu University, Finland, in 1998. He is now working as a project manager in the Centre for Wireless Communications at Oulu University, and is working on his Ph.D. in the area of mobility management and system architecture issues in fourth-generation wireless networks. His professional interests include IP protocol evolution, wireless optimizations, location-based services, and real-time architectures.

JUHA-PEKKA MAKELA (juha-pekka.makela@ee.oulu.fi) received his M.Sc. degree in electrical engineering from Oulu University, Finland, in 1997. He is currently a research scientist in the Centre for Wireless Communications at Oulu University, working for his Lich. Tech. degree. His interests are intelligent handoff techniques, ad hoc networks, and CDMA network issues.

ROMAN PICHNA (roman.pichna@nokia.com) received an Ing. degree from Slovak Technical University, Bratislava, Slovakia, in 1987 and a Ph.D. degree from the University of Victoria, Canada in 1996. He worked at the Centre for Wireless Communications, Oulu University, Finland until 1998. He is currently with Nokia Networks, Radio Access Systems Research in Helsinki. His professional interests are in the areas of cellular radio network simulations, medium access control protocols, network architecture design, and network protocols.

JARI VALLSTRÖM (jari.vallstrom@nokia.com) received an M.Sc. degree in electrical engineering from Oulu University, Finland, in 1995. He worked at the Centre for Wireless Communications, Oulu University, until 1998. He is currently with Nokia Mobile Phones. His professional interests are in the areas of GSM evolution, standardization, and network architectures.