Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Communications xx (xxxx) 1–23
computer communications
www.elsevier.com/locate/comcom

Mobility management across hybrid wireless networks: Trends and challenges

Farhan Siddiqui, Sherali Zeadally *

High-Speed Networking Laboratory, Department of Computer Science, Wayne State University, Detroit, MI 48202, USA

Received 14 July 2005; revised 31 August 2005; accepted 9 September 2005

Abstract

Future generation wireless networks are envisioned to be a combination of diverse but complementary access technologies. Internetworking these types of networks will provide mobile users with ubiquitous connectivity across a wide range of networking environments. The integration of existing and emerging heterogeneous wireless networks requires the design of intelligent handoff and location management schemes to enable mobile users to switch network access and experience uninterrupted service continuity anywhere, anytime. Real deployment of such mobility strategies remains a significant challenge. In this article, we focus on handoff management. We discuss in detail handoff decision and implementation procedures and present recent handoff techniques that aim at providing mobility over a wide range of access technologies. We also discuss some of the capabilities of mobile terminals that are necessary to implement seamless mobility over hybrid wireless networks. Furthermore, we also present and discuss limitations of recent handoff design architectures and protocols as well as outstanding challenges that still need to be addressed to achieve portable and scalable handoff solutions for continuous connectivity across wireless access networks.

Keywords: Wireless; Heterogeneous; Overlay; Mobility; Handoff; Multimode; Protocol

1. Introduction

Mobile wireless technology has gained tremendous popularity due to its ability to provide ubiquitous information access to users on the move. However, presently, there is no single wireless network technology that is capable of simultaneously providing a low latency, high bandwidth, and wide area data service to a large number of mobile users. *Wireless Overlay Networks* [1]—a hierarchical structure of room-size, building-size, and wide area data networks solve the problem of providing network connectivity to a large number of moving consumers in an efficient and scalable way. In an overlay

Abbreviations 3G, third generation; 3GPP, third generation partnership project; AAA, authentication, authorization and accounting; ACK, acknowledgement; ADSL, asymmetric digital subscriber line; AR, access router; ASCONF, address configuration; BS, base station; CA, congestion avoidance; CAR, candidate access router; CARD, candidate access router discovery; CD, communication daemon; CDMA, code division multiple access; CN, correspondent node; cSCTP, mobile stream control transmission protocol; CTAR, context transfer activation request; CTD, context transfer data; CTP, context transfer protocol; CTR, context transfer request; DAB, digital audio broadcasting; DCCP, datagram congestion control protocol; DECT, digital enhanced cordless telephone; DHCP, dynamic host configuration protocol; DMT, discrete multitone; DNS, domain name service; DSP, digital signal processors; DSSS, direct sequence spread spectrum; DVB, digital video broadcasting; EDGE, enhanced data rates for GSM evolution; FA, foreign agent; FDD, frequency division multiplexing; FHSS, frequency hopping spread spectrum; FS, fixed server; GPRS, general packet radio service; GSM, global system for mobile communication; GWFA, gateway foreign agent router; HA, home agent; HiperLAN, high performance radio local area network; HOPOVER, handoff protocol for overlay networks; HP, handoff-prepare; HSCSD, high-speed circuit-switched data; IETF, internet engineering task force; IMT, international mobile telecommunications; IP, internet protocol; L2, layer 2; LSS, location-service server; MN, mobile node; MR, multicast router; MSC, mobile switching center; mSCTP, mobile stream control transmission protocol; NACK, negative acknowledgement; NAR, next access router; OFDM, orthogonal frequency division multiplexing; PSTN, public switched telephone network; QAM, quadrature amplitude modulation; QoS, quality of service; QPSK, quadri-phase shift keying; RNC, radio network controller; RRC, radio resource control; SCTP, stream control transmission protocol; SDR, software-defined radios; SIP, session initiation protocol; SS, slow start; TCP, transmission control protocol; TDD, time division multiplexing; TDMA, time division multiple access; UMTS, universal mobile telecommunications system; VDC, virtual domain controller; VoIP, voice over internet protocol; WG, working group; WLAN, wireless local area network.

* Corresponding author. Tel.: +1 3135770731; fax: +1 3135776868.
E-mail address: zeadally@cs.wayne.edu (S. Zeadally).

0140-3664/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.comcom.2005.09.003

network, lower levels are comprised of high bandwidth wireless cells that provide a small coverage area. Higher levels in the hierarchy provide a lower bandwidth but a much wider access network. A mobile device with multiple wireless network interfaces can access these networks as it moves between different network environments. Next generation wireless systems typically constitute different types of access technologies [15]. The heterogeneity that will characterize future wireless systems instigates the development of intelligent and efficient handoff management mechanisms that can provide seamless roaming capability to end-users moving between several different access networks.

2. Wireless overlay networks

Fig. 1 shows a typical structure [2] of wireless overlay networks. First, the networks' service areas are overlapped. For example, the General Packet Radio Service (GPRS) network acts as an umbrella network to the Wireless Local Area Network (WLAN) network. Even the different cells of the same network overlap. This overlapping can be utilized to reduce service disruption, by simultaneously connecting to different subnets of the same access technology during transition from one network to another. Second, the networks support different data rates and cell sizes. For instance, IEEE 802.11b WLAN supports a data rate of 11 Mbps and GPRS a much lower data rate of about 9.6 Kbps. Third, because of the different characteristics of the networks involved, it is not possible to compare the signal powers received from the base stations of different networks to decide which network to connect to. Fourth, each network may offer a different level of reliability, security, quality of service etc. As mobile hosts move across different networks, a mechanism for conveying the new IP address to the correspondent nodes is required. Also, the power consumed by the network interfaces is different for each network technology and is directly proportional to the transmitted power. For example, the Code Division Multiple Access (CDMA) transmitted power is much higher as compared to WLANs.

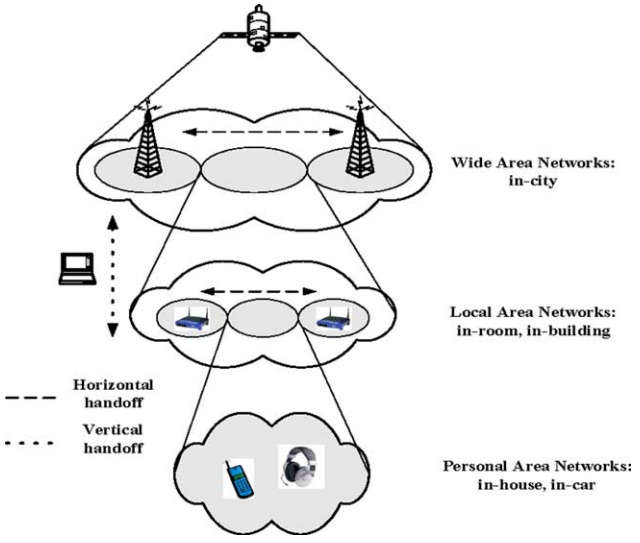

Fig. 1. Wireless Overlay Networks.

Currently, several wireless technologies and networks exist that capture different needs and requirements of mobile users. For high-data-rate local-area access, WLANs are satisfactory solutions [8]. For wide-area communications, traditional cellular networks may provide voice and data services. For worldwide coverage, satellite networks have been used extensively in military and commercial applications. Since, different wireless networks are complementary to each other, their integration will empower mobile users to be connected to the system using the 'best available' access network that suits their needs. Next generation wireless systems are envisaged to be a combination of a plethora of networking technologies. Fig. 2 illustrates a typical network architecture for next generation heterogeneous systems.

Some key features [3] of next generation heterogeneous access networks include:

- High usability with anytime, anywhere connectivity.
- Support for multimedia services with low transmission cost.
- Integrated access networks with a common IP-based core.
- Use of multimodal devices (capable of supporting various types of network access technologies).
- Support for telecommunication, data and multimedia services.
- Support for personalized services.
- Support for integrated service access from various service providers.

Table 1 shows some of the technologies, which will be part of future heterogeneous systems along with their characteristics.

3. Handoffs in wireless overlay networks

Handoff is the process by which a mobile terminal keeps its connection active when it migrates from the coverage of one network access point to another. Different types of handoffs can occur in wireless overlay networks.

3.1. Horizontal vs. vertical handoff

Handoffs that occur between the access-points of the same network technology and are termed *horizontal handoffs* or *intra-system handoffs* [4]. In other words, horizontal handoffs occur between homogeneous cells of a wireless access system, example between two cells of a cellular system, etc. Handoffs that occur between different access-points belonging to different networks (example WLAN to GPRS) are referred to as *vertical handoffs* or *inter-system handoffs*. Thus, vertical handoffs are implemented across heterogeneous cells of access systems, which differ in several aspects such as bandwidth, data rate, frequency of operation, etc. The different characteristics of the networks involved make the implementation of vertical handoffs more challenging as compared to horizontal handoffs. The terms horizontal and vertical follow from the overlay network structure that has networks with increasing cell sizes at higher levels in the hierarchy. Vertical handoffs are

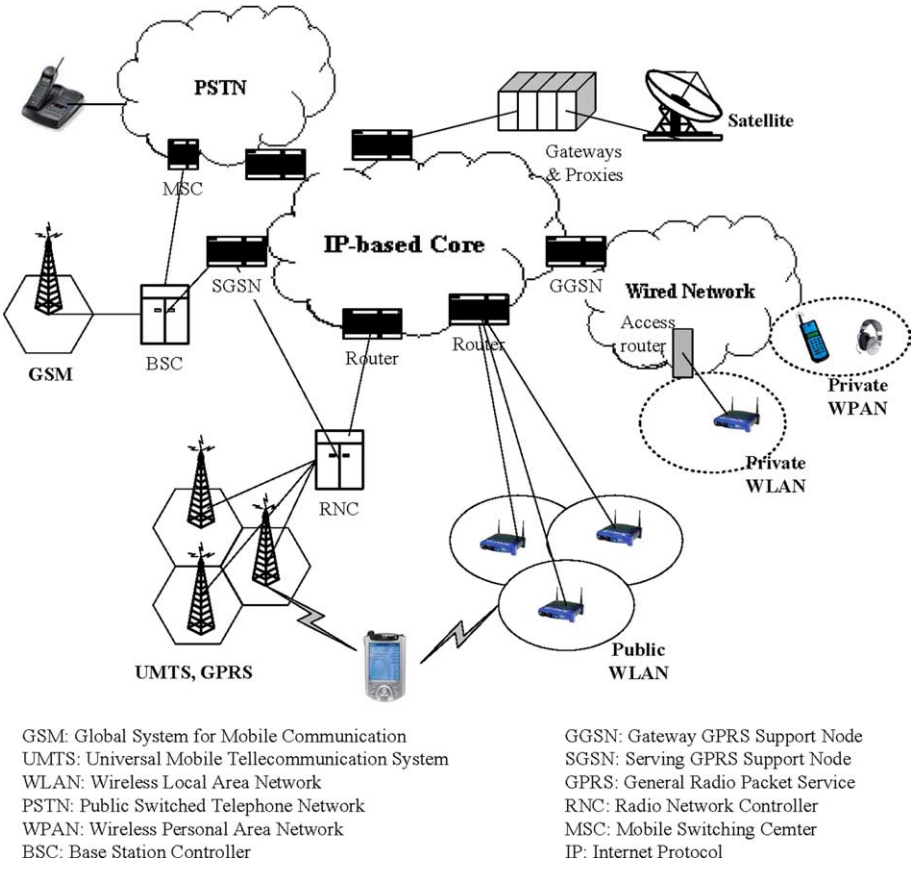

Fig. 2. A typical Architecture of Future Heterogeneous Access Systems.

generally of two types namely, upward and downward handoffs.

3.2. Upward-vertical handoff vs. downward vertical handoff

An *Upward-vertical handoff* is a handoff to a wireless overlay with a larger cell size and generally lower bandwidth per unit area. So, an upward-vertical handoff makes a mobile device disconnect from a network providing faster but smaller coverage (example WLAN) to a new network providing slower but broader coverage. A *downward vertical handoff* is a handoff to a wireless overlay with a smaller cell size, and generally higher bandwidth per unit area. A mobile device performing a downward vertical handoff disconnects from a cell providing broader coverage to one providing limited coverage but higher access speed [5]. A vertical handoff may be to an immediately higher or lower overlay, or the mobile host may 'skip' an overlay. For example, a mobile may handoff from an in-room network directly to a wide-area network, or vice versa.

3.3. Anticipated vs. unanticipated handoff

Anticipated handoffs are those which the mobile device will always want to perform, such as horizontal handoffs and upward vertical handoffs. For example, in the case of Wireless LANs (WLANs), a link layer trigger may indicate the presence of a new WLAN cell and the loss of coverage from the previous WLAN cell. In this case, the mobile device will definitely want to make a handoff to the new cell to continue the service. Similarly, all upward vertical handoffs can be anticipated, which means that a mobile device will always want to handoff to a network higher in the overlay (example switching to GPRS after receiving a link layer trigger indicating weak coverage of a WLAN). However, in the case of downward vertical handoffs, a link layer trigger can indicate to a mobile device that it is now under the coverage of a new network (example WLAN) and the mobile node may wish to execute the handoff. Downward vertical handoffs may be anticipated or unanticipated, such that a mobile device may already be under the coverage of the new network but may prefer to postpone a handoff based on requirements of the applications running on the mobile node, and may execute a handoff later, being already aware of the coverage status of the new network [4].

3.4. Hard vs. soft handoff

A handoff is *hard* if the MN can be associated with only one access point at a time. A *soft* handoff occurs if the MN can communicate with more than one access points during handoff. For example, if the MN is equipped with multiple network interfaces, it can simultaneously connect to multiple access-points in different networks during soft handoff [6]. Soft handoff may also be referred to as *make before break*handoff in

Table 1
Characteristics of heterogeneous wired and wireless access systems

	Access network type	Frequency	Data rate	Coverage	Cost	Technology
Wireless technologies	Bluetooth	2.4 GHz ISM band	Max. 721 kbps	0.1–10 m	Low	DSSS, FHSS
	IEEE 802.11g	2.4 GHz	54 Mbps	30–150 m	Low	OFDM
	IEEE 802.11b	2.4 GHz	11 Mbps	Up to 100 m	Low	DSSS
	IEEE 802.11a	5 GHz	20 Mbps	50–300 m	Low	OFDM, TDD
	HiperLAN2	5 GHz	54 Mbps	150 m max.	Low	OFDM
	IMT2000, UMTS	2 GHz	Max. 2 Mbps	30 m–20 Km	High	FDD, TDD
	IEEE 802.20	Below 3.5 GHz	Up to 9 Mbps	20 Km	High	OFDM
	IEEE 802.16	10–66 GHz	Max 70 Mbps	Over 50 Km	Low	OFDM
	GSM, GPRS, HSCSD, EDGE	900,1800,1900 MHz	9.6–384 Kbps	Up to 35 Km	High	TDMA, FDD
	Satellite	Up to 14 GHz	Max 144 Kbps	Several Kilometers	High	
	DAB	176–230 MHz; 1452–1467.5 MHz	1.5 Mbps	Up to 100 Km	Low	OFDM
	DVB-T	$<$860 MHz	5–31 Mbps	Up to 100 Km	Low	OFDM
	DECT/DECT Link	1880–1900 MHz	Up to 2 Mbps	Up to 50 m	Low	TDMA/TDD
Wired technologies	ADSL	Baseband	$<=$6.144 Mbps (down) 0.640 Mbps (up)	2–6 Km	Low	DMT
	Cable modem	60–860 MHz (down) 10–40 MHz (up)	$<$40 Mbps (down)$<$ 10 Mbps (up)	5–20 Km	Low	FDD/QAM/ QPSK

IMT, International Mobile Telecommunications; HSCSD, High-Speed Circuit-Switched Data; UMTS, Universal Mobile Telecommunication System; DAB, Digital Audio Broadcasting; GSM, Global System for Mobile Communication; DVB, Digital Video Broadcasting; EDGE, Enhanced Data Rates for GSM Evolution; ADSL, Asymmetric Digital Subscriber Line; DECT, Digital Enhanced Cordless Telecommunications; GPRS, General Radio Packet Service; HiperLAN, High Perfprmance Radio Local Area Network; OFDM, Orthogonal Frequency Division Multiplexing; TDMA, Time Division Multiple Access; FDD, Frequency Division Multiplexing; DSSS, Direct Sequence Spread Spectrum; TDD, Time Division Duplex; QAM, Quadrature Amplitude Modulation; QPSK, Quadri-Phase Shift Keying; FHSS, Frequency Hopping Spread Spectrum; DMT, Discrete Multitone.

which the mobile node's connection may be created at the target base station before the old base station connection is released. On the other hand, in the case of hard handoff or *break before make*handoff the new connection may be set up after the old connection has been torn down.

3.5. Mobile-controlled handoff vs. network-controlled handoff vs. mobile-assisted handoff

Under *network-controlled* handoff, the network makes the decision for handoff, while under *mobile-controlled* handoff, the mobile node must take its own signal strength measurements and make the handoff decision on its own. Under mobile-assisted handoff, the decision to handoff is made by the mobile node in cooperation with the network.

4. Vertical handoff process

The vertical handoff process may be divided into three phases [5]:

1. Network Discovery
2. Handoff Decision
3. Handoff Implementation

4.1. Network discovery

This is the process where a Mobile Node (MN) searches for reachable wireless networks. A MN with multiple interfaces must activate the interfaces to receive service advertisements, which are broadcasted by different wireless technologies. The MN will know a wireless network is reachable if its service advertisements can be heard. The simplest way to discover reachable wireless networks is to always keep all interfaces on. However, keeping an interface active all the time consumes the battery power even without receiving/sending any packets (as shown in Table 2). Therefore, to avoid keeping the idle interfaces always on is critical. Also the discovery time should be low so that the MN can benefit faster from the new wireless network.

The *power efficiency* and the *system discovery time* are the most critical considerations for system discovery methods' performance. The interface may be activated periodically to receive service advertisements. The activating frequency directly affects the system discovery time. The MN that activates the interfaces with high frequency may discover the reachable network quickly but its battery may run out very soon. The MN that activates the interface with low frequency may increase the power efficiency, but it may discover the reachable wireless networks slowly. There exists a tradeoff between the power efficiency and the system discovery time [5].

Table 2
Power consumption by 3G-CDMA and WLAN NICs

Technology	Transmit mode (W)	Receive mode (mW)	Idle mode (mW)
3G: CDMA 1X wireless modem NIC	2.8	495	82
ORINOCO IEEE 802.11B NIC	1.3	900	740

4.2. Handoff decision: traditional vs. next generation handoff strategies

Handoff decision is the ability to decide when to perform the handoff. A decision for vertical handoff may depend on several issues relating to the network to which the mobile node is already connected and to the one that it is going to handoff. For example, the decision to perform mobile-controlled handoffs may be made by a vertical handoff agent, sitting in the mobile device based on policies such as network bandwidth, load, coverage, cost, security, QoS, or even user preferences. User preference is important when performing vertical handoffs. For instance, if the new network to which a mobile device performs a handoff does not offer security; the user may still decide to use the old network. Depending upon coverage, a user may wish to use a secure and expensive link for his official email traffic (example using GPRS) but may still opt for a cheaper link to access web information (example WLAN) [4] (Table 3).

Traditionally, handoff decision has been performed based on the channel quality indicated by the received signal strength and other measurements, and the availability of resources in the new cell. This is done periodically so that degradations in signal strength below a prescribed threshold can be detected and handoff to another radio channel or cell can be initiated. However, the use of the signal strength as the only criteria to make a handoff decision limits the ability of the terminal to initiate a handoff for other reasons such as security requirements, etc. Also, traditional handoff does not allow user selection of networks, and assumes that there is only one choice for access technology. In a heterogeneous environment, user choice is a desirable amenity. Traditional handoff link transfer only concerns the delivery of packets to the new point of attachment. On the contrary, handoff in a heterogeneous environment demands the delivery of the context of the information flow between the mobile node and the network as well. Context may include security associations, QoS guarantees, authentication headers, and so on. Traditional handoff protocols are developed for homogeneous systems that rely on a common signaling protocol, routing technique, and mobility management standard. In heterogeneous environments, mobile nodes and network routers must be able to interoperate with different networks, and corresponding protocols and standards.

4.3. Handoff metrics in heterogeneous networks

Handoff metrics are used to indicate whether or not a handoff is needed. In traditional handoffs, only signal strength and channel availability are considered. In next generation heterogeneous wireless environment, new handoff metrics will have to be considered in connection with signal strength. Some of these metrics [7] include:

4.3.1. Cost

Cost is a major consideration to users since different network operators may employ different billing strategies. These variations in billing plans may affect the user's choice of handoff.

4.3.2. Network conditions

Network-related parameters such as traffic, available bandwidth, network latency, and congestion (packet loss) may need to be considered for effective network usage. Use of network information in handoff decisions can also be useful for load balancing across different networks, possibly relieving congestion in certain systems.

4.3.3. Battery power

Battery power may be a significant factor for handoff in some cases. For example, when the battery level is low, the user may choose to switch to a network with lower power requirements, such as an ad hoc Bluetooth network.

4.3.4. Application types

Different types of applications require different levels of reliability, latency, and data rate. The user applications running on a mobile device may also influence the handoff decision.

4.3.5. Mobile node conditions

Mobile node conditions include dynamic factors such as velocity, moving pattern, and location information.

4.3.6. User preferences

User preferences can be used to cater special requests for one type of system over another.

Next generation handoff will have to incorporate many of the above criteria rather than only signal strength in order to perform a handoff decision. Developing a handoff decision

Table 3
Comparison between traditional and next generation handoff strategies

	Traditional handoff	Next generation handoff
Handoff metrics	Signal strength	Bandwidth, cost, user-preference, network conditions, security, etc.
Radio link transfer requirements	Delivery of packets to the new point of attachment	Delivery of packets as well as contextual information (such as security associations, QoS guarantees, authentication headers, etc.)
Protocols	Relies on common signaling protocols, routing techniques, and mobility management standards	Requires mobile nodes and network routers to interoperate between different networks with various protocols and standards
Network types	Handoff occurs between homogeneous cells/networks (e.g. a handoff between two WLANs)	Handoff occurs between heterogeneous cells/ networks (e.g. handoff between a WLAN and a GPRS network)
Terminal types	Terminal equipped with a single access technology interface suffices	Multimode terminal (equipped with multiple access interfaces required)

function that takes into account numerous factors remains a research challenge.

4.4. Handoff implementation

Handoff implementation requires the actual transfer of data packets to a new wireless link in order to reroute a mobile user's connection path to the new point of attachment. It requires the network to transfer routing information about the mobile user to the new (or target) access router for the proper forwarding of packets. Since, next generation heterogeneous networks will operate in an environment of multiple standards and networks, transfer of packets to a new wireless link will also involve transfer of additional contextual information in order to enable the mobile node to move through different networks, while maintaining its data flows. The desired goal of transferring the context of a mobile node to the new network is to minimize the delay in re-establishing the mobile node's traffic flows. However, if the context transfer delay is so large as to have the same effect of the complete re-establishment, or large enough to increase the overall handoff call dropping rate, the advantages of context transfer are lost. A mechanism to allow for inter-network and/or inter-service-provider agreements to support fast intersystem handoffs while avoiding an unreasonable amount of inter-network signaling exchanges to validate or institute the adjustment in services is presently a crucial research problem [7].

Some of the motivations behind performing context transfer between nodes in an IP access network are:

1. The success of a time sensitive service such as VoIP, video, etc. in a mobile environment depends heavily on the minimization of the impact of the traffic redirection in order to deliver the host's IP traffic to the new point of access.
2. During the establishment of the new forwarding path, nodes along the new path must be prepared to provide similar forwarding treatment to IP packets.
3. In order to replicate the context of one forwarding node to another forwarding node, an open, standard solution for context transfer must be developed.

Context transfer is further discussed in detail in the following section.

4.4.1. Context transfer for seamless handoffs in heterogeneous wireless networks

Handoff may result in a change in the access router to which the mobile node is connected to. In many cases (when the mobile node moves across the access networks), the change in the communication path may also include a change in the communications network. For example, a host may move from a WLAN to a cellular network resulting in a change in the type of the wireless link.

Whenever the routing path of a mobile node changes due to handoff, the data flows (IP packets) of the host need to be re-routed by the access network to the mobile node's new point of attachment. This change of the routing path and re-establishment of routing information during host mobility is handled by *mobility protocols* (example Mobile IPv4 [35] and IPv6 [36]). The access networks may also need to establish and keep service state information (referred to as service *context* [37]) necessary to process and forward packets in a way that suits specific service requirements. A context is initially established in a network through protocol exchanges with the mobile node. However, when mobile nodes change their routing path, we need to ensure that nodes along the new path continue to provide similar treatment to IP packets as was provided along the old routing path. To provide seamless mobility, both the IP level connectivity and relevant context information needs to be quickly re-established.

If the mobile node is required to re-establish these service contexts using the same signaling process it used to initially establish them, delay-sensitive real time traffic (VoIP, video, etc.) as well as TCP-based applications may be seriously impacted. This impact is due to the large delay that would be introduced because of protocol exchanges with the mobile node. Thus, it is not possible to re-establish IP connectivity and service context information within the stringent time-constrain imposed by time-sensitive applications [39].

Context *transfer* [38] is a solution proposed by the Seamoby Working Group (WG) to reduce the handoff time by transferring the information related to the mobile node from the current access router to the next access router over the wired network, avoiding using the limited wireless bandwidth resources. An alternative to re-establishing service contexts when the mobile node moves to a new subnet is to transfer the existing service contexts from the host's Current Access Router to the New Access Router (NAR) of the destination subnet. Such services whose contexts may be transferred from one router to another are referred to as *context transfer candidate services* [37]. Examples of such services include Quality of Service (QoS), Authentication, Authorization and Accounting (AAA), and header compression state established and maintained between the mobile node and the Access Router (AR) to reduce the large IP header overhead of short (example VoIP) packets over bandwidth-limited wireless links.

Earlier works on maintaining context associations while moving across different wireless networks were mainly focused on setting up protocol state *after* handoff by signaling new state information over the wireless link. With context transfers, information is transferred from the old AR to the new AR through wired links. This approach results in faster and more secure transfer than signaling over the slow and error-prone wireless links. With context transfer, it would be possible to keep practically all handoff-related signaling within wired links of the access network [40].

To achieve seamless mobility across various networks and access technologies, the mobile node needs to have the information about access routers to which the mobile node could make a handoff to. These access routers are called Candidate Access Routers (CARs). The information that the mobile node needs to have in order to select a CAR and perform a handoff includes the identities (IP addresses) of CARs as well as their capabilities for supporting newly routed

data flows. This procedure of finding the access routers and their capabilities for the purpose of making a handoff is called Candidate Access Router Discovery (CARD) [41]. The Seamoby WG is involved in the development of a protocol that provides fast mobility within an access network. This is achieved by transferring the context of all IP-flows from one access router to another during handoff and eliminating the need to repeat the initial signaling process when the mobile node moves across IP subnets.

4.4.2. Candidate access router discovery and context transfer protocol

Handoffs in IP-based networks involve changes of the access point at the link layer and routing changes at the IP layer. For example a mobile node can initiate the handoff as soon as it detects a layer 2 (L2) id of a new access point during the link layer scan. A reverse address translation mechanism is needed to retrieve the IP address of the candidate access router based on its L2 id. Support for address translation is provided by the CARD protocol. The mobile node communicates to its current access router the L2 ids of the CARs; the AR performs the translation and sends the required information back to the mobile node. After a CAR has been found, the handoff process must be fast in order to support time-sensitive services.

To avoid delays in re-initializing all service states on the new access router, previously stored protocol state information can be transferred to the new subnet in advance through the wired network in order to quickly re-establish the service. *Context* is the information required for re-establishing the service at the new AR. *Context transfer* is the movement of the context from one router to another as a means of re-establishing context transfer candidate services on a new subnet without having to perform entire protocol exchanges with the mobile node from the beginning. In order for context transfer to be successful, the new AR must be capable of supporting the rerouted flow (must support the same context-transfer candidate services as the sending router); otherwise, even a timely transfer of flow state can lead to service disruption due to a lack of support at the NAR. The mobile node should have the means to discover the capabilities of the AR to which it is going to attach after handoff to be sure to receive the desired support. Such means are provided by the CARD protocol. In addition to mapping link layer IDs to IP addresses, the protocol can optionally be used to request from the current AR the capabilities of the CARs to eventually decide, whether it is worthwhile to perform a handoff to that AR. Capabilities may include flow related parameters, such as, QoS support or security and authorization information.

Context transfer is most useful when the context can be transferred in advance so that the new AR can establish quickly the service and make it available to the mobile node once it has performed the handoff. To initiate proactive context transfer, the context transfer solution must define the events that trigger the transfer of the context to the new subnet, so that all the involved network entities can receive and correctly process the information necessary to guarantee seamless handoff. An important feature in the common framework is the timeliness of the messages exchanged. The applications that will make use of CARD and context transfer can have requirements on the minimum interpacket delay and packet losses. Timeliness can be achieved by properly optimizing the message exchange, avoiding unnecessary messages, especially on the wireless link, and minimizing the number of entities involved in signaling. However, there are limits to the applicability of CARD and context transfer. Although CARD is especially useful for those network configurations where several ARs are close to each other, attention must be paid to avoid overwhelming the wireless link with messages aiming to perform address translation [40].

Context transfer is not always possible, even though the CAR can support the transferred contexts. An example could be when the mobile node crosses the border between two administrative domains: in such a case, the new network provider might require the mobile node to be re-authenticated from the beginning rather than to accept its transferred contexts. Security is a major concern in this environment; it is especially important that the entities exchanging context or router identity have authenticated each other, and the communication is encrypted. In the following sections, we discuss the main features and operation of the Candidate Access Router Discovery and Context Transfer protocols.

4.4.3. CARD protocol

CARD [35] and Context Transfer Protocol (CTP) [42] were designed by the Internet Engineering Task Force (IETF) Seamoby WG to enable context information to be seamlessly exchanged during handoffs. The CARD protocol allows mobile nodes to discover information about the AR candidates for the handoff. The two main functionalities of the CARD protocol are support for reverse address translation from the link layer ID of the AR to the IP address and discovery of the capabilities of the candidate AR. The actual mapping between link layer and IP layer addresses is done at the current AR. The CARD protocol carries the needed information between the mobile node and AR. The ARs maintain a local CAR table where the link layer IDs of the CARs, with the corresponding IP addresses, are mapped. In the table each AR is associated with a set of capabilities, refreshed periodically. The CARD specification does not mandate any mechanism for populating the local CAR table; two dynamic approaches, one relaying on a centralized entity coordinating the IP address and another distributed one, are mentioned. The table is updated using the CARD Request and CARD Reply messages, which are exchanged between the mobile node and its current AR or between the current AR and the CARs. Capabilities are carried as options in the CARD messages and expressed as a list of attribute-value pairs. The mobile node can specify a set of capabilities in which it is interested. The current AR can perform filtering of information and can provide only the CARs matching the preferences with subsequent saving of bandwidth. Examples of capabilities are resource availability information, wireless link layer technologies supported, and existing security mechanisms. The request for capabilities is optional;

the only necessary request instantiated by the mobile node is address translation.

The CARD process is initiated by the mobile node upon receipt of a CARD trigger, such as the link layer ID of a wireless access point; the mobile node sends to its current AR the MN-AR CARD Request message, specifying one or more access point ID and a list of desired capabilities. The mobile node informs the current AR about the need for CAR capabilities by raising a flag in the header of the option field. If no link layer IDs are provided, the current AR will reply with the IP addresses of the CARs currently present in its local table. If, for example, the mobile node requests a list of the capabilities supported by the CAR by specifying an access point's link layer ID in the Request message, then the AR resolves the IP address of the CAR based on the L2 ID received and sends to the specified CAR an AR-AR CARD Request indicating all the CAR capabilities whose timers have expired in the current AR local table. The CAR responds with the AR-AR CARD Reply reporting the requested capabilities. The current AR updates the capabilities entries in its table and sends back to the mobile node the IP address of the CAR, along with the requested capabilities with the MN-AR CARD Reply message.

4.4.4. Context transfer protocol

A context transfer protocol has been developed by the IETF Seamoby working group [25]. According to this protocol, as the mobile node moves from its previous access router to the new access router, the corresponding information about each of the mobile node's data-flows is forwarded between the access routers. Each data-flow is categorized into feature contexts, which allow the network to indicate and provide the particular context information needed per data-flow. For example, a particular mobile user may be downloading streaming video while conducting a voice transmission. The context required to continue the voice call may be authentication information only, while the context needed for the video service may be QoS and header compression, in addition to authentication.

CTP defines two ways to initiate context transfer operations: *proactive* and *reactive*. In the former case, the current AR initiates the context transfer before the actual handoff after receiving a Context Transfer Activate Request (CTAR) message from the mobile node or due to a network decision. The current AR sends a Context Transfer Data (CTD) message to the new AR to start the context transfer. In the latter case, the new AR receives the CTAR message from a new mobile node that includes the IP address of the old AR and information about the contexts to be transferred. The new AR sends a Context Transfer Request (CTR) message to the old AR, which replies with a CTD message that starts the transfer of context information.

In the first case, context transferred before actual handoff execution, implies proactive transfer of context information, and it is clearly more desirable than reactive transfer. However, proactive context transfer is not always possible. First, the current AR must know in advance that a handoff is going to be performed so that it can transfer context information a priori to the new AR. Scanning for other access points from the mobile node or listening for potential mobile nodes from access points may not be possible with all wireless links. Moreover, the handoff may be forced suddenly due to access point failure, for example. Second, context information that is updated at high rates cannot be sent proactively, as it would be obsolete at the new AR after mobile node handoff. When proactive context transfer cannot be used, having reactive context transfer, as in the second case, is still better than having no context transfer at all, because services need not be re-established from the beginning.

For proactive context transfers initiated by the mobile node, the CTAR message contains the IP addresses of the new AR and the mobile node on the current AR, the list of the contexts to be transferred, and a token that authorizes the transfer. The CTD message contains the transferred contexts and the old IP address of the mobile node, as well as parameters for the new AR to verify the authorization token. The CTAR message is always sent by the MN to the new AR, even if the proactive context transfer is initiated by the current AR based on a network decision. This allows the new AR to have means to compare the authorization token received from the mobile node with the CTAR message with the parameters received from the current AR with the CTD message. Optionally, the new AR can acknowledge the received context with a CTD Reply message (CTDR), reporting the outcome of the context transfer process.

For the reactive context transfer case, the mobile node solicits the new AR to start the context transfer with the CTAR message, providing the parameters described above. The new AR requests context transfer from the old AR with the CTR message, in which it provides the mobile node's current IP address, the contexts to be transferred, and the authorizing token generated by the mobile. The context is transferred with the CTD message and acknowledged with the CTDR message. The signaling diagrams show that the context transfer protocol creates only a few messages to be exchanged. Moreover, most of the signaling exchange happens on the wired links between ARs, and only one or two messages are sent through the wireless links. The size of messages can be kept low, optimizing the use of bandwidth in the wired link too. In the CTAR message, in fact, the mobile node can specify the context types the current AR must transfer, the default being all, leading to smaller message size. Context transfer can also be network initiated; in the proactive case, the current AR after a proper trigger (not specified by the CTP specification) can begin the transfer, sending a CTD to the new AR. In the reactive case, the new AR, again after a trigger, sends a CTR request to the current AR.

5. Terminal requirements for future heterogeneous wireless networks

As future wireless systems are expected to support several heterogeneous access networks, we need to develop terminals that that have the capability to access these diverse technologies. In this context, terminals and devices capable of supporting different types of access technologies are being

designed. Current multimode devices are of two different types:

- The first type includes those devices capable of supporting multiple access systems by incorporating several network interface cards and the appropriate software for switching between these network interfaces. These devices are often referred to as 'multimode terminals'.
- The second type includes those devices which use adaptable software modules that operate on a generic, reconfigurable hardware platform consisting of Digital Signal Processors (DSPs) and general purpose microprocessors, used to implement radio functions such as generation of transmitted signal (modulation) at transmitter and tuning/detection of received radio signal (demodulation) at receiver. Such devices are referred to as Software-defined Radios (SDRs).

5.1. Multimode wireless terminals

Multimode wireless terminals [10] are devices supporting multiple radio access technologies and allow reception of data over multiple system bearers with different characteristics. An intelligent multimode terminal should be able to decide *autonomously* the active interface that is best for an application session and to select the appropriate radio interface as the user moves in and out of the vicinity of a particular technology. The decision regarding the switching of the interface and the handoff of the active sessions to the newly active interface may be decided based on user preference setup (may also be referred to as 'user policy') on the multimode terminal regarding the interface usage. For example, if the user has setup some preference to indicate that the PDA should default to using the fastest and cheapest interface available, the terminal will switch to the WLAN interface whenever the PDA enters into the WLAN coverage area. However, the multimode terminal must also take into account other factors such as the QoS requirements of the running applications, etc.

5.1.1. Requirements for multimode terminal operation

Some of the requirements that need to be fulfilled in order to build intelligent multimode devices include:

1. The terminal should operate with minimal inputs from the user. From a user experience, it is preferable to carry out these decisions in an automated way rather than having to query the user every time a new interface becomes available, or an old interface disappears.
2. The terminal should be able to handle session handoffs from one interface to the other based on user policy.
3. Radio access interfaces should be selected based on user preferences, application QoS requirements, and, information about the network.
4. The requirements of applications should be determined and then decide whether an application could benefit from changing interfaces.
5. Traffic should be smoothly transferred while changing the active interface in a way that is transparent to the user, that is, as seamlessly as possible.

5.1.2. Deployment issues for multimode terminals

5.1.2.1. Network detection. The multimode terminal must be able to detect the availability of a new network (example WLAN connectivity). For some interface technologies, resources are not exhausted even if the interface is permanently on and scanning for network access coverage. In the case of IEEE 802.11 WLANs, this approach is infeasible due to the amount of power needed to keep the interface actively scanning for access points. So, the issue that arises is how to activate the WLAN interface when in the vicinity of a hotspot. Currently, this activation of the interface is left to the user who manually enables their WLAN card when in a hotspot. However, since one of the multimode terminal requirements is to minimize user interaction, a desirable goal in this case, would be however, to automate this process. One solution that would achieve this result would be to advertise the presence of nearby services (or coverage of other technologies) via the currently active interface. For example, the presence of a hotspot could be sent via GSM, which would then activate the WLAN interface automatically.

5.1.2.2. Network information. Finding out information about the network to support interface selection decisions. For example, if the user specified that the mobile terminal should default to using the cheapest connection, then for the terminal to be able to decide which connection is the cheapest, some charging information must be made available by the network. Service providers may choose to advertise certain pricing information allowing the prices of the different networks to be compared. This information also needs to be formatted or standardized between operators so that it is easy to compare. Other information such as authentication methods which allow the terminal to determine in advance if switching access networks is required may also be advertised. We need to investigate how to make this information available to mobile users.

5.2. Software-defined radios

Software-Defined Radio (SDR) [20–22,26] is a technology that uses adaptable software and flexible hardware platforms to address the problems that arise from the constant evolution and technical innovation in the wireless industry, particularly as waveforms, modulation techniques, protocols, services and standards change [16]. It is receiving enormous recognition and generating widespread interest among wireless users and developers as it promises to solve these problems by implementing the radio functionality as software modules running on a generic hardware platform. SDR technology aims to take advantage of re-programmable hardware modules to build an open-architecture based on radio system software. SDR facilitates the software implementation of some of the

functional modules in a radio system such as modulation/demodulation, signal generation, coding, and link-layer protocols. This helps in building reconfigurable software radio systems where dynamic selection of parameters for each of the above-mentioned functional modules is possible. A complete hardware based radio system has limited utility since parameters for each of the functional modules are fixed. A radio system built using SDR technology extends the utility of the system for a wide range of applications that use different link-layer protocols and modulation/demodulation techniques. The SDR system can be reconfigured depending on the software module being used. Also, the software modules that implement new services/features can be downloaded over-the-air onto the handsets. This kind of flexibility offered by SDR systems helps in dealing with problems due to differing standards and issues related to deployment of new services/-features. In a nutshell, Software-Defined Radio (SDR) refers to the technology wherein software modules running on a generic hardware platform consisting of DSPs and general purpose microprocessors are used to implement radio functions such as generation of transmitted signal (modulation) at transmitter and tuning/detection of received radio signal (demodulation) at receiver. SDR technology can be used to implement military, commercial and civilian radio applications. A wide range of radio applications likes Bluetooth, WLAN, GPS, Radar, WCDMA, GPRS, etc. can be implemented using SDR technology.

5.2.1. Motivation for SDR development

As mobile communications networks evolve to accommodate accelerating demand for wireless voice and data services, equipment manufacturers and network operators are being forced to confront many difficult, and simultaneous, challenges. Those challenges include: the support for an explosive subscriber growth using multiple standards, modes and frequencies, the proliferation of equipment and service platforms, meeting the demand for wireless Internet and information services; supporting the higher cost and longer lifetime of third-generation (3G) wireless terminals; adapting to the increasing competition between service providers, and overcoming the cost and scarcity of the wireless spectrum.

To address the above challenges, SDR offers the wireless industry an alternative vision where every consumer has a single, personalized terminal that works everywhere and to which new services are instantly downloaded. SDR has generated tremendous interest in the wireless communication industry for the wide-ranging economic and deployment benefits it offers. Some of the factors that are fueling such a strong interest in SDRs include:

- Commercial wireless network standards are continuously evolving from 2 to 2.5G/3G and then further onto 4G. Each generation of networks differ significantly in link-layer protocol standards causing problems to subscribers, wireless network operators and equipment vendors. Subscribers are forced to buy new handsets whenever a new generation of network standards is deployed. Wireless network operators face problems during migration of the network from one generation to next due to presence of large number of subscribers using legacy handsets that may be incompatible with a newer generation network. The network operators also need to incur high equipment costs when migrating from one generation to next. Equipment vendors face problems in rolling out newer generation equipment due to the short time-to-market requirements.
- The air interface and link-layer protocols differ across various geographies (for example, European wireless networks are predominantly GSM/TDMA based while in USA the wireless networks are predominantly IS94/CDMA based). This problem has inhibited the deployment of global roaming facilities causing great inconvenience to subscribers who travel frequently from one continent to another. Handset vendors face problems in building viable multi-mode handsets due to high cost and bulky nature of such handsets.
- Wireless network operators face deployment issues while rolling-out new services/features to realize new revenue-streams since this may require large-scale customizations on subscribers' handsets. SDR technology enables implementation of radio functions in networking infrastructure equipment and subscriber terminals as software modules running on a generic hardware platform. This significantly eases migration of networks from one generation to another since the migration would involve only a software upgrade. Furthermore, since the radio functions are implemented as software modules, multiple software modules that implement different standards can co-exist in the equipment and handsets. An appropriate software module can be chosen to run (either explicitly by the user or implicitly by the network) depending on the network requirements. This helps in building multi-mode handsets and equipment resulting in ubiquitous connectivity irrespective of the underlying network technology used. SDR technology supports over-the-air upload of software modules to subscriber handsets. This helps both network operators as well as handset manufacturers. Network operators can perform mass customizations on subscriber's handsets by simply uploading appropriate software modules resulting in faster deployment of new services. Manufacturers can perform remote diagnostics and provide defect fixes by just uploading a newer version of the software module to consumers' handsets as well as network infrastructure equipment. However, SDR technology has some drawbacks such as higher power consumption, higher processing power requirement, and higher initial costs. SDR technology may not be suitable for all kinds of radio equipment due to these factors. Hence, these factors should be carefully considered before using SDR technology in place of a complete hardware solution. For example, SDR technology may not be appropriate in pagers while it may offer great benefits when used to implement base-stations [18,19].

5.2.2. Key features of SDR technology

5.2.2.1. Reconfigurability. SDR allows co-existence of multiple software modules implementing different standards on the same system allowing dynamic configuration of the system by selecting the appropriate software module to run. This dynamic configuration is possible both in handsets as well as infrastructure equipment. The wireless network infrastructure can reconfigure itself to the subscriber's handset type or the subscriber's handset can reconfigure itself to network type. SDR technology facilitates the implementation of future-proof, multi-service, multi-mode, multi-band, multi-standard terminals and infrastructure equipment. Reconfiguration also helps in debugging errors on a mobile terminal or a base station by downloading appropriate software to reconfigure the hardware. Also, hardware equipment needs to be replaced when new functions are added to a particular device. SDR's ability to reconfigure solves this problem.

5.2.2.2. Multimode operation. SDR provides multimode operation [17], which is essential for next generation wireless systems. It helps in saving infrastructure costs. For example, if one system is suitable for covering outdoor area and another system is suitable for indoors; coverage can be expanded without additional investment in either system if the systems cooperate and dual-mode terminals become available.

5.2.2.3. Ubiquitous connectivity. SDR enables the implementation of air interface standards as software modules and multiple instances of such modules that implement different standards can co-exist in infrastructure equipment and handsets. Such benefits enable the global roaming facility. If a terminal is incompatible with the network technology in a particular region, an appropriate software module needs to be installed onto the handset (possibly over-the-air) resulting in seamless network access across various geographies. Further, if the handset used by the subscriber is a legacy handset, the infrastructure equipment can use a software module implementing the older standard to communicate with the handset.

5.2.2.4. Interoperability. SDR facilitates implementation of open architecture radio systems. End-users can seamlessly use innovative third-party applications on their handsets as in a personal computer system. This enhances the appeal and utility of such handsets.

6. Recent vertical handoff management techniques proposed for heterogeneous wireless networks

Several techniques have been proposed for performing handoffs while roaming across heterogeneous wireless access networks. These approaches operate at different layers of the network protocol stack. Most of these approaches are based on a modification to the Mobile IP protocol and are implemented at the network layer [7,9,11,23,28]. Other handoff strategies include SIP-based handoff (operating at the application layer) [29], and SCTP-based handoff (operating at the transport layer) [27]. These are also some recent schemes for handoff management proposed to achieve seamless mobility for TCP connections [30,33]. Other types of handoff schemes are based on efficient energy consumption of network interfaces [5,24]. In the following section, we briefly describe these approaches as well as their benefits and limitations (Table 4).

6.1. Mobile IP

Mobile IP [7,11] is a mobility management protocol proposed to solve the problem of node mobility by redirecting packets to the mobile node's current location. The Mobile IP architecture is shown in Fig. 3. Its main components include a Home Agent (HA) and a Foreign Agent (FA). HA is a router on a mobile node's home network, which encapsulates datagrams for delivery to the mobile node when it is away from home, and maintains current location information for the mobile node. FA is a router on a mobile node's visited network (foreign network) that provides routing services to the mobile node when registered. The FA decapsulates and delivers datagrams, tunneled by the mobile node's home agent, to the mobile node. When a mobile node moves out of its home network it must obtain another IP address because according to the traditional IP protocol a node's address is fixed by geographical location. So, in mobile IP, a mobile host uses two IP addresses: a fixed home address (a permanent IP address assigned to the host's network) and a *care-of-address*—a temporary address from the

Table 4
Emerging technologies for seamless mobility across heterogeneous wireless access networks

Technology	Description
Internet protocol (IP)	Provides a backbone and a scalable solution to interconnect different networks
Mobile IP	Enables mobile devices to maintain their IP addresses while moving between different networks
Mobile IPv6	Provides enhanced features to Mobile IP for optimizing the mobility process
Mobile stream control transmission protocol (mSCTP)	Transport layer solution that provides seamless roaming capability across different networks
Session initiation protocol (SIP)	Application layer solution that provides both, personal and terminal mobility across different networks
4G networks	An all-IP architecture that integrates several different networks with a plethora of services and capabilities
Intelligent software	Enables the creation of smart terminals that can make network selection decisions with minimal user intervention
Software-defined radio (SDR)	Elements of a wireless network whose operational modes and parameters can be changed via specialized software inside the DSPs

- Discovering the care-of address
- Registering the care-of address
- Tunneling to the care-of address

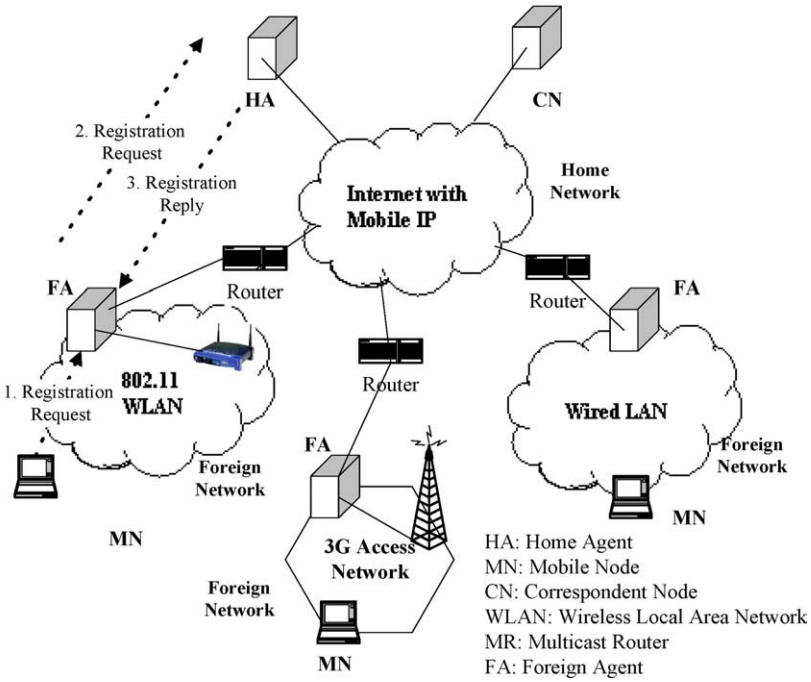

Fig. 3. Mobile IP Network Architecture.

new network (i.e. foreign network) that changes at each new point of attachment. When the mobile node moves, it has to first discover its new care-of-address. The care-of-address can be obtained by periodic advertising from the foreign agent through broadcasting. The mobile node then registers its care-of-address with its home agent by sending a registration request to its home agent via the foreign agent. The home agent then sends a registration reply either granting or denying the request. If the registration process is successful, any packets destined for the MN are intercepted by the HA, which encapsulates the packets and tunnels them to the FA where decapsulation takes place and the packets are then forwarded to the appropriate MN. Tunneling involves doing IP-within-IP or IP encapsulation. The home agent encapsulates the packets into its IP packet (IP inside IP). The inner packet is the original, untouched packet; the outer packet is addressed to the care-of address and sent through the usual Internet routing mechanisms. Once the foreign agent receives packets, they are stripped of their outer layer and the original packet is extracted. Thus, the Mobile IP process can be summarized in three steps:

- Discovering the care-of address
- Registering the care-of address
- Tunneling to the care-of address

There are certain routing inefficiencies in Mobile IP

1. Packets sent from a CN to an MN are first intercepted by the HA and then tunneled to the MN. However, packets from the MN are sent directly to the CN. This triangular routing problem results in communication routes significantly longer than the optimum routes and introduces additional delay for packet delivery [12].
2. When the MN moves form one subnet to another, the new FA cannot inform the old FA about the movement of the MN. Hence, packets already tunneled to the old care-of-address are lost.
3. Mobile IP is not a satisfactory solution for highly mobile users. Mobile IP requires a MN to send a location update to it's HA whenever it moves from one subnet to another. This location registration is required even though the MN does not communicate with others while moving. The signaling cost associated with location updates may become very significant as the number of MNs increase. Moreover, if the distance between the visited network and the home network is large, the signaling delay is long.

6.2. Mobile IPv6

Improvements have been proposed and adopted for Mobile IP under the title of Mobile IP version 6. Mobile IPv6 eliminates triangular routing and enables the correspondent node to reroute packets on a direct path to the mobile node. This process is referred to as *route optimization.* In addition, Mobile IPv4 also suffers from a lack of security constructs for authorization, authentication, and accounting, as well as for source routing [11]. Mobile IPv6 includes embedded binding updates and care-of address configuration for the execution of

location updates and processing the change in the mobile node's address. The newer version also includes authentication header processing to provide validation of mobile nodes. Finally, IPv6 has a fourfold increase in IP address space, which may be useful for developing new mobile node addressing schemes. Regardless of the type of network, issues such as addressing, route optimization, and authentication are part of the challenging overall problem of creating an efficient handoff mechanism that satisfies the seamless mobility needs of the user population while enabling advanced processing and optimization operations at the network. An improvement over Mobile Ipv6 is *Hierarchical Mobile Ipv6* [13] which is proposed to minimize the amount of signaling between the correspondent node and the HA.

6.3. Network layer handoff approaches

6.3.1. HOPOVER

HandOff Protocol for OVERlay networks (HOPOVER) [9] is a Mobile-IP based approach that handles both macro-level (vertical handoffs) as well as micro-level (horizontal handoffs) mobility. It is designed to address the problems of high handoff frequency due to fast moving mobile nodes, which cannot be handled by the traditional Mobile-IP protocol due to the high signaling overhead that is incurred. HOPOVER is scalable for a large number of mobile devices and expanding networks. It also incurs low signaling overhead. Fig. 4 shows the HOPOVER architecture.

The Handoff procedure in the HOPOVER scheme consists of three steps:

6.3.1.1. Handoff Preparation. When a Mobile Node (MN) decides it may encounter a handoff (based on the comparison of base station beacons it receives), it starts with the following handoff-preparation processes.

- The MN chooses a small group of neighboring Base Stations (BSs) and sends them a Handoff-Prepare (HP) packet. Each BS forwards the packet to its gateway router. The BS selection is based on signal strength, bandwidth, pricing and other factors, assuming such information is available from BS beacons. When such information is unavailable, the MN simply makes a random selection.
- The authentication server of the new network verifies the validity of the authentication information included in the HP packet. If it is invalid, a HP_NACK (negative Ack of the HP) is sent to the MN, and no further handoff preparation work is performed. If it is valid, the routing state along the path from the gateway router to the chosen BSs is set.
- Based on the resource reservation information included in the HP packet, resources are reserved in target cells and along the path from the MN's current sender to the target cell. Each of the chosen BS allocates a buffer for the MN and prepares to buffer packets that are in transit to the MN.
- Each of the new BSs sends the old BS a HP_ACK packet. Upon receiving such a packet, the old BS adds the corresponding new BS to the *forward list* for that MN and begins forwarding packets to the new BS. On completing these signaling procedures, routing information along the path from the new gateway routers to the new BSs is set up and packets in transit have been buffered for the MN.

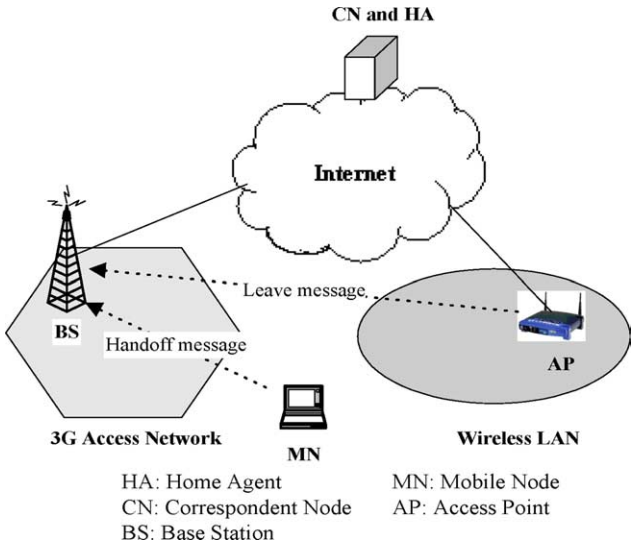

Fig. 4. HOPOVER Architecture.

6.3.1.2. Handoff. When the MN decides, it is actually moving to another cell, the actual handoff process is performed using the following procedures.

- The MN sends a Handoff message to the BS of the target cell. The new BS then begins forwarding packets to the MN including the buffered ones.
- The new BS sends a Leave message to the old BS and all the 'handoff preparing' BSs.
- The old BS records the MN's current network so that it can forward packets there. It stops forwarding packets to other BSs, which are on the forwarding list.
- The old BS removes the MN from its 'current MN' list.
- Other handoff-preparing BSs remove the related routing information, allocated buffers and buffered packets.
- The old BS and other handoff-preparing BSs delete themselves from the resource reservation tree they joined for that MN by sending a Release message. Thus, the network releases those reserved resources.

6.3.1.3. Updating mobile IP information. After handoff, the previous BS maintains the forwarding address for the MN. To avoid wasted resources and additional delay, sometimes, the MN's Mobile IP FA information should be modified to reflect the MN's current address. Timers are used to make sure that the Home Agent is contacted only if the MN stays in the new cell for long time.

After a handoff, both the old and new BSs set up a timer for the MN. If after the timer expires, the MN is still in the new network then the new BS sends a Mobile IP update packet to the MN's HA, so that the new BS becomes the new FA for the

MN. Also, the new BS notifies the old BS to remove the forwarding information for that MN.

If before the timer expires, the MN moves back to the previous network, then in that case, no contact with the Mobile IP HA is performed. Both timers are removed.

If before the timer expires, the MN moves to a third network. The second BS sends Mobile IP update packet to the MN's HA to make itself the new FA of the MN. At the same time, the second BS begins monitoring the MN's stay in the third network.

6.3.1.4. Limitations of HOPOVER. The HOPOVER handoff process requires base stations to maintain excessive information about mobile nodes. This information includes a list of BSs to forward packets to, buffering queues for packets in transit, routing and resource reservation parameters, as well as information regarding current network characteristics. The handoff process using the HOPOVER scheme also calls for complete standardization among different network service providers so that the various network entities can exchange signaling messages among themselves. This requires major modifications to the existing infrastructure.

6.3.2. Hierarchical approach

The *Hierarchical approach* [23] is a Mobile-IP based scheme proposed to improve vertical handoffs in heterogeneous systems. It defines five network entities: Home Agent (HA), Foreign Agent (FA), Gateway Foreign Agent Router (GWFA), a Base Station (BS), Multicast Router (MR) and a Mobile Node (MN). The HA is a router in the mobile node's home network. The GWFA is a router implementing the role of a FA and a multicast router in order to manage the visitor mobile nodes. It can also implement the role of an HA if its domain is the home network. The GWFA router periodically broadcasts agent advertisement messages containing its IP address and assigns a multicast address unique within its domain to the MN. The GWFA router manages macro and micro cells and uses an overlay network. The BS is a network-layer router having two wireless interfaces: wireless and wired. MN accesses the Internet via the BSs over the wireless links. The BSs can buffer the last few IP packets sent to the MN. Fig. 5 shows the network architecture of the Hierarchical approach.

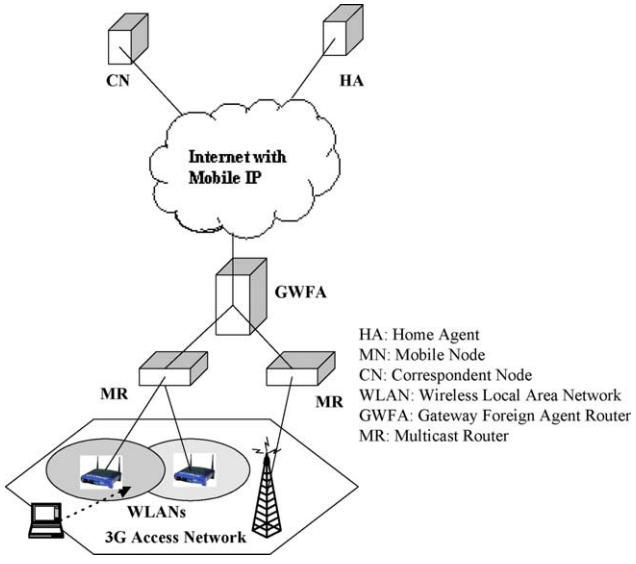

Fig. 5. Architecture of Hierarchical Handoff Scheme.

When the MN is away from its home network, the HA intercepts the packets addressed to it and sends these packets through a tunnel to the current mobile node's attachment. The MN would previously have registered its GW router care-of-address with it's HA. The GWFA router decapsulates packets sent by the HA and forwards those packets to the MN as a multicast stream using the multicast address assigned to this MN. A small group of BSs are selected by the MN to listen to its multicast address for packets encapsulated and sent by the GWFA router. One of the BSs is selected by the MN to be the forwarding BS; it decapsulates the packets sent by the GWFA router and forwards those packets to the MN. The other BSs are buffering BSs; they hold a small number of packets from the GWFA router in a circular buffer. BSs send out periodic beacons similar to Mobile IP foreign agent advertisements. The MN listens to these packets and determines which BS should forward packets for the MN, which BSs should buffer packets in anticipation of a handoff, and which BSs should not belong to the multicast group at all. When the MN initiates a handoff, it instructs the old BS to move from forwarding to buffering mode, and the new BS to move from buffering to forwarding mode. The new forwarding BS forwards the buffered packets that the MN has not yet received.

6.3.2.1. Limitations of the hierarchical approach. The Hierarchical approach involves registration and packet forwarding mechanisms similar to the Mobile IP protocol and this incurs additional delay as well as signaling overhead since packets are intercepted by the HA and then forwarded to the mobile node.

6.3.3. OmniCon

The OmniCon approach [28] is a modification to the Mobile IP protocol and is designed to support vertical handoffs in heterogeneous networks. It provides support for packet scheduling and buffering mechanisms to accommodate different transmission characteristics of the networks involved. Fig. 6 shows the OmniCon architecture.

The OmniCon approach modifies the MN and the GPRS FA with two new modules:

(i) Decision Module
(ii) Communication Daemon

The Decision Module is a WLAN link availability monitoring system incorporated into the MN to constantly monitor the WLAN signal strength, quality and noise level. The MN and the FA are modified to support TCP tunneling. The MN and the GPRS FA are also enhanced with OmniCon Communication Daemons (CDs), which establish a TCP connection with each other over the GPRS link.

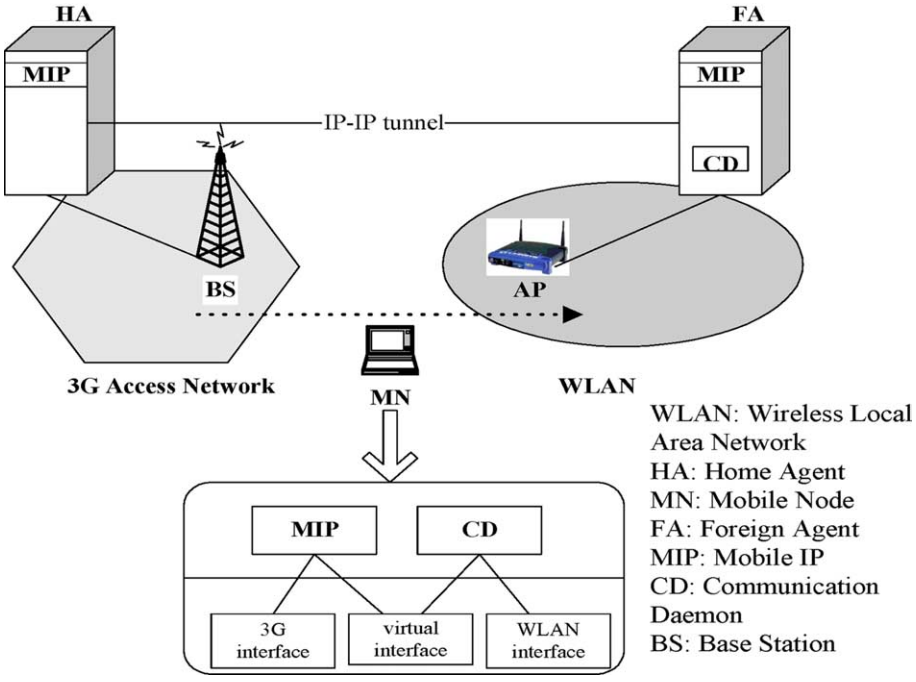

Fig. 6. OMNICON Network Architecture.

For vertical handoffs, OmniCon implements a virtual network interface called *tcptun*. This virtual interface communicates with the Mobile IP software and the CDs. During system initialization, the *tcptun* is assigned the same IP address as the WLAN network interface. The Mobile IP software listens on this interface for mobile agent advertisements. On reception of GPRS foreign agent advertisements on *tcptun*, the Mobile IP software registers with the FA using the same interface. Once the registration is successful, the routing table entries are updated to set the *tcptun* interface as the default interface for all outbound packets bearing the Mobile IP address as the source address. These outbound packets are read by the CD and are tunneled over the TCP connection to the GPRS FA, which takes care of further routing.

Whenever, the MN detects that the WLAN signal strength has fallen below a particular threshold, OMNICON triggers a handoff from the WLAN interface to the virtual interface. The handoff triggering is carried out by sending multiple foreign agent advertisements on behalf of the GPRS FA through the virtual interface, up to the TCP/IP stack of the MN. This enables Mobile-IP to carry out a network layer handoff and start using the virtual interface. Mobile IP switches back to the WLAN interface by holding off the foreign agent advertisements on the GPRS link. In order to avoid multiple handoffs between the two networks, the decision module uses two signal-strength threshold values, with the maximum value referred to as the high-watermark and the minimum value referred to as the low-watermark. A handoff from the WLAN to the GPRS is triggered only if the signal strength falls below the low-watermark. On the other hand, a handoff in the reverse direction (GPRS to WLAN) is triggered only when the signal strength improves above the high-watermark.

6.3.3.1. Limitations of the Omnicon Approach. The Decision Module in the OmniCon approach makes the handoff decision based only on the detection of the WLAN signal performed by the WLAN link availability monitoring system. It does not consider the requirements of the running applications or user preferences to switch between access networks.

6.4. Transport layer handoff approaches

Implementation of handoff techniques at the *transport layer* requires a means to detect and reconfigure mobile hosts as they move from one network type to another. This includes the detection of new networks and the allocation of new IP addresses. These tasks are often handled by Dynamic Host Configuration Protocol (DHCP) or Router/Neighbor Discovery methods. A handoff protocol at the transport layer is required to implement a method that can dynamically rebind a connection's IP address. Transport layer handoff techniques operate by exploiting services from network and data-link layers for network detection and IP address management. Hence, handoff techniques implemented at the transport layer are often referred to as cross-layer approaches [43]. Recent transport-level handoff management protocols that have been proposed include the Stream Control Transmission Protocol (SCTP) [45] and the Datagram Congestion Control Protocol (DCCP) [46].

SCTP provides a connection-oriented reliable service. A connection between two SCTP end-points is called an *association*. Multi-homing is a prominent feature of SCTP. Multi-homing allows an association to maintain multiple IP addresses. In the case of SCTP, an end-point of an SCTP association can be mapped to several IP addresses. Among those addresses, one address is used as the primary address for

current transmission and reception. Other addresses (secondary) can be used for retransmissions. Multi-homing feature of SCTP provides a basis for mobility support since it allows a mobile node (MN) to add a new IP address, while holding an old IP address already assigned to it. In addition to SCTP's multi-homing feature, an extension to SCTP called mSCTP (mobile SCTP) allows functions such as ADDIP, DELETEIP that enable dynamic addition/removal of IP addresses to an SCTP end-point [44]. SCTP's built-in support for multi-homed endpoints is especially useful for applications that require high network availability and helps to recover from link failures without interrupting the data transfer [49].

Another SCTP-based mechanism called cSCTP (cellular SCTP) [47] is used to implement soft handoffs and provides better performance compared to mSCTP. During handoff using cSCTP, the correspondent node (CN) sends packets to the primary address (old network) and to the primary address of the new network. This transfers the same data via two different paths, to reduce the probability that the MN would miss the data packets sent by the CN. In contrast, for mSCTP, before the MN sets the new IP address to the primary IP address of the association, data packets are sent to the old IP address. If the MN is not reachable by the old IP address anymore, then retransmissions will be sent to the new IP address. These retransmissions result in increased delays and degrade the performance of SCTP. cSCTP is designed to eliminate packet loss by using two IP addresses in parallel to duplicate packet transmission during handoff.

Transport layer provides services that handle congestion control. It is therefore crucial that the transport layer be aware of path changes of a mobile node due to mobility, so that the rate of data transfer may be adjusted. For instance, a mobile node's sending rate (before a movement between networks) may be too fast for the new network path, causing substantial packet loss if the transport layer does not reinitialize its congestion control state for the new network path. Similarly, in the case of TCP, a slow start threshold may prevent the transport layer from efficiently utilizing a new path of higher bandwidth. These issues indicate that for efficient operation of mobility schemes implemented at any layer of the network protocol stack, congestion-control transport protocols require at least some modifications if connections are to persist smoothly across attachment point changes.

There are several benefits associated with implementing handoffs at the transport layer [48]. Some of these benefits include:

1. Simplified network infrastructure. There is no requirement of a home network or additional infrastructure beyond DHCP and Domain Name Server (DNS), which are already well deployed as part of IP networks.
2. Inherent route optimization. With mobility anchored at the transport layer, there is implicit route optimization. Packets move directly from the end source to the end destination, with no indirection. There are no triangular routes.
3. Smooth handoffs. With transport layer mobility, transport protocols are explicitly aware of the changes in their network attachment status. This allows transport protocols to take proper action in order to smooth transitions (such as pausing transmission during the handoff and resetting congestion-control state) into new networks
4. Immune to spurious agents. Transport layer mobility architecture does not use HAs or FAs and is therefore immune to spurious agents.
5. Location privacy. With transport layer mobility, corresponding nodes always know the mobile node's current location. However, this information is hidden from the home network, as well as from the foreign network. No intermediate active agents are required to maintain the state information of any mobile node. DNS is used to handle location management and to provide access to a mobile node's current location.

However, one of the drawbacks of transport layer handoff approach is their dependence on other layers for location management. Another drawback is that if each transport protocol is to implement binding updates, then each of the protocols requires an authentication scheme to prevent spoofing. Ensuring the security of individual authentication schemes could be tedious and prone to errors [43,48].

In the following section, we discuss the procedure used by SCTP to implement a vertical handoff between a UMTS and WLAN network.

6.4.1. SCTP-based vertical handoff

SCTP was originally designed as a specialized transport protocol for call control signaling in Voice over IP (VoIP) networks and has been specified by the third Generation Partnership Project (3GPP) to carry call signaling traffic in UMTS. In the base version of SCTP, the endpoints exchange IP addresses before the SCTP association is established, and these IP addresses cannot be changed during the session. However, in the integrated UMTS/WLAN environment, an MN may not have a fixed, previously known IP address. mSCTP allows endpoints to add, delete or change IP addresses during an active SCTP association using address configuration (ASCONF) messages. mSCTP is used for mobility management in hybrid wireless networks.

The SCTP-based vertical handoff [27] approach uses the multi-homing feature of SCTP so that a mobile client can have two IP addresses during the vertical handoff, one from the UMTS and the other from the WLAN. Fig. 7 shows the protocol architecture of the SCTP-based scheme. Both, the MN and the Fixed Server (FS) need to implement mSCTP. The MN also needs to support both, UMTS and WLAN at the physical and data link layers. The handoff procedure using mSCTP consists of the following steps:

1 Addition of IP address
2 Triggering of vertical handoff
3 Deletion of IP address

The FS is configured with an IP address, say, FS_IP. The MN is also allocated an IP address, UMTS_IP, in a UMTS cell

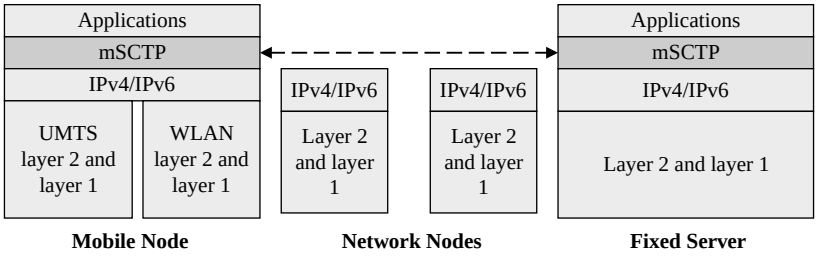

Fig. 7. Protocol Architecture using mSCTP.

and using this IP address it communicates with the FS via mSCTP. When the MN moves into a WLAN cell covered by a UMTS cell, it acquires a new IP address, WLAN_IP (assigned to it via router advertisements heard on the new network), and starts to add the WLAN IP address. The MN informs the FS of its new IP address by sending an ASCONF message to the FS with parameters set to 'add IP address' and WLAN_IP. The vertical handover triggering process allows the MN to trigger a handoff based on some decision rules. The UMTS-to-WLAN handoff is triggered by the MN sending an ASCONF message with parameters set to 'set primary address' and WLAN_IP. After the MN receives an acknowledgment (ACK) from the FS, the WLAN becomes the primary choice, and the traffic between the MN and the FS is routed through the WLAN. The WLAN-to-UMTS handoff is triggered by the MN sending an ASCONF message with parameters set to 'set primary address' and UMTS_IP. After the MN receives an ACK from the FS, the UMTS becomes the primary choice, and the traffic between the MN and the FS is routed though the UMTS. If the MN loses the signal from the LAN cell, it starts the delete IP address process. The MN sends an ASCONF message with parameters set to 'delete IP address' and WLAN_IP to request that the FS release the address WLAN_IP from its host routing table. After the MN receives an ACK from the FS, it deletes WLAN_IP from its address list, and WLAN_IP is released from the association.

6.4.1.1. Limitations of SCTP-based handoff scheme. The approach is based on the mSCTP protocol (which is mainly used for client-server services) where a client initiates a session with a fixed server. To support peer-to-peer services, mSCTP must be used with additional location management schemes which the original mSCTP protocol lacks.

6.5. Application layer handoff approach

The SIP-based handoff [29] approach is an application-layer solution to mobility management in heterogeneous networks. SIP is a scalable text-based protocol that offers a number of benefits, including extensibility and the provision for call/session control. The main entities in a SIP are user agents, proxy servers and redirect servers. A user is generally identified using an email like address user@userdomain, where user is the username and userdomain is the domain or numerical address. There exist various methods defined in SIP-INVITE, ACK, BYE, OPTIONS, CANCEL, and REGISTER. Terminal mobility requires SIP to establish connection either during the start of a new session, when the terminal or the MN has already moved to a different location, or during the middle of a session. The former situation is referred to, as pre-call mobility while the latter is known as mid-call mobility. Performing a vertical handoff during an ongoing session is similar to mid-call mobility. Fig. 8 shows how vertical handoff is performed using SIP. For mid-call mobility management, once the MN moves into a new network, it sends re-invites to the correspondent nodes (CNs) to participate in the call by sending a SIP_INVITE message. This INVITE message uses the same call identifiers as in the original call setup and contains the new IP address of the new location. Once the CN gets the updated information about the MN, it sends an acknowledge message and begins data transmission.

6.5.1. Limitations of SIP

Since, SIP is an application layer protocol the processing of SIP messages in the intermediate and destination servers may take considerable amount of time and introduce unacceptable handoff delays. Also, implementation of an application layer mobility scheme may require each application to be modified to support the mobility technique, which may not favor ease of deployment of this strategy.

6.6. Energy-efficient Handoff Approaches

6.6.1. Adaptive vertical handoff

The *Adaptive* scheme [23] for performing vertical handoffs is proposed to enable efficient discovery of wireless systems

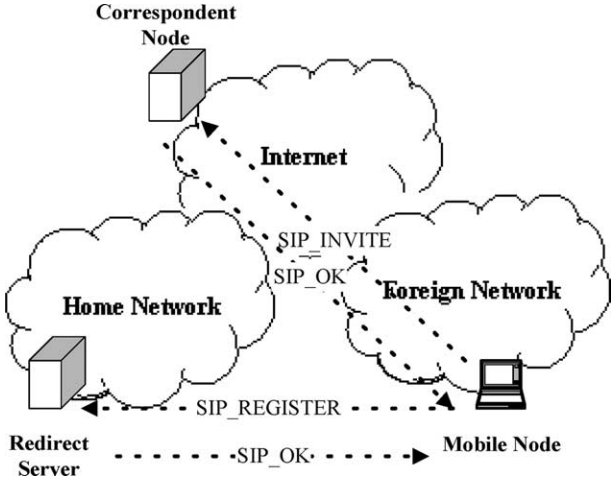

Fig. 8. SIP-based Vertical Handoff.

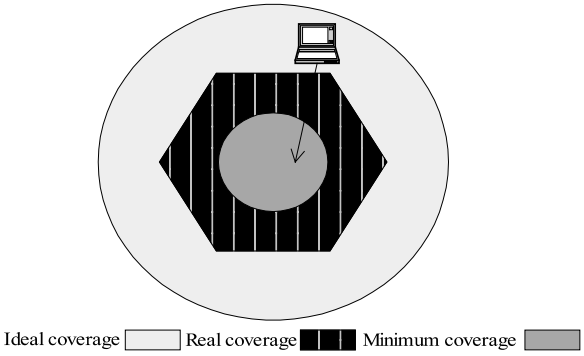

Fig. 9. Ideal Coverage vs. Minimum Coverage.

and to minimize the power consumption of a mobile device (caused due to unnecessary interface activation) during the handoff process. It introduces two new concepts: *the ideal coverage area* and the *minimum coverage area.* The *ideal coverage area* is the smallest circle that covers the real coverage and the *minimum coverage area* is the largest circle that can be included by the real coverage. These concepts are depicted in Fig. 9. With the Adaptive approach, the operator is required to publish information about the position of each base station, the radius of ideal coverage, the radius of the minimum coverage at the Location-Service Server (LSS).

The handoff process in this scheme is performed in two steps.

Interface Activation. The MN obtains the list of possibly reachable wireless networks from the LSS. It checks its current position and the possibly reachable wireless networks list periodically. If the MN finds that it is under the ideal coverage of a specific wireless network (except the current one), it activates its corresponding interface. Since, the interface is activated only when the MN enters the ideal coverage, the probability of receiving the service advertisement increases, thereby avoiding unnecessary interface activation and fast detection of reachable networks. Once the MN discovers the reachable wireless network, it begins to evaluate it using a utility function that quantifies the QoS provided by the wireless network from the viewpoint of running applications on the MN. It is assumed that a performance agent residing in the base station periodically announces the available resources information using beacons. If a MN is associated with a GPRS network and discovers the presence of a WLAN, the MN makes the handoff decision based on the utility ratio of the two wireless networks. The utility function of each application should be predefined and may be dependent upon factors such as effective bandwidth, cost, power consumption, etc. For example, if a utility function contains bandwidth and movement speed as factors, the reasons for the utility ratio's change could be as follows:

1. Increase in Uwlan/Ugprs: increase in effective bandwidth of the WLAN, decrease in the movement speed of the MN.
2. Decrease in Uwlan/Ugprs: decrease in effective bandwidth of the WLAN, increase in the movement speed of the MN.

Handoff Implementation. When the MN finds the target wireless network, it observes it for a 'stability period' which is defined as follows: Tstability = Utility target/Utility current. If the MN observes that the target network yields a better utility ratio for a time period equal to the stability period, then it implements the handoff.

Limitations of the Adaptive Vertical Handoff Scheme. The approach requires each operator to publish the information regarding the ideal coverage of its networks on the Location-Service Server, and to update this information upon changes in the ideal coverage areas or upon deployment of new networks. It also requires mobile nodes to query the LSS to obtain the information regarding the network coverage areas. This is not a very feasible strategy since the mobile node needs to know which LSS to query and also the IP address of each LSS. Furthermore, querying the LSS may require additional authentication schemes and introduce further delays in the handoff process.

6.6.2. WISE

The Wise Interface SElection [24] approach for vertical handoff between 3G and WLAN systems is proposed to reduce the energy consumption in mobile terminals, without degrading the throughput. It implements a mobile-assisted handoff based on the tight-coupling [32] approach in which a single service provider controls both 3G as well as WLAN networks.

Fig. 10 shows the network architecture of WISE. It introduces a new conceptual object called the Virtual Domain Controller (VDC) in a 3G-core network, which acts as a central point for controlling both, 3G as well as WLAN networks. The VDC is designed such that it is capable of managing the uplink and the downlink separately. It obtains the information about the downlink load from a base station or an access point and the information about the uplink load from a mobile node. Based on this information, the VDC balances the network load efficiently.

WISE aims to dynamically switch an activated network interface to another by comparing the energy consumption of the interfaces and selecting the interface that consumes the least amount of energy, unless that operation degrades the throughput after taking into consideration the network load. WISE operates in cooperation with the mobile nodes and the network. A mobile

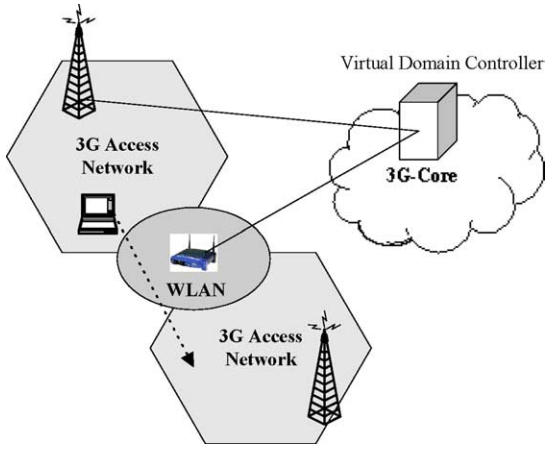

Fig. 10. WISE Network Architecture.

terminal makes a decision as to which network interface will consume less amount of energy under the current pattern of data traffic. If the selected network interface is different from the activated one, the MN sends a vertical handoff request to the VDC.

The VDC accepts or rejects its request after taking into consideration both the 3G networks and the WLANs. The VDC may reject the request for the handoff if that operation makes the overall network performance degrade. If the VDC accepts the handoff request, it triggers a handoff. Therefore, when a VDC receives a request for a vertical handoff, it checks the network load on the requested link on the target network by measuring the queue length of the associated BS or AP. Based on this information, The VDC makes a decision and sends a reply to the MT indicating in the message whether it has accepted/rejected the handoff. If the request to handoff is granted by the VDC, the MN triggers a handoff, else it waits for a specified amount of time and repeats the above process.

Limitations of the WISE Approach. The decision to handoff is performed based on the energy consumption of the interfaces and the network load. However, requirements of the applications running on the mobile terminal are not considered. (For example, some applications may require security.) A 3G connection may be able to fulfill its security requirements; however, switching to a WLAN to reduce the energy consumption by the active interface may provide the application with a greater bandwidth but not satisfy its security requirements.

6.7. TCP-based handoff strategies

A TCP-based handoff approach [30] is proposed to improve TCP performance during handoff in hybrid wireless networks. It focuses on hiding packet losses on the wireless link from the sender, so that the sender does not regard these errors as network congestion and decrease its data rate. In a hybrid network integrating different types of wireless mobile networks, one of the issues is to adjust the sender's data rate after moving into a new network environment. Each mobile network within a hybrid wireless network has a quite different capacity and coverage. Since WLAN and cellular networks provide different bandwidths to a user, a drastic change in data rate may occur when a MN is handed off. The latency of the vertical handoff is longer since it may need an authentication procedure when a MN enters into a new network. Considering these facts, a TCP sender needs to temporarily halt its data transmission during handoff to avoid a timeout and packet losses. In the TCP-based approach, the sender temporarily halts data transmission and stops its timeout timer. When the handoff is completed the data transmission is restarted in the *TCP slow start* state.

Basically, TCP operates in two states; Slow Start (SS) state in which a sender exponentially increases its data rate and Congestion Avoidance (CA) state in which a sender linearly increases its data rate. The main idea of this approach is to freeze TCP during handoff and to restart its data transmission in SS state after handoff. Since, the network environment is drastically changed in a new wireless network after vertical HO, it improves performance to start in a SS state and estimate the available bandwidth, rather than to start with the same bandwidth as before the vertical HO. The physical layer in MN measures the strength of the received signal and reports it to the radio resource control (RRC) layer in the MN's protocol stack. The RRC layer then determines the handoff triggering time and notifies the impending handoff to the TCP layer. In this approach, TCP uses an optional field in TCP header to identify the handoff situation. The optional field is defined as following.

HO optional field = 00: No Handoff
HO optional field = 10: Horizontal Handoff
HO optional field = 11: Vertical Handoff

The TCP receiver sends an ACK, setting HO optional field to the proper value according to the handoff type when a handoff is impending. It sends an ACK with 00 in HO optional field as soon as handoff is completed so that the sender can resume the data transmission without waiting for a timeout. This ACK message to inform the sender of an impending HO can be delivered multiple times to improve reliability. The TCP sender monitors the HO option field and adjusts its congestion window size. If it detects a horizontal handoff is occurring, the TCP sender stops the timeout timer and suspends data transmission until the handoff is completed. When the handoff is completed, the TCP sender resumes data transmission at the CA state with the same congestion window size as before handoff, since the MN moves into the same wireless environment. If a vertical HO is occurring, the TCP sender stops the timeout timer and holds data transmission until the handoff is completed. When the handoff is completed, the TCP sender resumes data transmission at the SS state.

Another proposed TCP-based handoff approach [33] achieves good TCP performance by maintaining a copy of the last outgoing or last incoming TCP ACK packet. Three copies of these ACKs are retransmitted to the sender or inserted to the mobile node's incoming TCP buffer to trigger the TCP fast retransmit algorithm and to immediately resume any active TCP connections after the handoff. This approach is used in cases where the mobile node performs a continuous handoff in which the network coverage of two or more base stations overlap and the mobile node has a pre-defined set of base stations to handoff to. However, in the non-continuous handoff case, acceptable TCP performance is achieved by stopping the sender from retransmitting until another point of attachment is found. This is achieved by setting the TCP receive window size to zero in the last outgoing or last incoming TCP ACK packet before the mobile node performs a handoff. This halts the transmission of TCP data. When a new point of attachment is found, the mobile node is made to advertise a non-zero TCP receive window size causing its TCP-connections to be resumed.

We present a summary of the above approaches for handoff management in Table 5. It has been observed that most of the proposed handoff schemes are mobile-controlled (handoff decision is made by the mobile node independently). Also, most of the existing approaches use a simple handoff decision

Table 5
A comparison of vertical handoff schemes

Handoff approach	Type of handoff	Handoff latency	Basis for handoff decision	Description
HOPOVER	Mobile-controlled	Moderate	Reception of WLAN signal	Addresses the problem of high handoff frequency that cannot be handled by traditional Mobile IP
WISE	Mobile-assisted	Low	Energy consumption of interfaces, network load	An energy-efficient interface selection method for vertical handoffs in tightly integrated systems
Hierarchical method	Mobile-controlled	Low	Reception of WLAN signal	A fast handoff scheme that reduces upward vertical handoff latency, packet loss and disruption
OMNICON	Mobile-controlled	Moderate	Reception of a strong WLAN signal	Improves mobile IP handoff by supporting packet scheduling and buffering
TCP-based method	Mobile-controlled	Low	Received WLAN signal strength and velocity	Improves the performance of TCP connections during vertical handoffs
SCTP-based	Mobile-controlled	Moderate	Reception of WLAN signal	Improves handoff delay and throughput using the multihoming capability and dynamic address configuration of SCTP
SIP-based	Mobile-controlled	High	Reception of WLAN signal	An application-layer solution for vertical handoffs in heterogeneous networks
Adaptive method	Mobile-controlled	Low	Energy consumption of interfaces, and QoS offered by networks	An effective system discovery method to reduce power consumption during vertical handoffs

method (based upon one or two factors only) and do not provide an intelligent decision strategy that takes into account important factors such as cost, security, user-preferences, application requirements, etc.

7. Research challenges for mobility management in future heterogeneous wireless networks

Several issues [3,14] still need to be resolved in order to build a complete solution for seamless roaming across various radio access systems. First, it is imperative that the user application session persists without timing out during mobility. However, current handoff management techniques do not provide such persistence and may time out before the handoff to a new subnet is achieved. This requires the user to have to restart the application at the new point of attachment. Since in a wireless environment, it is common to go out of range or coverage for a few seconds to a minute and handoff between subnets takes a finite amount of time (while handling mobility), it is necessary to ensure that the application session state is maintained. Second, mobile users want fast Internet access and seamless roaming capability without the complexities of configuring devices, inputting authentication parameters, entering and changing user preferences, updating parameters and receiving bills from multiple service operators [30]. User demand for new services featuring seamless roaming and hassle-free authentication and configuration requires new technologies that will simplify device configuration and authentication in order to make seamless roaming possible.

Next generation heterogeneous systems will be expected to provide end-users with a usage model for wireless devices that will allow them to roam seamlessly among different networks, using multimodal wireless devices. Such terminals will be able to roam and communicate freely and access the Internet across both WLANs and WWANs. The task of managing authentication between client devices and networks and configuration of various parameters should become automatic and transparent to the user. This is a highly desirable goal, but much work remains to make this a reality. Technology needs to be improved in multiple areas, including IP (Internet Protocol) address management, billing management, roaming services, radio technology and security. To summarize, we list some of the main issues involved in supporting seamless access across hybrid wireless networks.

7.1. Multimodality

To be able to access different types of network technologies, it is essential to develop multimode terminals that have the capability to adapt to various technologies differing in the frequency of operation, data rate, access technology, etc. In addition, these devices should be smart in order to shift the network selection process from the user to the terminal.

7.2. Efficient and seamless roaming

7.2.1. Detection of network coverage

Mobile devices must be able to detect easily and efficiently the presence of a new network coverage area by processing signals sent from different wireless systems, which differ in access protocols and are incompatible with each other.

7.2.2. Selection of the most appropriate access network

Wireless networks differ in terms of various factors, and the task of selecting the most appropriate network is complex with respect to the needs and services of an end-user.

7.2.3. Handoffs

Moving across various network technologies presents the need to design intelligent and efficient handoff techniques with minimum latency and packet losses. Handoff management techniques should allow mobile users to roam among multiple wireless networks in a manner that is completely transparent to applications and disrupts connectivity as little as possible. In addition, in hierarchically structured wireless systems, the choice of the best wireless network for location and handoff management poses a significant challenge since different overlay levels might have widely varying characteristics. Moreover, in traditional mobile systems only horizontal handoff has to be performed where as in next generation wireless access systems, both horizontal and vertical handoff should be performed.

7.3. Quality of service (QoS)

7.3.1. QoS requirements

Next generation wireless systems will consist of various access technologies with differing parameters interconnected with a common IP-backbone. On the other hand, mobile terminals will run various types of multimedia applications. These applications will have varying requirements, which may not be satisfied by the best-effort IP framework. Providing proper QoS including satisfactory bandwidth, throughput, reliability, perceived quality, and costs in a heterogeneous mobile computing environment presents a major challenge. Mobility management in such environments also introduces new issues such as timely service delivery, QoS negotiation during inter-system handoff, etc. The main problem in providing QoS in a mobile environment is the coupling of QoS and mobility management, i.e. how to keep providing the same level of quality to the packet flow during and after the handoff. No currently available technology exists that would allow fully seamless IP mobility. Also, these exists a broad range of QoS mechanisms, which makes it impossible for an application designer to select a single QoS framework that is supported across different types of networks.

7.3.2. Security

The existence of heterogeneous wireless access systems makes it very challenging to provide a constant level of security to data flows during host mobility. This requires the development of adaptive security mechanisms.

7.4. Billing and pricing

Existence of multiple networks gives rise to the presence of several network operators. It is imperative to develop billing [34] and accounting mechanisms to provide ease and satisfaction to the end-user.

7.5. Transfer of contextual information

7.5.1. Context transfer

Handoff in heterogeneous environments requires the transfer of contextual information for information flow between the mobile node and the network. Context transfer is designed to allow access routers to exchange state information regarding a mobile node's packet treatment. A context transfer protocol aims to minimize the impact of transport/routing/security-related services on the handoff performance. When a mobile node moves to a new subnet, it needs to continue services that have already been established at the previous subnet. Such services are called 'context-transfer candidate services' [31] and include Authentication, Authorization and Accounting (AAA) profiles, IPSec states, QoS policy, etc. A context-transfer protocol needs to be developed to allow quick re-establishment of context-transfer candidate services at the new network and thus enable seamless operation of application streams during mobility.

7.5.2. Security of context transfer

Sensitive context information must be protected to preserve user privacy and maintain security. The security context shared between different domains represents a level of trust between them. If context is transferred to another intermediate device, whether in the same domain or different domains, ideally, the same level of trust must be in place as between the intermediate device and the authentication server. Network mobility management can be optimized if contexts are transferred from one entity to another securely.

7.6. Inter-service provider compliance

An effective handoff scheme must be capable of being deployed on the existing network infrastructure which consists of different networks owned by different service providers and should not require major modifications to the existing networks.

8. Conclusion

In this article, we have presented a detailed discussion of various components of handoff management as well as recent handoff techniques for mobility across heterogeneous wireless access networks. We have also discussed about various features of multimodal mobile terminals that will enable the deployment of efficient mobility solutions. We have presented several handoff implementation issues and challenges that still need to be addressed to allow seamless mobility in future radio access systems. We found that current roaming techniques are not well suited to support continuity of application sessions in heterogeneous mobile environments. Extensive research and technological improvements in the areas of radio transmission, mobility management, and implementation of multimodal devices will lead to the development of an adaptable network environment where users will be able to use self-configuring devices that will enable them to realize seamless global roaming efficiently and cost-effectively. These systems are expected to provide users the ability to execute various mobile multimedia applications while moving across different types of networks environments. Work is in progress to migrate the existing networks to an optimized system in which several different types

of technologies can interoperate and co-exist in a way so as to provide a completely seamless mobility experience to end-users.

9. Uncited references

[8]. [15]. [18]. [19].

Acknowledgements

This work was supported by grants from Microsoft (Seattle), Sun Microsystems (Palo Alto) and Ixia Corporation (Calabasas). We express our gratitude to the anonymous reviewers and the editor for their remarks and suggestions which helped to improve the paper.

References

[1] M. Stemm, R. Katz, Vertical handoffs in wireless overlay networks, ACM Mobile Networking, (MONET), Special Issue on Mobile Networking in the Internet 3 (4) (1998) 335–350.

[2] R. Inayat, R. Aibara, K. Nishimura, A seamless handoff for dual-interfaced mobile devices in hybrid wireless access networks, Proceedings of the 18th IEEE International Conference on Advanced Information Networking and Applications 1, 2004 pp. 373–378.

[3] Y.H. Suk, Y. Kai Hau, Challenges in migration to 4g mobile systems, IEEE Communications Magazine 41 (12) (2003) 54–59.

[4] Chakravorty R., Vidales P., Patnapongpibul L., Subramanian K., Pratt I., Crowcroft J., On Inter-network Handover Performance using IPv6, http://www.cl.clam.ac.uk/users/rc277/handovers.pdf

[5] C. Wen-Tsuen, L. Jen-Chu, H. Hsieh-Kuan, An adaptive scheme for vertical handoffs in wireless overlay networks Proceedings of the First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services, 2004, pp. 111–112.

[6] H. Haijie, C. Jianfei, Improving TCP performance during soft vertical handoff Proceedings of the 19th International Conference on Advanced Information Networking and Applications, 2005, pp. 329–332.

[7] J. McNair, Z. Fang, Vertical handoffs in fourth generation multinetwork environments, IEEE Wireless Communications 11 (3) (2004) 8–15.

[8] D. Axiotis, T. Al-Gizawi, K. Peppas, E. Protonotarios, F. Lazarakis, C. Papadias, P. Philippopoulos, Services in internetworking 3G and WLAN environments, IEEE Wireless Communication 11 (5) (2004) 14–20.

[9] D. Fan, L. Ni, A. Esfahanian, HOPOVER: a new handoff protocol for overlay networks, Proceedings of the IEEE International Conference on Communications, 5, 2002, pp. 3234–3239.

[10] McCann S., Groting W., Pandolfi A., Hepworth E., Next Generation Multimode Terminals, http://www.roke.co.uk/download/papers/next_generation_multimode_terminals.pdf

[11] C. Perkins, Mobile networking in the internet, Mobile Networks and Applications 3 (1998) 319–334.

[12] I. Akyildiz, X. Jiang, S. Mohanty, A survey of mobility management in next-generation all-IP-based wireless systems, IEEE Wireless Communications 11 (4) (2004) 16–28.

[13] N. Montavont, T. Noel, Handover management for mobile nodes in Ipv6 networks, IEEE Communications Magazine 40 (8) (2002) 38–43.

[14] A. Mohammad, A. Chen, Seamless mobility requirements and mobility architectures, Proceedings of IEEE Global Telecommunications Conference, 3, 2001, pp. 1950–1956.

[15] N. Deshpande, Enabling seamless roaming between wireless networks, Intel DeveloperUPDATE Magazine (2002).

[16] J. Hoffmeyer, Radio software download for commercial wireless reconfigurable devices, IEEE Communications Magazine 42 (3) (2004) 526–532; **Summary:** Commercial wireless end user terminals are becoming increasingly complex in order to integrate in a common package distinct separate capabilities formerly implemented in multiple single-function, single-band devices. Software is an increasingly impor

[17] N. Nakajima, R. Kohno, S. Kubota, Research and developments of software-defined radio technologies in Japan, IEEE Communications Magazine 39 (8) (2001) 146–155; **Summary:** Software-defined radio technologies are attractive for future mobile communication systems because of reconfigurable and multimode operation capabilities. The reconfigurable feature is useful for enhancing functions of equipment without replacing har.....

[18] P.B. Kenington, Emerging technologies for software radio, Electronics and Communication Engineering Journal 11 (2) (1999) 69–83.

[19] M. Dillenger, S. Buljore, Reconfigurable systems in a heterogeneous environment, Software Defined Radio: Architectures, Systems and Functions, Wiley, 2003.

[20] Dimitris M., Seamless Mobility, ARC Group, Motorola, http://www.motorola.com/mot/doc/1/1874_MotDoc.pdf

[21] SDR Forum, http://www.sdrforum.org

[22] Software Defined Radio, White Paper, WIPRO Technologies, August, 2002, http://www.broadcastpapers.com/broadband/WiproSDRadio.pdf

[23] H. Badis, A. Khaldoun, Fast and efficient vertical handoffs in wireless overlay networks, Proceedings of the 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 3, 2004, pp. 1968–1972.

[24] N. Minji, C. Nakjung, S. Yongho, C. Yanghee, WISE: Energy-efficient interface selection on vertical handoff between 3G networks and WLANs Proceedings of the 15th IEEE International Symposium, IEEE, 2004.

[25] Context Transfer, Handoff Candidate Discovery and Dormant mode Host Alerting, Seamoby WG, IETF, http://www.ietf.org/html.charters/seamoby-charter.html

[26] Software Defined Radio, White Paper, WIPRO Technologies, August, 2002

[27] M. Li, Y. Fei, V. Leung, T. Randhawa, A new method to support UMTS/WLAN vertical handover using SCTP, IEEE Wireless Communications 11 (4) (2004) 44–51.

[28] S. Sharma, B. Inho, Y. Dodia, C. Tzi-cker, Omnicon: a mobile IP-based vertical handoff system for wireless LAN and GPRS links Proceedings of the Second IEEE Annual conference on Pervasive Computing and Communications, 2004, pp. 155–164.

[29] N. Banerjee, W. Wu, K. Basu, S. Das, Analysis of SIP-based mobility management in 4G wireless networks, Computer Communications (2004).

[30] S.-E. Kim, J.A. Copeland, TCP for seamless vertical handoff, Proceedings of the IEEE Global Telecommunications Conference, 2, 2003, pp. 661–665.

[31] Georgiades M., Wang H., Tafazolli R., Security of context transfer in future Wireless Communications, http://www.ambient-networks.org/docs/Sceurity_of_Contex_Transfer_for_Future%20_Mobile_Communications.pdf

[32] A. Salkintzis, Interworking between WLANs and third-generation cellular data networks, Proceedings of the 57th IEEE Semiannual Vehicular Technology Conference, 3, 2003, pp. 1802–1806.

[33] P. Vidales, L. Patanapongpibul, G. Mapp, A. Hopper, Experiences with Heterogeneous Wireless Networks, Unveiling the Challenges, www-lce.eng.cam.ac.uk/～pav25/publications/HetNets04-Vidales.pdf

[34] Real-time Revenue Management of Mobile Content & Entertainment, Inside Billing, 2004, http://www.portal.com/news_events/press_releases/release_2003/microsoft.htm

[35] C. Perkins, IP Mobility Support for Ipv4, RFC 3344, August 2002

[36] D. Johnson, Mobile Support in IPv6, RFC 3775, June 2004

[37] J. Manner, M. Kojo, Mobility Related Terminology', RFC 3753, August 2004

[38] J. Kempf, RFC Reasons for Performing Context Transfers between Nodes in an IP Access Network, RFC 3374, September 2002

[39] H. Duong, A. Dadej, S. Gordon, Proactive context transfer in WLAN-based access networks Proceedings of the second ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots, 2004.

[40] S. Leggio, J. Manner, K. Raatikainen, Achieving seamless mobility in IP-based radio access networks, IEEE Wireless Communications 12 (1) (2005).

[41] M. Leibsch, A. Singh, RFC Candidate Access Router Discovery', Internet Draft draft-ietf-seamoby- card-protocol-08.txt, September 2004

[42] J. Loughney, M. Nakhjiri, C. Perkins, R. Koodli, RFC Context Transfer Protocol. Internet Draft, draft-ietf-seamoby-ctp-11.txt, February 2005.

[43] W. Eddy, At what layer does mobility belong?, IEEE Communications Magazine 42 (10) (2004) 155–159.

[44] M. Chang, M. Lee, S. Koh, Transport layer mobility support utilizing link signal strength information, IEICE Transaction on Communications E87-B (9) (2004) 2548–2556.

[45] L. Ong, J. Yoakum, An Introduction to Stream Control Transmission Protocol, RFC 3286, May 2002

[46] E. Kohler, M. Handley, S. Floyd, Datagram Congestion Control Protocol, Internet-draft, (2005)

[47] I. Aydin, W. Seok, C. Shen, Cellular SCTP: a transport-layer approach to internet mobility Proceedings of 12th International Conference on Computer Communications and Networks (ICCCN 2003), Dallas, Texas 2003.

[48] W. Eddy, J. Ishac, M. Atiquzzamman, An Architecture for Transport Layer Mobility, Internet-draft, 2004.

[49] S. Fu, M. Atiquzzaman, L. Ma, W. Ivancic, Y. Lee, J. Jones, J. Lu, TraSH: A Transport Layer Seamless Handover for Mobile Networks, University of Oklahoma Technical Report OU-TNRL-04-10, January 2004.