

MCIP: A 3G/IP Interworking System Supporting Inter-Cluster Soft Handoff*

XIN LIU¹ and WEIHUA ZHUANG²

¹295 Phillip Street, Waterloo, Ontario, Canada N2L 3W8

E-mail: xliu@rim.com

²Centre for Wireless Communications (CWC), Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

E-mail: wzhuang@bbcr.uwaterloo.ca

Abstract. In this paper, a novel 3-tier Mobile Cellular IP (MCIP) access network is proposed for interworking between a third generation (3G) wireless cellular system and a wireline Internet Protocol (IP) based network. An inter-cluster hard handoff scheme and an inter-cluster soft handoff scheme are proposed, based on the 3-tier MCIP system model, the core network protocol stacks, and the underlying MCIP routing algorithm. The core network protocol stack is presented to integrate the 3G radio interface and the IP-based core network, and to provide the access network with capability to support soft handoff macroscopic space diversity. The MCIP hard and soft handoff schemes are compared with the hard handoff schemes used in the Cellular IP and HAWAII access networks. The MCIP access network is more efficient in terms of signaling cost, but has the same scalability as Cellular IP and HAWAII. Both MCIP hard and soft handoff schemes enable IP packets to be delivered within the MCIP access network in-order without loss and duplication, a highly desired attribute for real-time multimedia applications. The advantages of supporting soft handoffs and quality-of-service (QoS) provisioning for real-time services are achieved at slightly increased system complexity.

Keywords: cellular mobile communications, handoff, Internet Protocol (IP), wideband code-division multiple access (CDMA), wireless and wireline interworking

1. Introduction

The next-generation wireless networks are evolving toward a versatile Internet Protocol (IP) based network that can provide various real-time multimedia services to mobile users. Recent studies on the interworking between the third generation (3G) wireless systems and IP-based wireline networks (3G/IP) have been intensive [1–4]. Although a general consensus on the infrastructure for the 3G/IP interworking is yet to be reached, it is expected that the Mobile IP enabled Internet will service as the backbone network to provide global coverage, while the front-end 3G wireless segments will support seamless user roaming. The 3G wireless networks will adopt micro/pico-cellular architectures for various advantages including higher data throughput, greater frequency reuse, and location information with finer granularity [5]. In this environment, the handoff rate grows rapidly, including both intra-cluster handoff and

*This work was supported by a research grant from the Natural Science and Engineering Research Council (NSERC) of Canada. The authors wish to thank the anonymous reviewers for their helpful comments and suggestions which improve the presentation of this paper.

inter-cluster handoff.¹ On the other hand, 3G networks are expected to support a wide range of multimedia services with different quality-of-service (QoS) requirements, which are sensitive to packet loss, delay, and delay jitter. Consequently, provisioning of seamless fast handoff is extremely crucial for the successful deployment of 3G/IP interworking systems.

According to the number of base transceiver stations (BTSs) simultaneously involved during a radio link transfer process, a handoff can be distinguished as a hard handoff or a soft handoff. Soft handoff is supported only by code-division multiple access (CDMA) technology, while hard handoff is applicable in any cellular systems. In a hard handoff, the mobile node (MN) communicates with only one BTS at any time instant when crossing from one cell to another. There is a definite decision to switch an ongoing call from one BTS to another. On the contrary, during a soft handoff, an MN transmits/receives different copies of the same radio signal simultaneously to/from more than one BTSs when crossing from one cell to another. The collection of all the BTSs connected with the MN at a given time is called the active set. During the soft handoff process, the MN monitors the received signal levels broadcast from neighboring BTSs, compares them to a set of thresholds, and reports accordingly back to the network. Based on this information, the network informs the MN to add or remove BTS(s) from its active set.

A unique feature of the CDMA systems is the support of soft handoffs, which can effectively increase the capacity, reliability, and coverage range of the wireless systems at the cost of higher complexity [6–8]. The implementation of fast soft handoff is also important for the efficient delivery of real-time services, e.g., low-rate video streaming to mobile users. However, to our best knowledge, very limited research on inter-cluster soft handoff in an IP-based 3G wireless network is available in the open literature.

Mobile IPv6 [9] provides a simple and scalable global mobility solution for connecting mobile users to the Internet; however, it does not emphasize the support of fast and seamless handoffs in the wireless mobile domain. The 3G wireless systems, on the other hand, are expected to offer smooth mobility support but are built on complex networking infrastructures that lack the flexibility offered by IP-based solutions. There are considerable research activities and proposals, such as Cellular IP [10, 11], HAWAII [12], Hierarchical Foreign Agents [13], etc., to overcome the Mobile IP registration latency. Among these solutions, Cellular IP supports passive connectivity and paging which are fundamental features for improving scalability and minimizing power consumption of mobile terminals. However, Cellular IP aims at providing high-speed packet radio access to the Internet with the design principle of lightweight nature. Two characteristics of Cellular IP prevent it from working as the protocol connecting 3G wireless access points to the Internet for real-time services to the mobile users. First, handoffs are handled by the IP layer itself, independent of the radio interface. If the handoff happens in the course of the transmission of a long IP packet, the packet will be lost and the loss cannot be recovered for real-time services; Second, only mobile-controlled hard handoff schemes are supported. Without modification to its routing and handoff schemes, Cellular IP cannot support soft handoffs [14]. The same problems exist in HAWAII and Hierarchical Foreign Agents solutions as well.

In this paper, we propose a novel 3G/IPv6 interworking system model and investigate fast inter-cluster handoffs in the system. We also present a fast inter-cluster soft handoff

¹ Throughout this paper, a cluster is defined as a radio access network, referred to as Universal Terrestrial Radio Access Network (UTRAN) in the Universal Mobile Telecommunication System (UMTS) and Radio Access Network (RAN) in cdma2000 systems.

scheme based on the so-called *Mobile Cellular IP (MCIP)* routing, which is motivated by Cellular IP, but supports QoS guarantee for real-time services. In Section 2, a novel 3-tier 3G/IP interworking system model (i.e., the MCIP access network) is proposed. Within the framework of the MCIP access network, a core network protocol stack is designed to integrate the 3G radio interface and the IP-based core network, and to provide the access network with capability to support soft handoff. A simplified radio interface model based on the wideband CDMA (WCDMA) proposals is also discussed. Section 3 is dedicated to the MCIP routing and inter-cluster handoff schemes. Both soft handoff and hard handoff supported by the MCIP access network are discussed. Section 4 compares the signaling efficiency of the MCIP soft and hard handoff schemes with those of the Cellular IP and HAWAII protocols. Concluding remarks of this research are given in Section 5.

2. System Model and Protocol Architecture

2.1. THREE-TIER SYSTEM MODEL

Under the assumption that the backbone network is the mobile IPv6 enabled Internet, the proposed system model is illustrated in Figure 1, which has three tiers. The first tier is a Gateway router which connects the MCIP access network to the Internet backbone; The second tier is a mesh of base station controllers (BSCs) which communicate with each other by regular IP routing and MCIP routing; The third tier consists of BTS clusters, each connecting to a

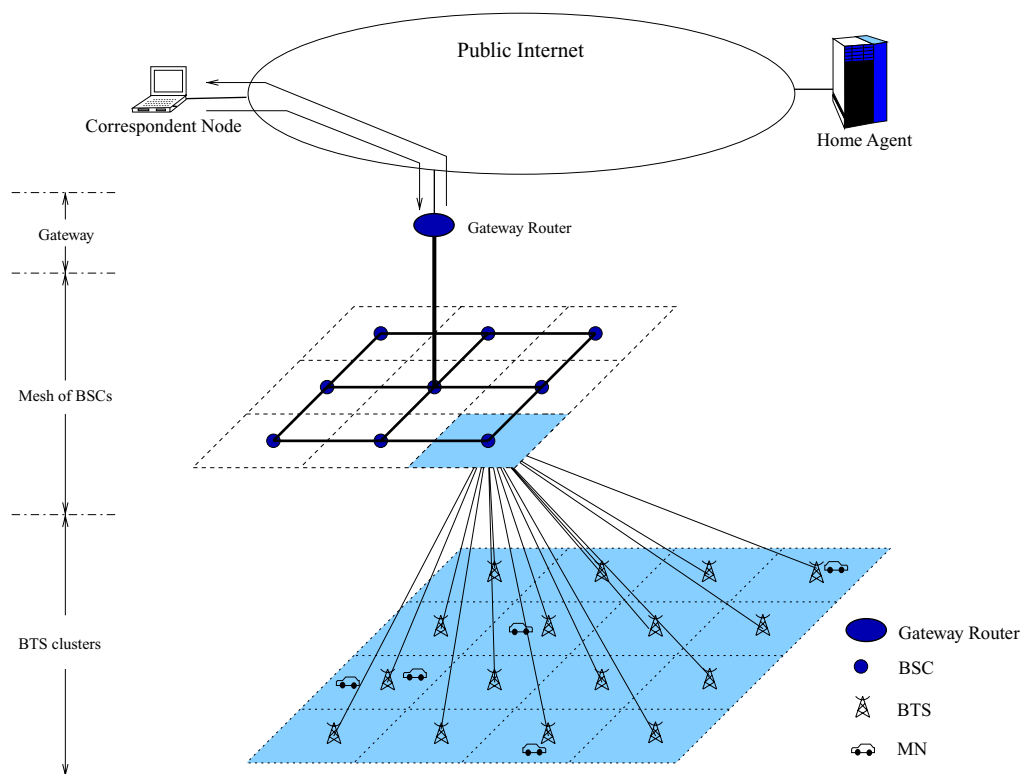


Figure 1. The 3-tier system model.

BSC. All the MNs within the Gateway domain use the Gateway's IP address as their *care-of address*. IP packets originated from an MN are routed to the Gateway router by MCIP routing within the Gateway domain, and from there to its correspondent node (CN) by Mobile IPv6; IP packets destined to the MN are routed by Mobile IPv6 from the CN to the Gateway router, and then routed by MCIP to the MN.

A BSC in the system model has two roles: a router built on top of both the regular IPv6 and the MCIP routing engines, and a base station controller as defined in 3G systems (also called *Radio Network Controller* in 3GPP specifications and *Selection and Distribution Unit* in 3GPP2 specifications). In the network layer, a BSC executes regular IP routing and MCIP routing algorithms to route IP packets to other system elements, such as BSCs or the Gateway router. If the BSC is the *serving BSC* of an MN, it works as the *edge router* of the MN and performs the Layer 2 (i.e., medium access control (MAC) sublayer and radio link control (RLC) sublayer) processing of the IP packets for the MN. The BSC and BTSs connected to it comprise a *Radio Access Network* (RAN). The BSC controls all the BTSs within the RAN, i.e., it is responsible for all radio-related functionalities in every cell, such as load and congestion control, admission control, packet scheduling, power control, etc. In this regard, the BSC is also referred to as the *controlling BSC* of the BTSs in the RAN. Handoffs between different cells within the same RAN are handled by the Intra-BSC handoff control algorithms inside the controlling BSC. As in the GSM/GPRS systems, a BSC can control up to several hundred BTSs, and it may co-locate with one of its controlled BTSs [15]. To facilitate inter-cluster soft handoffs, adjacent BSCs are connected by direct links [16, 17].

The BTSs are the access points for MNs in the MCIP access network. Aside from performing Layer 1 functions (such as modulation/demodulation, scrambling/descrambling, spreading/despreading, coding/decoding, etc.), a BTS plays two other roles in the MCIP system. First, it works as a *bridge* in the radio interface between an MN and its serving BSC, delivering basic data units that are received or to be transferred over the air. There is no IP layer connection between a BTS and MNs in the cell. Secondly, it takes part in the radio resource management within its cell under the control of its controlling BSC, such as forward link close-loop fast power control, measurement of air interface load, etc. For this purpose, a BTS keeps an IP layer connection with its controlling BSC to obtain system parameters and submit measurement reports.

The protocol architecture used in the MCIP access network is shown in Figure 2, where APP stands for Application Layer, RRM for Radio Resource Management, TCP for Transmission Control Protocol, UDP for User Datagram Protocol, DCH-FP for Dedicated Transport Channel – Frame Protocol, and PHY for Physical Layer. Two interactive and overlapping protocol stacks are defined in the MCIP system: the *radio interface* protocol stack, and the *core network* protocol stack. The radio interface and core network protocols work in conjunction with each other to ensure that the IP packets received by the MCIP Gateway from the CN arrive at the MN, and that the IP packets originated from the MN are delivered to the Gateway, in-order without loss and duplication.

2.2. SIMPLIFIED RADIO INTERFACE

In MCIP, the radio interface protocol stack is mainly defined within a radio access subsystem (i.e., the RAN). The left side in Figure 2 illustrates the radio interface protocol stack in a RAN, where the controlling BSC of the BTS is also the serving BSC of the MN. The serving BSC for an MN is the BSC that performs Layer 2 processing of the IP packets for

the MN to/from the MAC sublayer Transport Blocks (TBs). The radio resource management operations, such as the mapping of radio bearer parameters into air interface transport channel parameters, the handoff decision, and outer loop power control, are executed in the serving BSC [16]. An MN attached to the MCIP access network has one and only one serving BSC.

The overall function of the radio interface is to provide bandwidth-on-demand services to the IP layer. It comprises the CDMA layer [16, 18–22], the Layer 2 including the MAC sublayer [16, 23] and the RLC sublayer [16, 24], and the RRM layer [25]. The radio interface is the most flexible and complex part of the radio access system; the complete discussion of it is beyond the scope of this work. To evaluate the handoff performance, we give a simplified radio interface model here, as shown in Figure 3, which can be considered as a subset of WCDMA air interface. As illustrated in Figure 3, two Radio Bearers (RBs) are used by the air interface to provide services to the IP layer: one for user traffic, and the other for control signaling. Each of the RBs is served by a single *Dedicated Physical Channel (DPCH)* at the CDMA layer. The bit rate of the signaling RB is 7.2 kbps. There are two cases considered for the user traffic RB: variable rate of 144k, 72k or 36 kbps for packet data traffic (non-real-time service), and 7.2 kbps constant rate for packet switched voice traffic (real-time service). It is worth to note that the radio interface is no longer divided into “Control Plane” and “User Plane” as that in the WCDMA air interface.

The UMTS system, which utilizes the WCDMA standard as its radio interface, is built on top of a complex networking infrastructure. Its control signaling and user data are processed by two different network layer protocols, and may be routed via different system elements. In order to accommodate the two different protocol stacks, the WCDMA radio interface is divided into Control Plane and User Plane, where the services offered by the RLC to the network layer in the control plane are called *Signaling Radio Bearers* and the services in the user plane are called *Radio Bearers*. On the contrary, the MCIP system is built on top of IPv6. The RRM layer is considered as a specific *application layer protocol* running on top of UDP/IP, the same as other application layer protocols, such as the SNMP (Simple Network Management Protocol)

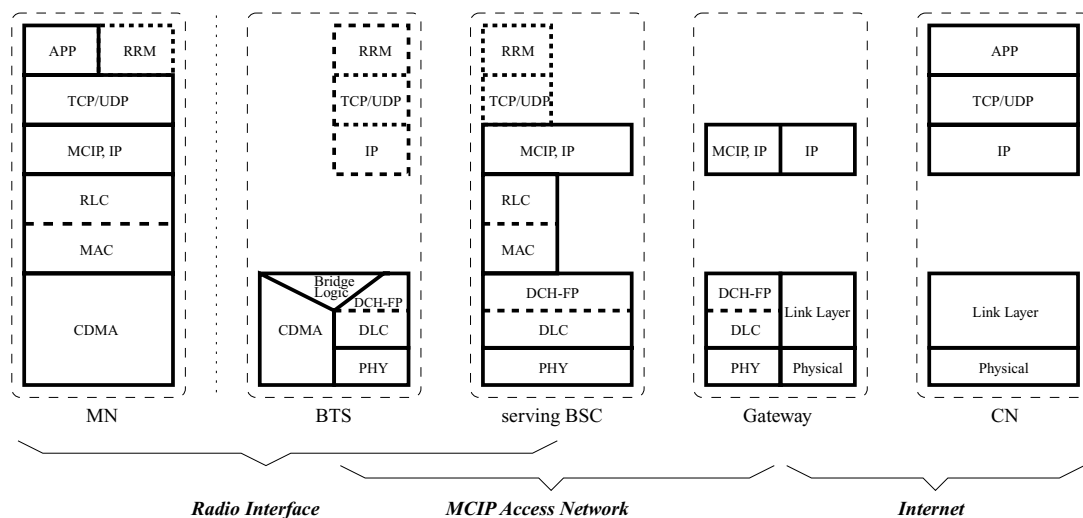


Figure 2. Protocol stacks in the MCIP system elements.

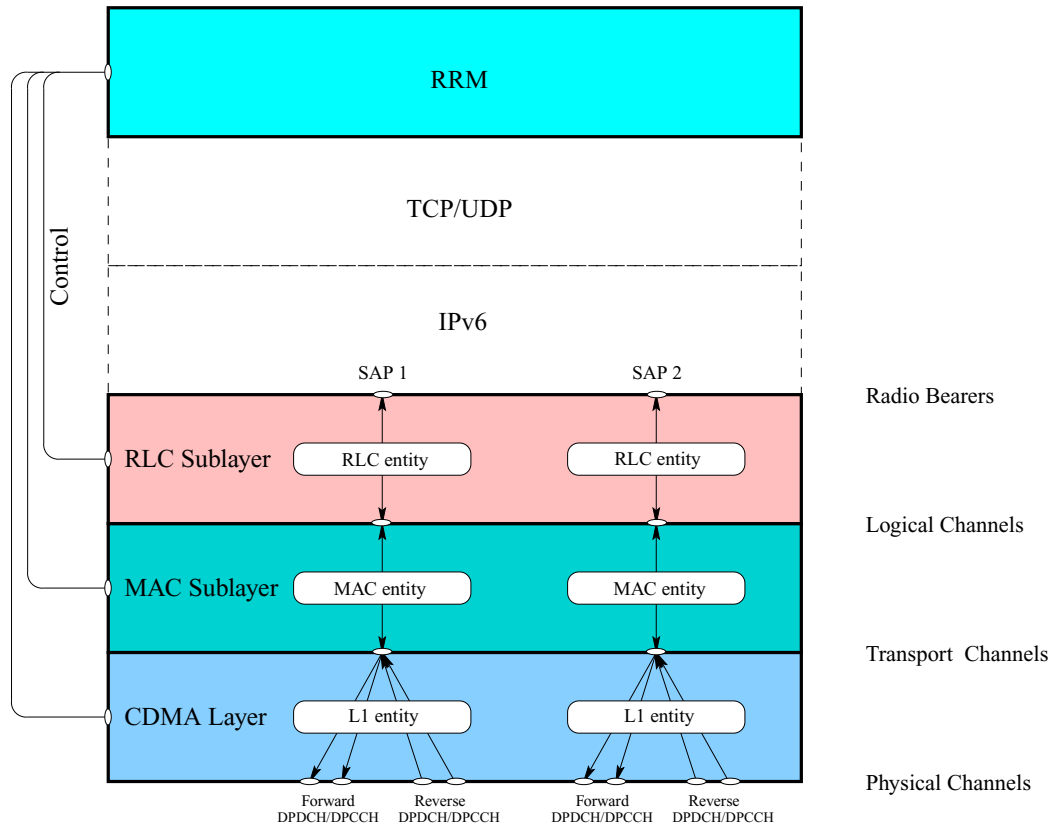


Figure 3. Simplified radio interface model.

running in the Internet routers. Therefore both the user data and control signaling traffic are processed by the same protocol stack. The radio interface has no knowledge of which RB is used by signaling and which RB serves the user traffic. For presentation simplicity, throughout this paper, the term *signaling RB* refers to the RB carrying the signaling traffic, and *user RB* refers to the RB serving the user traffic.

Unlike the wireline TCP/IP networks which employ strictly layered protocol architecture, that is, data exchange vertically between different protocol stack layers happens only between adjacent layers, the MCIP air interface is a loosely layered protocol architecture. The RRM protocol has control access to the RLC sublayer, the MAC sublayer and the CDMA layer, as well as transmits/receives application layer messages to/from the TCP/UDP layer. The RRM layer also has the capability to receive measurement reports from the RLC, MAC and physical layers. For example, when an MN needs to initiate a connection to the network, the MN sends a connection setup request message via the random access channel to the BSC. The RRM at the BSC performs call admission control based on the availability of the current radio resources, where the availability information is obtained by measurement performed at the physical layer and MAC sublayer. If the request is admitted, the RRM takes the responsibility to select appropriate RLC, MAC and physical layer entities and sets appropriate parameters to these entities according to the QoS requirements [16, 26].

In summary, the proposed radio interface can be mapped to those in 3G standards. For example, the RAN corresponds to the Universal Terrestrial Radio Access Network (UTRAN),

the BSC corresponds to Radio Network Controller (RNC), and the BTS corresponds to Node B in UMTS. The main difference from the 3G standards lies in the network layer. In addition, the proposed loosely layered protocol architecture facilitates efficient mobility and resource management which is adaptive to the status of the lower layers. Details of how the different layers of the radio interface coordinate among each other to provide variable bit rate services to the upper layers via a dedicated channel, and the design of the MAC PDUs (Protocol Data Units) and the RLC PDUs are given in [14].

2.3. CORE NETWORK PROTOCOLS

In Cellular IP access systems or Internet, the Data Link Layer (Layer 2) is required only to provide service to the network layer, i.e., to offer IP packet delivery services between two nodes connected by a direct link; but in the MCIP access system, since RAN is introduced to work as the radio access subsystem, the situation becomes more complicated. Examining the radio interface shown in Figure 2, the CDMA layer terminates at the MN and the BTS, but the MAC sublayer and RLC sublayer span from the MN to its serving BSC. Some mechanism is needed to *extend the transport channels* from the BTS to the serving BSC. Therefore, a *bridge logic*, which works as the proxy of the MAC sublayer entities, is placed on top of the CDMA layer at the BTS, and a *DCH-FP (Dedicated Channel – Frame Protocol)* sublayer is introduced to Layer 2 of the core network protocol stack. By doing so, in addition to providing packet data transmission services to the network layer (IP layer), Layer 2 is able to provide transparent packet data transmission service between the *bridge logic* at the BTS and the MAC sublayer at the serving BSC. As observed from Figure 2, the DCH-FP sublayer and DLC sublayer comprise Layer 2 in the core network. The DLC sublayer can be a data link layer protocol which supports metropolitan area networking, such as Asynchronous Transfer Mode (ATM) and Frame Relay.

The interconnection of the radio interface and the core network is as shown in Figure 2. Although the DCH-FP sublayer is not part of the radio interface, it connects the CDMA layer and the MAC sublayer together, or in other words, it extends the transport channel from the BTS to the serving BSC. The DCH-FP sublayer integrates the radio interface protocol stack and the core network protocol stack smoothly [14].

2.4. SPACE DIVERSITY PROCEDURES IN SOFT HANDOFF

The radio interface protocol stack in Figure 2 should be extended to two or more RANs for inter-cluster handoffs, as illustrated in Figure 4 for two RANs during an inter-cluster soft handoff.

Although there is no limit for the number of BTSs in an active set, it has been shown that no significant additional benefit can be achieved by having more than two BTSs [27]. Therefore, the maximum active set size is chosen to be 2 hereafter for presentation clarity. During a soft handoff, the MN is in the overlapping area of two cells. The information exchanges between the MN and the core network take place via two air interface channels from two BTSs simultaneously. Macroscopic space diversity provided by the different physical locations of the two BTSs is utilized to improve system performance.

The protocol stack for the space diversity combining/splitting procedure during an inter-cluster soft handoff is shown in Figure 4. The combining/splitting procedure in an intra-cluster

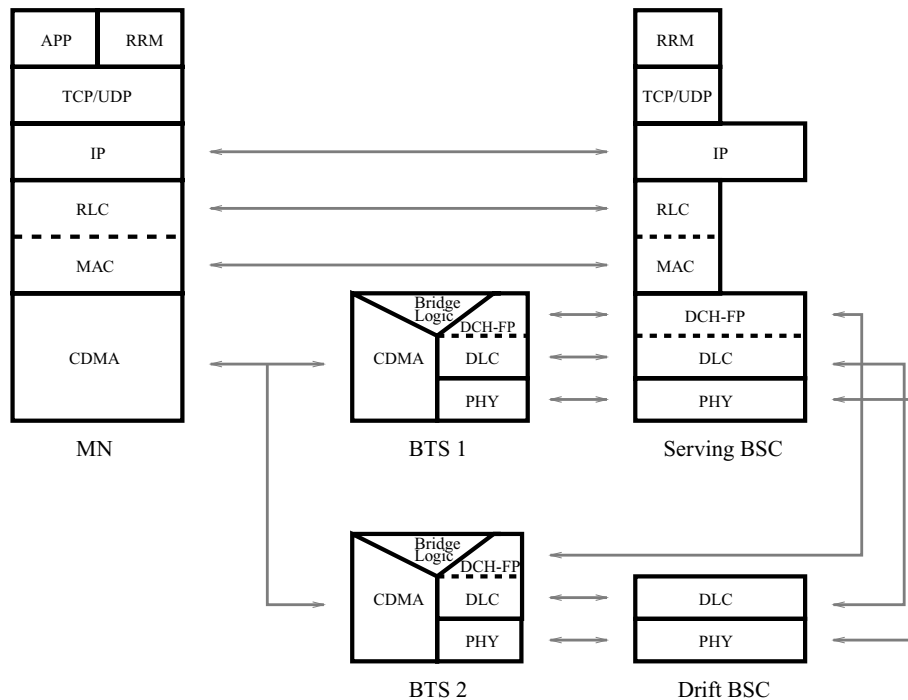


Figure 4. Protocol stack illustration for space diversity combining in soft handoff.

(intra-BSC) soft handoff is similar except that the drift BSC does not exist and the BTS 2 has a direct physical link to the serving BSC.

2.4.1. Layer 2 Procedures in Soft Handoff

As stated previously, the serving BSC for an MN is the BSC that performs the Layer 2 processing of the IP packets into/from the TBs. An MN connected to the MCIP system has one and only one serving BSC at any time. The *drift BSC* is a BSC, other than the serving BSC, that controls cells associated with the MN. In Figure 4, the drift BSC is the controlling BSC of BTS 2. During an inter-cluster soft handoff, the drift BSC participates in the macroscopic diversity combining/splitting by forwarding the DLC PDUs, each of which encapsulates a DCH-FP PDU containing a TB as its payload, between BTS 2 and the serving BSC. The drift BSC does not perform Layer 2 processing of the IP packets.

In the downlink direction, the RLC sublayer of the serving BSC segments and encapsulates the IP packets into RLC PDUs. One or more RLC PDUs are encapsulated into a TB at the MAC sublayer. Since both BTS 1 and BTS 2 are in the active set, the TB should be sent to both BTS 1 and BTS 2 simultaneously. The serving BSC knows the DLC sublayer addresses of both BTS 1 and BTS 2, because it tracks all the BTSs in the active set of the MN. The TB, its Transport Format Indicator (TFI) [16], and the transport channel ID (*tran. ID*) are passed down to the DCH-FP sublayer and encapsulated into a DCH-FP PDU at the serving BSC. The DCH-FP PDU is further encapsulated into two DLC PDUs each with a DLC sublayer destination address of either BTS 1 or BTS 2. The two DLC PDUs are sent to BTS 1 and BTS 2, respectively. At each of the two BTSs, the *bridge logic* receives a copy of the TB, associated TFI and the *tran. ID*. According to the *tran. ID*, the TB and TFI are passed down to the appropriate PDCH. After performing a series of physical layer processing at each BTS,

two CDMA radio signals are sent down to the MN from BTS 1 and BTS 2 simultaneously. At the MN, the two radio signals are combined by a *maximal ratio combining RAKE receiver* at the CDMA layer.

In the uplink direction, the radio signal transmitted from the MN is intercepted by both BTS 1 and BTS 2. Each received TB together with its CRC (cyclic redundancy check) result is encapsulated into a DCH-FP PDU at each BTS, and then routed to the DCH-FP sublayer at the serving BSC. The MAC sublayer at the serving BSC receives two copies of the same TB and CRC check result from two BTSs, and chooses the better one if they are different. The chosen TB and CRC check result are submitted to the RLC layer for further processing. This process is called *selection diversity*.

2.4.2. Synchronization Procedure in Soft Handoff

The WCDMA radio access network is an asynchronous system, that is, the BSC and BTSs within an RAN, and BSCs in different RANs are not exactly synchronized as those in cdma2000 networks [28]. A complete discussion of WCDMA system synchronization can be found in [29]. Here, we only discuss the synchronization procedure during soft handoff.

There is a need to adjust the transmission timing in soft handoff to allow coherent combining in the RAKE receiver at the MN, otherwise it would be difficult to combine the transmissions from different BTSs. The timing measurement is performed by the MN. Each BTS keeps a 12-bits BTS frame number (BFN) counter which is incremented by 1 every 10 ms. The BFN is broadcast in the primary common control physical channel (PCCPCH) of the cell. The timing difference between the current cell and the candidate cell is found from the BFNs broadcast in their PCCPCHs [29]. The measurement result is reported to the serving BSC before the new cell is added to the active set.

When in an active mode, the MN continuously searches for new BTSs on the current carrier frequency. During the search, the MN monitors the received signal levels from neighboring BTSs, compares them to a set of thresholds, and reports them accordingly back to the serving BSC. Based on this information, the serving BSC informs the MN to add or remove a BTS from its active set. From the cell-search procedure, the MN knows the frame offset of the PCCPCH of potential soft handoff candidate cells relative to the PCCPCH of the current BTS from the decoded BFNs [22, 30]. When a soft handoff is to take place, this offset (together with the frame offset between the downlink DPCH and the PCCPCH of the current BTS) is used to calculate the required frame offset between the downlink DPCH and the PCCPCH of the candidate BTS. This offset is chosen so that the frame offset between the downlink DPCHs originated from the old and new BTSs at the MN receiver is minimized. This calculated frame offset is reported by the MN to its serving BSC.

When a new BTS is to be added to the active set, the serving BSC of the MN commands it to adjust the downlink timing in steps of 256 chips based on the timing information received from the MN, such that the frame timing received at the MN will be within $T_0 \pm 148$ chips prior to the frame timing of the uplink DPCH/DPDCH at the MN [21], where T_0 is a constant defined to be 1024 chips. The 256-chips step restriction is used in order to preserve downlink orthogonality [31].

Note that the DPCHs from the two different BTSs in the active set use two different scrambling codes [31]. This requires additional RAKE fingers in the MN receiver [16, 30]. The serving BSC informs the MN the identification number (4 bits, [31]) of the scrambling code used by the new BTS before the MN receiver can combine the radio signals from the two downlink DPCHs.

In the uplink direction with selection diversity, the situation is much simpler: the same radio signal is received by the two BTSs. There is no need to adjust the frame timing.

3. MCIP Routing and Handoff Schemes

All MNs in an MCIP access network use the Gateway's IP address as their care-of address. Outside the Gateway domain, global mobility is handled by Mobile IPv6. Within the MCIP access network (i.e., inside the Gateway domain), micro-mobility management is integrated with the MCIP *routing*, *handoff* and *paging* schemes.

The same as traditional cellular systems, MNs attached to an MCIP access network are divided into two groups: *active* MNs and *idle* MNs. An MN stays in the active state when communicating with its correspondent node, and goes into the idle state after the communication is over. For an active MN, its exact location information is implicitly included in its *virtual path* from its serving BSC to the Gateway, and the virtual path is modified each time after the MN is handed off from one cluster to another. The proposed MCIP routing and handoff schemes ensure that IP packets destined to/originated from the MN arrive at the MN/Gateway in-order without loss and duplication. This is a desired characteristic for real-time applications, which does not exist in Cellular IP and HAWAII. Paging scheme is used by the network to track idle users' locations in the level of *paging area*. A paging area can be configured to an MCIP domain or part of it. When a call is intended for an idle user, the MN is waked up to the active state by a paging message broadcast in the paging area.

According to the different protocol layers at which signaling messages are processed, signaling messages are divided into two categories: *network layer signaling messages* and *application layer (RRM layer) signaling messages*. Network layer messages are different from applications layer messages in that a network layer message is an IPv6 packet with a *Hop-by-Hop options extension header* containing control information, therefore it is processed by each network node en route to the destination [32]. In the following, we make two assumptions: (1) For signaling messages transferred over the radio interface, the length of the IP packets is assumed to be 9 octets. Provided that the signaling RB is 7.2 kbps, a signaling message is conveyed over the air via a single 10 ms TB; (2) For signaling messages exchanged between the fixed part of the access network, the length of the IP packets is assumed to be less than 128 octets.

3.1. MCIP ROUTING

Unlike Cellular IP and HAWAII, which use *soft-state* paths to route IP packets from the Gateway to the MN actual location, the MCIP access network explicitly sets up a *virtual path* for each MN. A virtual path is created when an MN initiates a call or is paged by the network, and cleared after the call is completed.

Two types of routing algorithms run at each network node: *regular IP routing* and *MCIP routing*. User packets are routed by MCIP routing along the virtual paths. Signaling messages are routed by regular IP routing, with exceptions that *path-teardown* and *crossover-discovery* are routed by MCIP routing. Regular IP routing is necessary to set up, modify and clear the virtual paths. It is also used for delivering RRM layer signaling messages. When an intermediate node (BSC or the Gateway) receives an IP packet, it knows whether or not the packet is a signaling message by checking the *flow label* field.

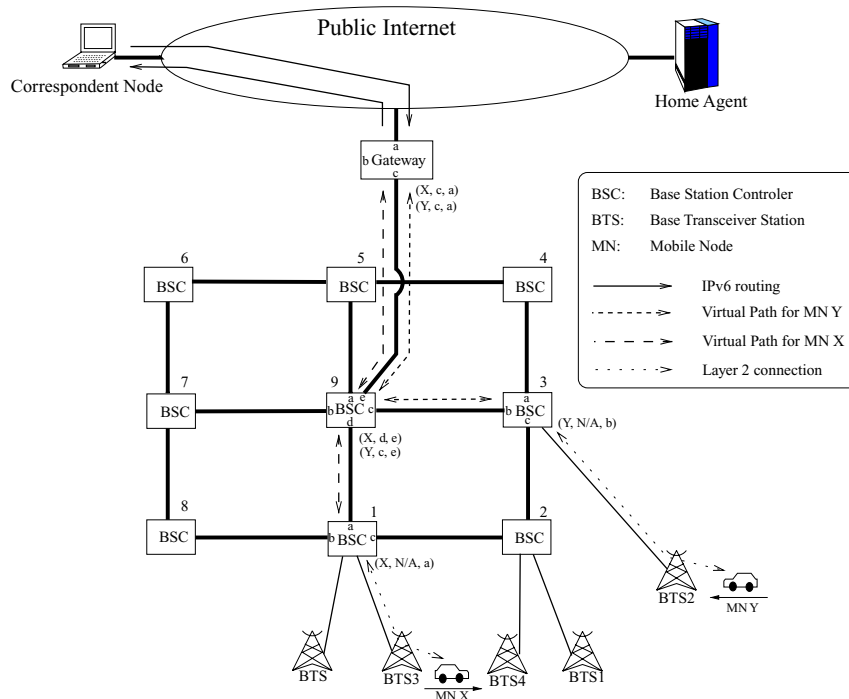


Figure 5. MCIP routing.

3.1.1. Virtual Path Setup

An MN knows the IP addresses of its serving BSC and the Gateway from the broadcasting channel after it is turned on or enters the Gateway domain. When a virtual path is needed, the MN generates a *path-setup* message with the Gateway as its destination. This message is first sent by the MN to its serving BSC over the air and then routed toward the Gateway by *regular IP routing*. This is an IPv6 packet with a *Hop-by-Hop Options extension header*. Thus it will be processed by each network node en route to the Gateway. The path taken by the path-setup packet is recorded by the intermediate nodes (BSCs) between the serving BSC and the Gateway. All user packets addressed to or originated from the MN will be routed on a hop-by-hop basis along the established virtual path thereafter. Viewing from the physical layer of the air interface, the *path-setup* message is delivered to the BTS over the Physical Random Access Channel [16, 18].

In the scenario illustrated in Figure 5, when MN X needs to set up a virtual path toward the Gateway router, a *path-setup* packet is generated and sent to BSC 1 over the radio interface. By regular IP routing the *path-setup* message is routed to BSC 9 through BSC 1's interface *a*. Thus BSC 1 keeps a triplet for MN X, $(X, N/A, a)$. The first item in the triplet, *X*, is the MN IP address. The second item, *N/A*, is the outgoing interface for packets addressed to MN X as well as incoming interface for packets sent by MN X. "N/A", which stands for "not available", is used here because BSC 1 is the *edge router* of MN X. The third item, *a*, is the incoming interface for packets addressed to MN X as well as outgoing interface for packets sent by MN X. Similarly, BSC 9 keeps a triplet for MN X (X, d, e) , and the Gateway keeps a triplet for MN X (X, c, a) . In this way, the virtual path between the MN X and the Gateway is kept in all the intermediate nodes (BSCs) and the Gateway router.

3.1.2. *Virtual Path Teardown*

A *path-teardown* packet has the same format as the *path-setup* packet, except the control information in the hop-by-hop options header is different. A *path-teardown* packet may be generated at any node (the MN, a BSC, or the Gateway) along the virtual path. But if the packet is originated from a BSC, and is to be sent to the Gateway, it uses the MN IP address as its source address so that the packet can be routed along the right virtual path². After receiving this packet, each network node clears the triplet in its cache associated with the MN.

It is possible that an MN loses connection to the network during a call. This happens due to a radio link disruption or when the MN moves out of the coverage. In this case, the serving BSC sends a *path-teardown* message toward the Gateway router to clear the virtual path. The *path-teardown* message is also used in path optimization process after an inter-cluster handoff as to be discussed. In the following discussion on handoffs, we assume that the MN is in the coverage of both the old BSC and the new BSC before the handoff procedure is initiated.

3.1.3. *An Example*

We describe the MCIP routing algorithm for a mobile to mobile communications case. Assume MN *X* has an IP packet to send to MN *Y*, and each of the MNs has established a virtual path to the Gateway as shown in Figure 5. MN *X* sends the packet over the air to BSC 1. BSC 1 compares the packet's destination address, *Y*, with all the mappings in its routing cache. Since no match is found, BSC 1 then compares the packet's source address, *X*, with the mappings. BSC 1 finds the triplet (*X*, N/A, *a*), and knows this packet is originated from MN *X* and should be sent to BSC 9 via interface *a* (upstream along the virtual path for MN *X*). After the packet arrives at BSC 9, BSC 9 finds a mapping for its destination address *Y*, (*Y*, *c*, *e*). Then BSC 9 knows this packet is destined to MN *Y*, and forwards it toward BSC 3 via interface *c* (downstream along the virtual path for MN *Y*). Thus, at the crossover node, BSC 9, the packet is diverted to MN *Y* instead of forwarded toward the Gateway. After BSC 3 receives the packet, it recognizes that the packet is destined for MN *Y*, because BSC 3 has a triplet (*Y*, N/A, *b*) which matches with the packet's destination address *Y*. Finally the packet is sent to MN *Y* over the air from BSC 3.

In the case that the MN communicates with a correspondent node (which may also be an MN, but outside the MCIP access network), packets destined to the MN are first routed to the Gateway via Mobile IPv6, and then forwarded to the MN along the virtual path inside the access network; packets originated from the MN are first delivered to the Gateway along its virtual path, and then sent to the correspondent node outside the access network via Mobile IPv6.

3.2. MCIP INTER-CLUSTER HANDOFF SCHEMES

Reliable and efficient handoff schemes can improve system performance significantly. Both hard handoff and soft handoff are supported in the MCIP access network. The decision on whether to use hard handoff or soft handoff is made by the system according to MN QoS requirements and network topology. According to the system elements involved during a handoff, there are two types of handoffs in the MCIP access network: inter-cluster handoff and intra-cluster handoff. Intra-cluster handoff is handled by the serving BSC itself, and no

² The same method is also used in the Cellular IP *indirect semi-soft handoff* after the candidate BS receives the *semi-handoff* message.

signaling message is transferred in the core network. In the following, we focus on the inter-cluster handoffs. Both the soft and hard handoff schemes adopt *mobile assisted* and *backward* handoff strategies. These are essential requirements for fast and seamless handoffs in 3G/IP interworking [27]. Handoff messages are application layer (RRM layer) messages in the MCIP access network. They are transferred in the network over UDP/IP protocols. IP packets carrying RRM messages have the same *flow label* value as *path-setup* and *path-teardown* messages, but they do not use extension headers.

At the target BSC, policy-based radio resource allocation algorithms may be performed if the requested radio resources are not available. Discussion of radio resource allocation algorithms is beyond the scope of this work. To simplify the presentation, we assume that the handoff request is simply queued at the target BSC if the requested resources cannot be allocated.

3.2.1. Soft Handoff

Handoff relies on the E_c/I_0 measurement performed over the physical pilot channels [16], where E_c is the received signal energy per chip of the coded pilot channel and I_0 is the power spectral density of interference plus noise within the channel bandwidth. The inter-cluster soft handoff consists of two phases: *add* and *drop*. To highlight the handoff process, a segment of the access network (Figure 5) is redrawn in Figure 6. MN X connects to the network via BTS 1, and is moving toward BTS 2. MN X periodically performs the E_c/I_0 measurement over the pilot channels from its surrounding BTSs, averages the measurement over an averaging window, compares the result to a set of thresholds, and reports back to its serving BSC, BSC 2 [16]. After MN X moves further into the overlapping area and some specific conditions are satisfied, the new BTS (BTS 2) is added to the *active set*; when MN X leaves the overlapping area, the old BTS (BTS 1) is dropped from the active set. Both the add and drop decisions are made by BSC 2.

During the add phase, the old BSC is still the serving BSC. It works as the edge router of the IP packets addressed to or originated from MN X. The IP packets destined to MN X are

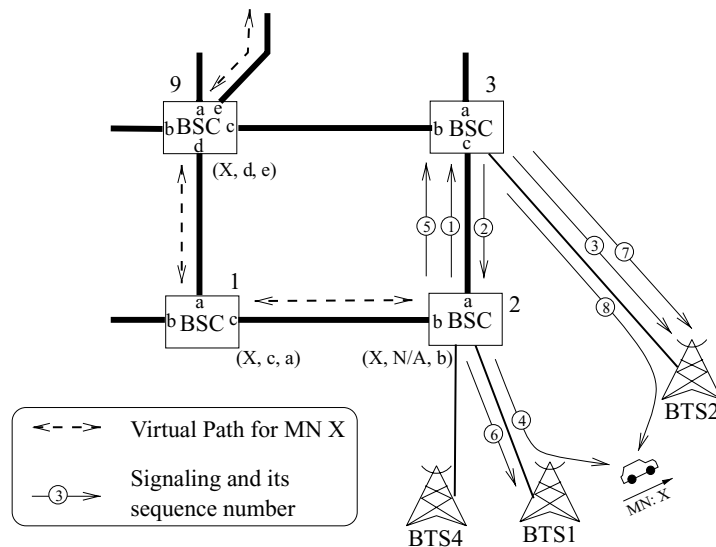
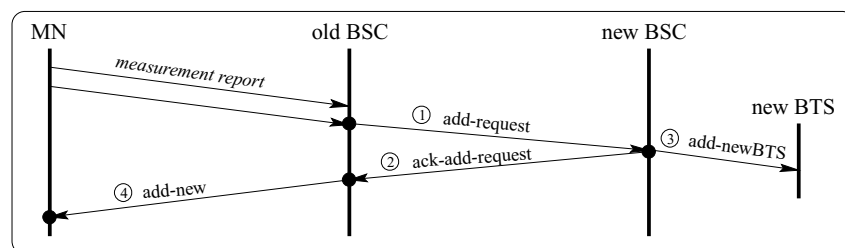
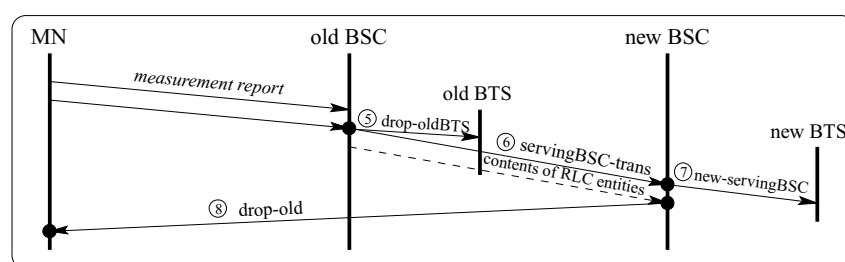


Figure 6. MCIP soft handoff process.



(a) Signalling in the add phase



(b) Signalling in the drop phase

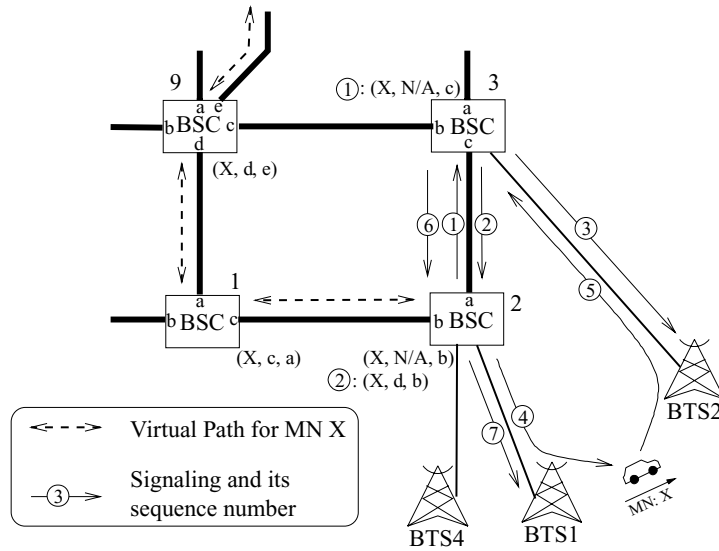
Figure 7. MCIP soft handoff signaling.

segmented into RLC PDUs in the RLC sublayer of the old BSC, formatted into TBs, and then sent to the MN via only the old BTS (before the new BTS is added to the active set) or via both the old and the new BTSs simultaneously (after the new BTS is added to and before the old BTS is dropped from the active set) over the air. The IP packets sent by MN X are assembled at the old BSC and sent to the Gateway along the virtual path. This proceeds until the old BSC transfers its serving BSC functions to the new BSC by sending a *servingBSC-trans* message which belongs to the drop phase signaling. As shown in Figure 7(a), the add phase has four RRM messages – (1) *add-request* from BSC 2 to BSC 3: It carries all parameters about the radio connection to MN X so that a second channel from BTS 3 to MN X can be established; (2) *ack-add-request* from BSC 3 to BSC 2: At BSC 3, upon receiving the request message, a radio resource allocation algorithm is performed. If available, the required radio resources at BSC 3 and BTS 2 are reserved for MN X . The downlink scrambling code ID assigned to the new channel and BSC 3's IP address are carried by this message. If the required resources are not available, the handoff request is queued by BSC 3 until resources are available or the communication session is forced to terminate; (3) *add-newBTS* from BSC 3 to BTS 2: It is sent if the radio resource allocation is successful. Upon receiving this message, BTS 2 sets up all parameters used by the new channel for MN X at the bridge logic, and is ready to participate in both the downlink and uplink space diversity procedures. The uplink diversity procedure starts at this moment; (4) *add-new* from BSC 2 to MN X : It informs the MN to add fingers to its receiver. After receiving the *ack-add-request* message, BSC 2 is ready to receive the second copies of the uplink TBs forwarded by BTS 2 via the DCH-FP sublayer and to execute the uplink macro diversity combining. The *add-new* message carries the ID number of BTS 2, the downlink scrambling code ID used by BTS 2, and BSC 3's IP address. After sending this message, BSC 2 forwards every downlink TB to both BTS 1 and BTS 2 via the DCH-FP sublayer.

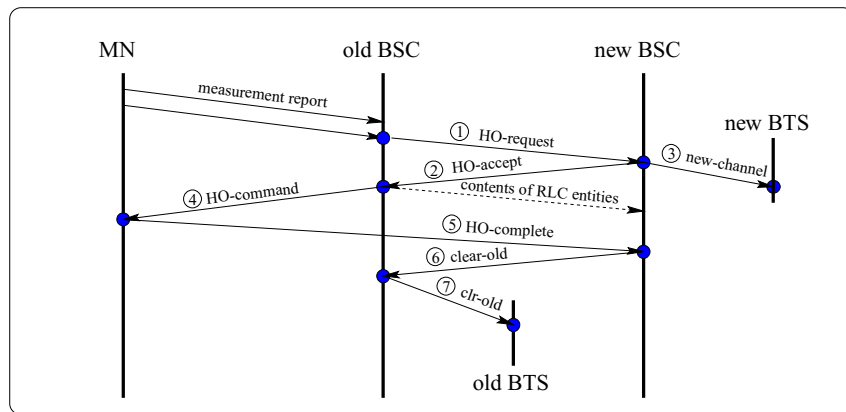
As shown in Figure 7(b), the drop phase signaling procedure is accomplished by four RRM messages – (1) *drop-oldBTS* from BSC 2 to BTS 1: It informs BTS 1 to release the old channel after the drop decision is made. BSC 2 also ends its function as the serving BSC, stops popping downlink TBs into the DCH-FP sublayer, withdraws resources of the old connection; (2) *servingBSC-trans* from BSC 2 to BSC 3: It informs BSC 3 to take over the function of serving BSC. Following it, BSC 2 encapsulates the contents and parameters of the corresponding RLC sublayer entities into an IP packet, and sends it to BSC 3. The IP packets awaiting for transmission in the IP layer are also forwarded to BSC 3; (3) *new-servingBSC* from BSC 3 to BTS 2: It informs BTS 2 to reconfigure the bridge logic so that the newly received TBs will be forwarded to BSC 3; (4) *drop-old* from BSC 3 to MN X: It informs the MN to drop BTS 1 from its active set. The ID number of BTS 1 is the only content of this message. MN X then removes from its receiver the fingers corresponding to the scrambling codes used by BTS 1. After the drop phase, the new BSC is the only BSC and the new BTS is the only BTS serving the MN X.

3.2.2. Hard Handoff

As illustrated in Figure 8, the handoff is accomplished by seven signaling messages – (1) *HO-request*: After the handoff decision is made, the old BSC sends an *HO-request* message to the new BSC. This packet carries MN X IP address, MN X address ID, CN IP address, CN address ID, the old and new BTS identification numbers, and the QoS requirements for the new radio channel, which are the same as those guaranteed by the old channel; (2) *HO-accept*: At the new BSC, upon receiving the HO-request message, a call admission control algorithm is performed on behalf of the new BTS (new cell). If the required radio resources are available, the handoff request is accepted, and the new BSC sends back to the old BSC an *HO-accept* message. MN X IP address, the old and new BTS ID numbers, and the ID numbers of the scrambling codes (both downlink and uplink) and the Transport Format (TF) Set used by the new channel are carried by the payload field of this message. If the required resources are not available, the handoff request is queued at the new BSC until resources are available or the communication session is forced to terminate; (3) *new-channel*: Immediately after sending the HO-accept message, the new BSC selects the appropriate RLC and MAC entities with appropriate parameters for the new channel, sets down the IP-address-vs-address-ID mappings used by the IP header compression algorithm, and sends a *new-channel* message to the new BTS commanding the new BTS to set up the requested new radio link (CDMA layer) to MN X. This message carries the radio resource related information, such as the ID numbers of the scrambling codes, the TF Set, and the fast-loop power control target used by the new channel. After these actions, the new BSC waits for the *HO-complete* message from the MN; (4) *HO-command*: Upon receiving the HO-accept message, the old BSC sends an *HO-command* packet to MN X over the old radio link to inform the MN that the new channel is ready. This message carries the ID numbers of the scrambling codes (both downlink and uplink) and the TF set used by the new channel, and the new BSC's IP address and address ID; (5) *HO-complete*: After receiving the HO-command message, MN X switches its transceiver to the new channel, and sends an *HO-complete*, which is an empty message, to the new BSC via the new BTS over the new radio channel; (6) *clear-old* and (7) *clr-old*: Upon receiving the HO-complete message, the new BSC sends a *clear-old* message to the old BSC. The old BSC then releases the RLC and MAC entities associated with the old channel, and sends a *clr-old* message to the old BTS. The old BTS, in turn, releases its resources used by the old channel. In this way, the old channel is completely removed. Note that these messages are Application Layer messages



(a) Signalling in the add phase



(b) Signalling in the drop phase

Figure 8. MCIP hard handoff.

(RRM messages), and are sent using the UDP protocol to a reserved port. They are routed by regular IP routing between network nodes.

We next discuss how the hard handoff process affects the user traffic. For presentation clarity, BTS 1, BTS 2, BSC 2, BSC 3 are referred to as old BTS, new BTS, old BSC, and new BSC, respectively. As discussed in Section 2.2, the user traffic and signaling traffic are carried by two separate RBs, each of which is approximately 15 kbps (smaller than 15 kbps due to RLC and MAC headers). In the downlink direction, the disruption is the time elapses since the old BSC starts sending HO-command message to the first RLC PDU (which encapsulates a segment of the HO-complete message) is correctly received by the new BSC. Before the old BSC sends the HO-command, it stops sending user traffic to the MN, and encapsulates its segmentation buffer, sending buffer, reassembling buffer, receiving buffer, and related RLC variables into an IP packet, and sends this IP packets to the new BSC. After the new BSC

receives this IP packet, the contents of the packet are released to the corresponding RLC entities in the new BSC. Then, the new BSC is ready to accept RLC PDUs from the MN. Normally, the new BSC and new BTS are ready earlier than the moment when MN receives the HO-command message, because the HO-command message is segmented and encapsulated into 3 RLC PDUs and transmitting them from the old BSC to the MN will take at least 30 ms. The first RLC PDU that the MN sends to the new BSC is a segment of the HO-complete message. After this PDU is correctly received by the new BSC, the new BSC's RLC entities know that they can start communication with the MN, because the MN is the only one who knows the scrambling code and spreading code used by the new channel. At this instant, the new BSC starts sending user traffic toward the MN. In the uplink direction, the user traffic disruption is the time elapses from the moment that the old BSC begins to send HO-command message to the moment that the entire message is received by the MN. This is shorter than the downlink disruption time.

3.2.3. Path Optimization

As shown in Figure 9(a), after MN X is handed off from BSC 1 to BSC 2, its virtual path is not the shortest one. A path optimization process is proposed to find the shortest path for the MN after it is handed off from one BSC (cluster) to another. If there are multiple shortest paths, it is desired to reuse the old path to the largest extent, that is, to maximize the *path reuse efficiency* defined as

$$\text{path reuse efficiency} = \frac{\text{the count of reused hops}}{\text{the hop count of the new path}}$$

There are two benefits to maximize the path reuse efficiency: (1) the disruption caused to the user traffic is minimized, because during the rerouting period the user traffic is disrupted; and (2) the signaling cost is minimized.

The crossover node refers to the BSC which is closest to the old BSC and where the old path from the old BSC to the Gateway intersects with one of the shortest paths from the new BSC to the Gateway. The handoffs can be classified into 3 cases in terms of the logical location of the crossover node, as shown in Figure 9. For cases (b) and (c), a path optimization process is not necessary because the new path is the optimum one. In case (b), where the crossover node is the old BSC, the new BSC checks its routing table, and finds out that one of its shortest path to the Gateway is via the old BSC. Therefore the old BSC will not initiate the path optimization process. In case (c), the new BSC knows that it is on the old path, that is, the new virtual path is a segment of the old one. The new BSC sends to the old BSC a *path-teardown* to clear the portion of the old path from the new BSC to the old BSC.

For the case in Figure 9(a), before the path optimization process, the *add-request* message during soft handoff (or *HO-request* during hard handoff) sent from the old BSC to the new BSC carries a parameter used by the path optimization scheme: the *hop count of the old path* from the old BSC to the Gateway. After receiving this message, the new BSC checks its routing table, and finds that the shortest path from it to the Gateway is shorter than the old path. A path optimization process is needed. In the *ack-add-request* message in soft handoff (or *HO-accept* in hard handoff), there is an *Optimization* flag, which is used to inform the old BSC to initiate the optimization process when asserted. The *ack-add-request* (or *HO-accept*) message also carries the hop count of the shortest path from the new BSC to the Gateway router. A path optimization process is always started from the old BSC. In soft handoff, it is initiated after the old BSC sends the *servingBSC-trans* message to the new BSC; in hard handoff, it is initiated

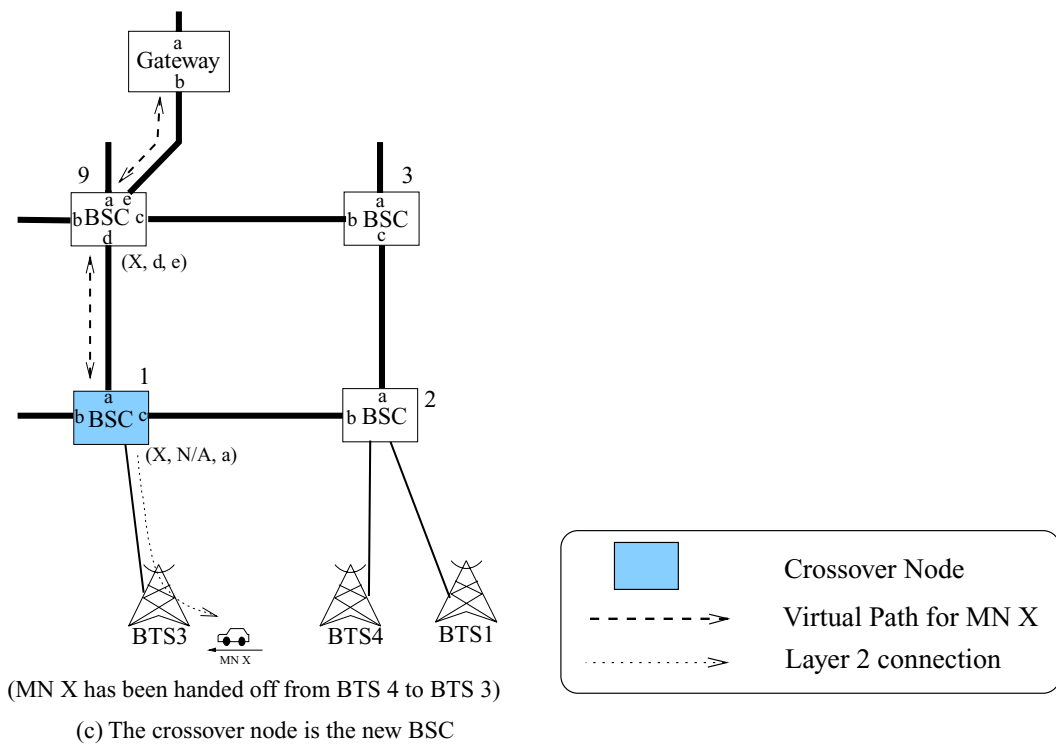
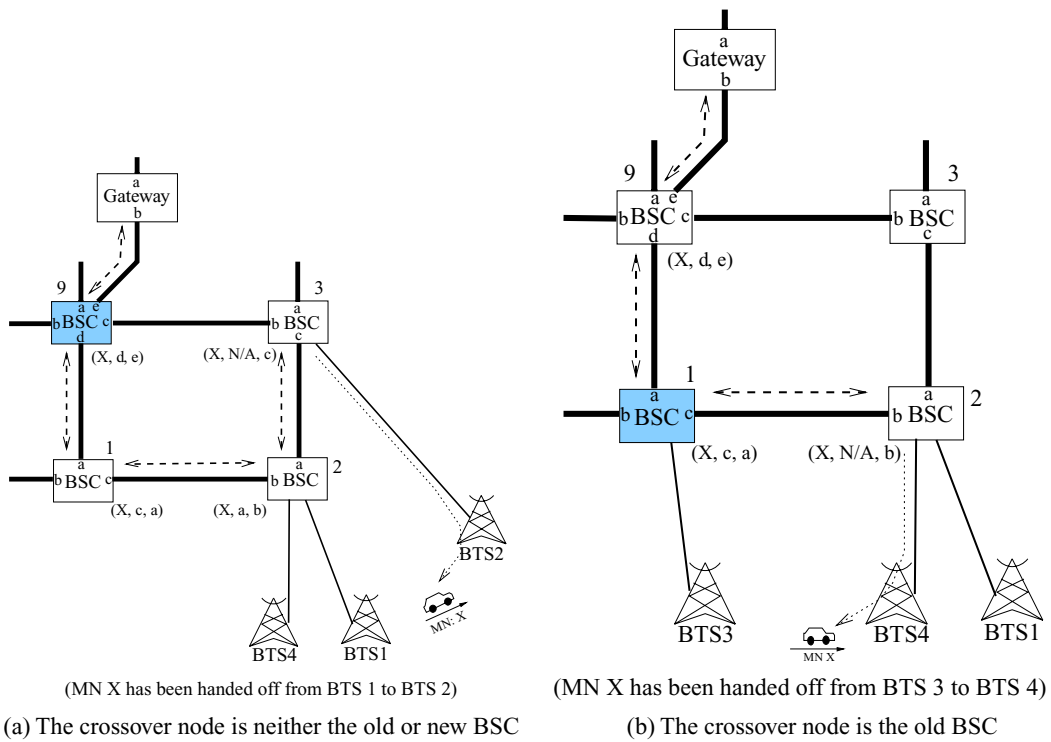


Figure 9. The 3 cases of the logical location of the crossover node.

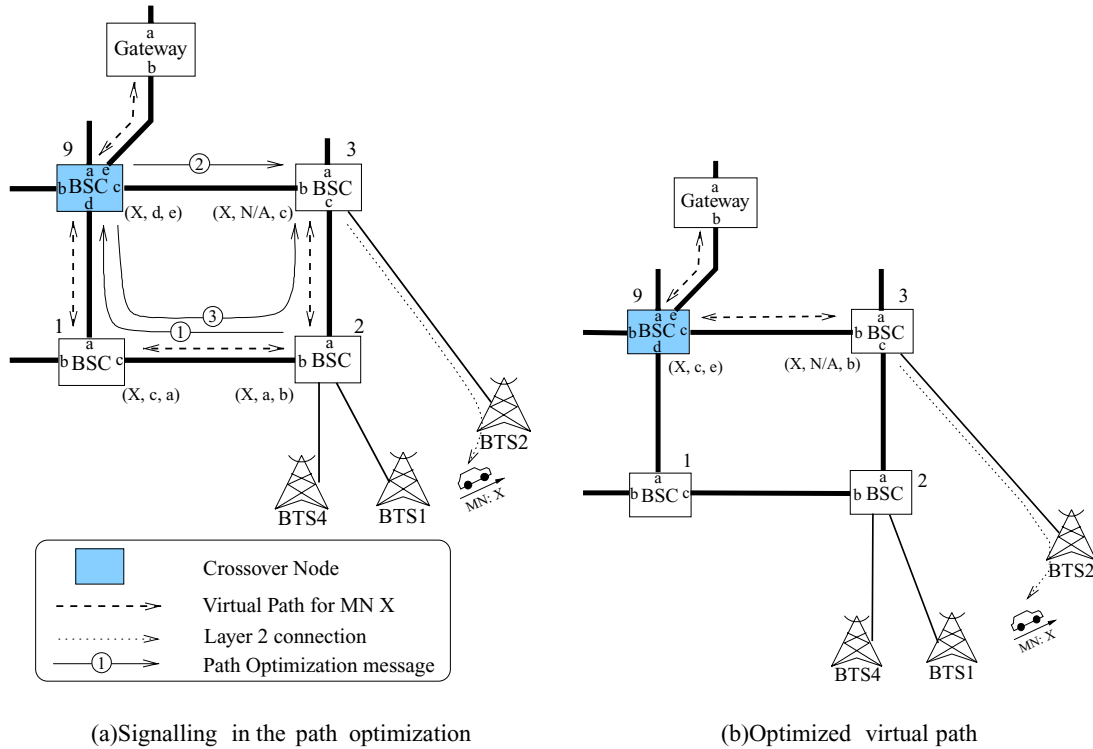


Figure 10. Path optimization when the crossover node is neither old or new BSC.

after the old BSC receives the *HO-accept* message. Note that the *Optimization* flag in the *HO-accept* message is not asserted for both cases (b) and (c).

The path optimization is accomplished by three signaling messages as shown in Figure 10(a): (1) *crossover-discovery*, (2) *partial-path-setup*, and (3) *path-teardown*. Note that, unlike the handoff signaling messages, path optimization messages are network layer signaling messages, and have the same *flow label* value as *path-setup* message. Each message carries a *hop-by-hop options header*, which bears the control information. Each path optimization message is to be processed by each intermediate node en route to its destination.

The old BSC generates a *crossover-discovery* packet with the IP address of the MN as its source address and the Gateway IP address as its destination address. The *crossover-discovery* packet is routed by MCIP routing toward the Gateway along the old path. This packet carries the hop count of the shortest path from the new BSC to the Gateway. At each node en route to the Gateway, the smallest hop count to the Gateway and the smallest hop count to the new BSC are found out in its routing table. If the sum of these two hop counts equals to the hop count carried by the message, the current node is the crossover node; otherwise it is not. If it is not the crossover node, the message is sent to next network node by MCIP routing. Finally the message arrives at the crossover node, which is BSC 9 in Figure 9(a).

The crossover node discards the *crossover-discovery* message, and generates the other two messages: *partial-path-setup*, and *path-teardown*. The *partial-path-setup* packet is routed by regular IP routing toward the new BSC, and a new partial path from the crossover node to the new BSC is created along the path taken by this message. The *path-teardown* message is

routed to the old BSC along the old path. Thus the old path is cleared, and *path-teardown* is forwarded to the new BSC by the old BSC to indicate the termination of the path optimization process. Note that the *path-teardown* packet normally arrives at the new BSC later than the *partial-path-setup*. During this time period, packets arrived along the new path should be queued at the new BSC until *path-teardown* is received to ensure that the packets are sent to the MN in-order. The new virtual path for case (a) is shown in Figure 10(b).

In hard handoff, the path optimization process is finished before the new BSC receives the *HO-complete* message from the MN. Thus no extra disruption is introduced to user traffic in the uplink direction.

To summarize, between the soft and hard handoff schemes, the soft handoff offers better link quality with the macro spatial diversity, reduces or eliminates the “Ping-Pong” effect, and needs a lower hysteresis margin in defining the active BTS set, resulting in a less handoff delay. However, the advantages are achieved at the cost of a more complex handoff procedure, and a likely increased forward link interference during soft handoff. As a result, soft handoff is preferred for real-time services and for better QoS provisioning.

4. Performance Analysis

4.1. SYSTEM CONFIGURATIONS

To compare the performance of MCIP handoff schemes with those used in Cellular IP and HAWAII, we first describe the system configurations. The network topologies used by the MCIP, Cellular IP and HAWAII access networks are shown in Figures 1, 11 and 12, respectively, and their cluster layouts are shown in Figure 13. The configuration parameters of these access networks are listed in Table 1. A *cluster* in a Cellular IP or HAWAII domain is defined to be the

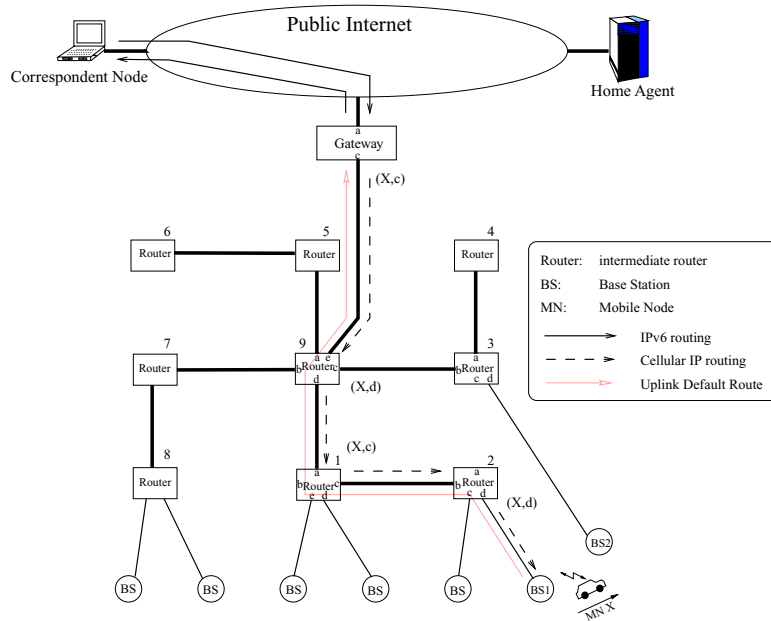


Figure 11. Cellular IP access network.

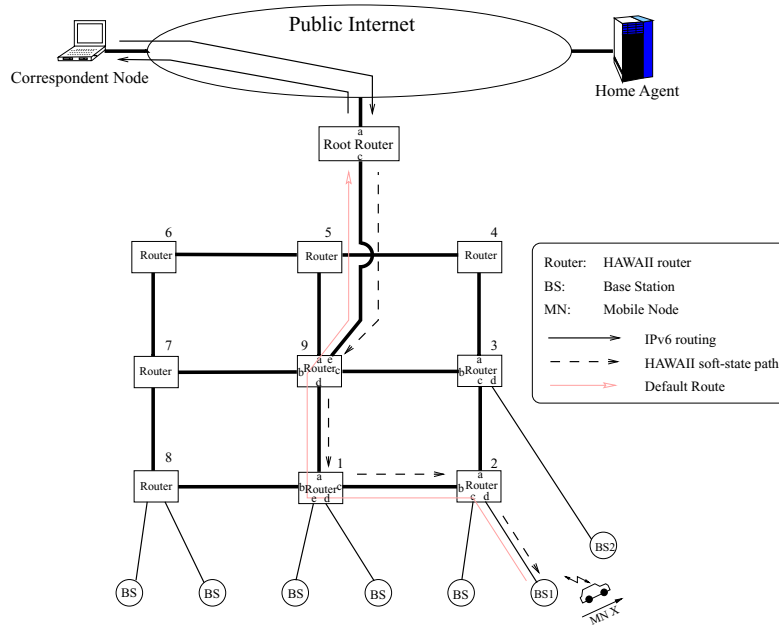


Figure 12. HAWAII access network.

group of cells, whose base stations are physically connected to the same intermediate router. A similar configuration is used in [12] with $n_1 = 140$, $n_2 = 7$ and $n_3 = 20$. We modify these parameters for the square coverage area of a cell, a cluster, or a domain. The *fluid flow mobility model* [33], which is also used in [12, 34], assumes that MNs are moving at an average velocity of v km/hr, and their direction of movement is uniformly distributed over $[0, 2\pi]$. Assuming that MNs are uniformly populated with a density of ρ users per km^2 in an area with boundary length L km, the rate of boundary crossing in s^{-1} , R , is given by

$$R = \frac{\rho v L}{3600 \pi} = 16.4 (\text{s}^{-1}).$$

Since each cluster is a square, the border crossing rate from (or to) one of its adjacent cluster is $R/4$.

4.2. SIGNALING COST FOR MOBILITY MANAGEMENT

The mobility management signaling cost refers to the total signaling cost for intra-cluster (inter-cell) handoffs, inter-cluster handoffs, and inter-domain migrations. Inter-domain mobility is handled by Mobile IPv6. With MCIP, when an MN migrates across the cluster boundary from one cluster to another in the same Gateway domain, 8 signaling messages can be used to complete a soft handoff, or 7 signaling messages to accomplish a hard handoff. We notice that, in both hard handoff and soft handoff, 3 messages are exchanged between the old BSC and the new BSC. Therefore, in terms of signaling cost to the core network, inter-cluster hard handoff and soft handoff have the same performance. For intra-cluster handoff, there is no signaling message exchange between BSCs because the handoff is handled by the serving BSC itself locally.

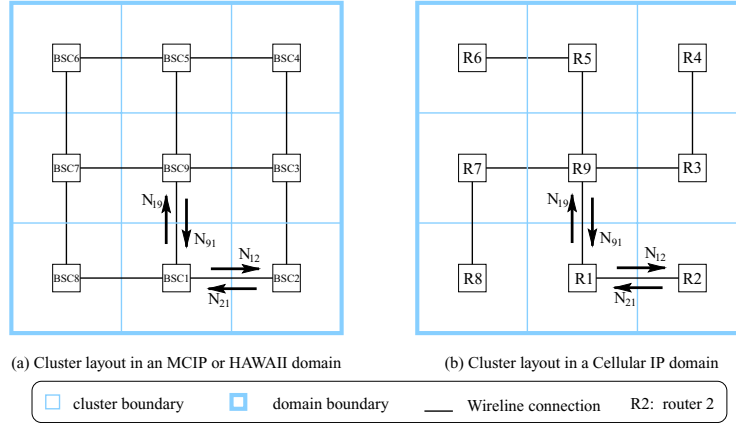


Figure 13. Cluster layout.

Table 1. Domain configuration values

Symbol	Definition	Value
n_1	BTSs (or BSs) per domain	144
n_2	BSCs (or second tier Routers) per domain	9
n_3	BTSs (or BSs) per cluster	16
L_c	Perimeter of a cell	10.6 km [12]
ρ	User density (active user)	39 per km ² [12]
v	Average user speed	112 km/hr [12]
L_R	Perimeter of a cluster [$4L_c$]	42.4 km
L_D	Perimeter of a domain [$3L_R$]	127.2 km
A	Coverage area of a domain [$(\frac{L_D}{4})^2$]	1,011.24 km ²
N	Number of active users in a domain [$A\rho$]	39,438

In the MCIP and Cellular IP access networks, MNs use the Gateway IP address as their *care-of address*, while in HAWAII an MN has to obtain a *care-of address* from the network. To simplify the calculation, the *care-of address* acquisition process in a HAWAII network is ignored. Note that in an MCIP domain an MN keeps the IP layer connection with its serving BSC, but in a Cellular IP or HAWAII domain an MN keeps the IP layer connection with its serving BS. To make a fair and meaningful comparison, we only compare their signaling cost at the second tier, that is, signaling cost on the links between BSCs in an MCIP domain and on the links between intermediate routers in a Cellular IP or HAWAII access network. Also, the metric we use in the comparison is *number of hops per second*, but not *number of signaling messages per second*. This is because a signaling message which travels several network nodes consumes more core network resources than a message which is delivered from a node to its next hop neighbor. Assume that, in an MCIP domain, the Gateway router is co-located with BSC 9 (Figure 1 and 5); in Cellular IP, the Gateway co-locates with Router 9 (Figure 11); in HAWAII, the domain root router is co-located with intermediate Router 9 (Figure 12). We also assume that, when an MN moves from one cell to another, its handoff request (or access request when the MN moves from outside the Gateway domain into the Gateway domain) is always accepted.

Table 2. Comparison of average mobility signaling cost

Number of messages	MCIP (soft)	MCIP (hard)	MCIP (no path opt)	HAWAII (inter-cluster)	Cellular IP (inter-cluster)	Cellular IP (intra & inter-cluster)
N_{19} (hops/second)	41.0	41.0	36.9	20.5	53.2	196.6
N_{91} (hops/second)	20.5	20.5	12.3	8.2	24.6	98.2
N_{12} (hops/second)	14.4	14.4	12.3	8.2	8.2	49.2
N_{21} (hops/second)	24.6	24.6	20.5	12.3	24.6	98.2
N_{HO} (hops/second)	558.0	558.0	459.2	278.8	442.4	1,768.8

Table 2 summarizes the average mobility signaling cost for the schemes, where N_{ij} denotes the mean number of signaling messages of various types sent from BSC_{*i*} to BSC_{*j*} (or from Router *i* to Router *j*) per second, and N_{HO} gives the total number of signaling cost in each access systems under comparison. The calculation of N_{19} , N_{91} , N_{12} and N_{21} is straightforward, and the total signaling cost in MCIP, Cellular IP and HAWAII systems can be calculated according to the following equations [14]:

$$\begin{aligned}
 N_{MCIP} &= 4(N_{19} + N_{91}) + 8(N_{12} + N_{21}) \\
 N_{CellularIP} &= 8(N_{19} + N_{91}) + N_{12} + N_{21} \\
 N_{HAWAII} &= 4(N_{19} + N_{91}) + 8(N_{12} + N_{21}).
 \end{aligned}$$

As seen from Table 2, in MCIP, the signaling is not symmetric on the two directions of a link, i.e., $N_{19} \gg N_{91}$ and $N_{21} \gg N_{12}$. This is mainly because of the effect of the inter-domain roaming (note that the Gateway router is co-located with BSC 9). When an MN moves into the MCIP domain under study, a *path-setup* message has to be sent from the serving BSC to the Gateway; when an MN moves out of the domain, a *path-teardown* message is sent to the Gateway from the serving BSC. Thus, the number of signaling messages from the surrounding BSCs toward BSC 9 are more than that in the reverse direction.

The signaling cost in MCIP is approximately twice as much as that in HAWAII, due to the following 3 reasons: First, HAWAII paths are soft-state. For inter-domain handoffs, when an MN moves out of the domain under consideration, it does not signal the intermediate nodes which cache entries for its soft-state path; it simply leaves the soft-state path over there for time-out. By using explicit signaling messages to set up and tear down a virtual path for each MN, MCIP protocol reduces the number of mappings cached by each BSC. This in turn reduces the route lookup time and improves the network scalability, which is not shown in Table 2. Second, in HAWAII there is no path optimization process. However, the path optimization is necessary (with the cost of 6 hops for each optimization process). Without it, all the user packets may have to travel 2 extra hops, which degrades the network resource utilization efficiency and causes a longer delay to the user traffic. Third, both MCIP soft and hard handoff schemes are mobile-assisted backward handoffs, which requires 3 signaling message exchanges between the old and new BSCs, while HAWAII MSF handoff scheme is a mobile-controlled forward handoff, which only needs 2 signaling message exchanges between the old and new BSs. The handoff scheme may work well in random access radio networks, but it is not suitable for 3G/IP interworking. Consider that an MN is connected to the network via a dedicated channel, which is the most common case for real-time services. When the MN moves into a new cell, it has to send the Mobile IP registration message (used as the handoff request in HAWAII) to the

new BS via the Random Access Channel (RACH) [16]. This message is likely to experience a long transmission delay due to a low transmission rate over RACH and an access delay due to collision(s). After the new BS performs a call admission control procedure, the MN is assigned a dedicated channel if the requested radio resources are available, but the assigned downlink scrambling code ID has to be sent to the MN via a downlink shared channel with a low bit rate. Another factor that contributes to the long handoff delay is that the new BS has no knowledge of the context used by the header compression algorithm at the old BS. Thus the first IP packet to be transmitted over the air of each traffic flow has to contain the full header [14]. The handoff delay could be tens of milli-seconds or even more.

Cellular IP is the least efficient in terms of the signaling cost due to handoffs. Both intra-cluster and inter-cluster handoffs cause signaling messages in the second tier, and every signaling message has to be sent from its origin to the Gateway first and then from there forwarded to its destination. Cellular IP handoff also suffers from a long handoff delay due to the RACH and IP compression problems as in HAWAII handoffs. Moreover, the old BS cannot release the radio resources used by the old channel to the MN quickly after the handoff, because there is no feedback mechanism from the MN or the new BS to inform the old BS that the MN has switched its transceiver to the new BS. This degrades the frequency spectrum efficiency.

4.3. SOFT HANDOFF LATENCY

The MCIP soft handoff scheme adopts the mobile assisted backward handoff strategy. The measurement is performed by the MN, while the add or drop decision is made by the serving BSC. Therefore, there is a latency from the moment that a BTS should be added to or dropped from the active set to the moment that the BTS is actually added or dropped by the MN. The handoff latency is analyzed under the following assumptions: (1) Packet data is transmitted in an error-free environment; (2) The propagation delay is ignored, i.e., only the transmission delay and queuing delay are considered; (3) There is only one significant radio path from each of the BTSs; (4) Each signaling message, or measurement report is encapsulated in a single radio frame. In the case that the signaling RB is 7.2 kbps, excluding the 2-octets compressed version of the upper layer header overhead, there are 7 octets that can be actually used to convey the status and control information in each message; (5) The message processing time is zero. Upon reception of a signaling message, a network node acts on the message immediately with zero response time.

The handoff latency in the add phase is called the *add phase completion time*. It refers to the time interval from the moment that the pilot channel signal from the new BTS becomes strong enough for the new BTS to be added to the active set to the moment that the MN actually receives the *add-new* command from the serving BSC. Similarly, the handoff delay in the drop phase is called the *drop phase completion time*. It is the time duration from the moment that the condition to drop the old BTS from MN's active BTS set is satisfied to the moment that the MN receives the *drop-old* command from the new BSC. Four mechanisms are employed in the MCIP access network to minimize the handoff delays in both the add and drop phases: DPCH frame timing scheduling, DCH-FP PDU prioritization, IP packet segmentation, and TB preemption [14].

It can be shown that both the add phase completion time and drop phase completion time are less than 40 ms [14]. This is far less than the handoff delays in GSM systems. Also, because we assume a separate signaling RB is used for signaling purpose, signaling messages do not

cause any additional transmission delay to user traffic, which normally can tolerate at most 200 ms end-to-end transmission delay [35].

The MCIP handoff schemes are much faster than those used in Cellular IP and HAWAII access networks mainly because, in Cellular IP and HAWAII, an MN has to set up a connection with the target BTS through a random access channel during handoff [14]. In addition, for real-time traffic, a significant advantage of MCIP over HAWAII and Cellular IP is that user packets are delivered within the Gateway domain in sequence without loss and duplication, which is critical for real-time multimedia services. In HAWAII or Cellular IP, in-sequence delivery is not guaranteed and loss and duplication are likely to happen during handoffs.

4.4. SCALABILITY

The network scalability refers to the capability of the radio access network to serve a large population of simultaneous users in a large region (a large network size). The coverage area of a domain (either an MCIP, a Cellular IP or a HAWAII domain) as in Table 1 is quite large: $A = 1,011.24 \text{ km}^2$. The scalability of a Cellular IP domain and a HAWAII domain has been demonstrated in [10] and [12], respectively. An MCIP domain has a similar scalability. The bottleneck of the network scalability is the Gateway router, because every IP packet has to pass the Gateway router. As in a HAWAII domain, the number of mappings at the Gateway router in an MCIP domain (i.e., the number of active users in the domain) is $\rho A = 39,438$. It is well within the capability of modern routers [12]. Furthermore, a majority of these MNs are completely specified in a particular domain/subnet, that is, they are within their home domains. In this case, perfect hashing is possible resulting in significantly reduced memory access for IP route lookup. Thus, route lookup for data forwarding can be done efficiently at the Gateway router [12].

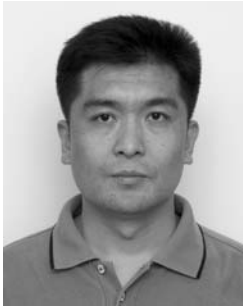
5. Conclusion

This paper proposes a novel 3-tier 3G wireless/IP interworking system model, the so-called MCIP access network, based on the characteristics of the 3G radio interface. Within the framework of the system model, an MCIP routing algorithm, an inter-cluster soft handoff scheme, and an inter-cluster hard handoff scheme are presented. Both MCIP hard and soft handoff schemes enable IP packets to be delivered within the MCIP access network in-order without loss and duplication. The MCIP soft and hard handoff schemes are compared with the Cellular IP and HAWAII protocols. It is shown that, in addition to their distinct QoS provisioning capability, MCIP handoff schemes are more efficient in terms of signaling cost and faster in terms of handoff delays. The MCIP advantages of supporting soft handoffs and QoS provisioning for real-time services are achieved at slightly increased system complexity.

References

1. L. Bos and S. Leroy, "Toward an all-IP-Based UMTS System Architecture", *IEEE Network*, vol. 15, pp. 36–45, Jan.–Feb. 2001.
2. G. Patel and S. Dennett, "The 3GPP and 3GPP2 Movements Toward an all-IP Mobile Network", *IEEE Personal Communications*, Vol. 7, pp. 62–64, Aug. 2000.
3. J. Yang and I. Kriaras, "Migration to all-IP Based UMTS Networks", *First International Conference on 3G Mobile Communication Technologies*, No. 471, pp. 19–23, 2000.

4. P. J. McCann and T. Hiller, "An Internet Infrastructure for Cellular CDMA Networks Using Mobile IP", *IEEE Personal Communications*, Vol. 7, pp. 26–32, Aug. 2000.
5. J.H. Lee, T.H. Jung, S.U. Yoon, S.K. Youm, and C.H. Kang, "An Adaptive Resource Allocation Mechanism Including fast and Reliable Handoff in IP-based 3Gwireless Networks", *IEEE personal communications*, vol. 7, pp. 42–47, Dec. 2000.
6. D. Wong and T. J. Lim, "Soft Handoffs in CDMA Mobile Systems", *IEEE Personal Communications*, vol. 4, pp. 6–17, Dec. 1997.
7. C. C. Lee and R. Steele, "Effect of Soft and Softer Handoffs on CDMA System Capacity", *IEEE Vehicular Technology Conference, 1999. VTC 1999 – Fall. IEEE VTS 50th*, vol. 47, pp. 830–841, Aug. 1998.
8. N. Binucci, K. Hiltunen, and M. Caselli, "Soft Handover Gain in WCDMA", *IEEE Vehicular Technology Conference (VTC) Fall 2000*, vol. 3, pp. 1467–1472, 2000.
9. C. Perkins and D. Johnson, "Mobility Support in IPv6", *Proc. ACM/IEEE Mobicom' 96*, pp. 27–37, Nov. 1996.
10. A.T. Campbell, J. Gomez, S. Kim, A.G. Valko, Chieh-Yih Wan, and Z.R. Turanyi, "Design, Implementation, and Evaluation of Cellular IP", *IEEE personal communications*, vol. 7, pp. 42–49, Aug. 2000.
11. Z.D. Shelby, D. Gatzounas, A. Campbell and C. Wan, "Cellular IPv6", Internet Draft, Draft-shelby-seamoby-cellularipv6-00.txt, Nov. 2000.
12. R. Ramjee, T.L. Porta, S. Thuel, K. Varadhan, and S.Y. Wang, "HAWAII: A Domain-based Approach for Supporting Mobility in Wide-area wireless Networks", *Seventh International Conference on Network Protocols*, pp. 283–292, 1999.
13. E. Gustafsson, A. Jonsson, and C. Perkins, "Mobile IP Regional Registration", Internet Draft, draft-ietf-mobileip-reg-tunnel-04.txt, Mar. 2000.
14. X. Liu, "Inter-cluster soft handoff in 3g/ip Interworking", Master's thesis, University of Waterloo, Waterloo, Ontario, 2002.
15. T.S. Rappaport, *WIRELESS COMMUNICATIONS: Principles and Practice*. Prentice Hall, 1996.
16. H. Homa and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. John Wiley & Sons, 2000.
17. J. Kempf, P. McCann, and P. Roberts, "IP Mobility and the CDMA Radio Access Network: Applicability Statement for Soft Handoff", Internet Draft, Draft-kempf-cdma-appl-02.txt, Sept. 2001.
18. 3GPP TS 25.211 v3.3.0 (2000-06), "Physical Channels and Mapping of Transport Channels onto Physical Channels (FDD) (Release 1999)"
19. 3GPP TS 25.212 v3.3.0 (2000-06), "Multiplexing and Channel Coding (FDD) (Release 1999)."
20. 3GPP TS 25.213 v3.3.0 (2000-06), "Spreading and Modulation (FDD) (Release 1999)."
21. 3GPP TS 25.214 v3.3.0 (2000-06), "Physical Layer Procedures (FDD) (Release 1999)."
22. 3GPP TS 25.215 v3.3.0 (2000-06), "Physical Layer - Measurements (FDD) (Release 1999)."
23. 3GPP TS 25.321 v3.8.0 (2001-06), "MAC Protocol Specification (Release 1999)."
24. 3GPP TS 25.322 v3.7.0 (2001-06), "RLC Protocol Specification (Release 1999)."
25. 3GPP TS 25.922 v4.0.0 (2001-03), "Radio Resource Management Strategies (Release 4)."
26. E. Dahlman, P. Beming, J. Knutsson, F. Ovesjo, M. Persson, and C. Roobol, "WCDMA - the Radio Interface for Future Mobile Multimedia Communications", *IEEE Transaction on Vehicular Technology*, 1998., Vol. 47, pp. 1105–1118, Nov. 1998.
27. A. H. Aghvami and P. Smyth, "Forward or Backward Handover for W-CDMA?", *First International Conference on 3G Mobile Communication Technologies*, No. 471, pp. 235–239, 2000.
28. 3GPP2 IS-2002.2, "Physical Layer Standard for cdma2000 Spread Spectrum Systems."
29. 3GPP TS 25.402 v4.3.0 (2001-12), "Synchronization in UTRAN Stage 2 (Release 4)."
30. ETSI/SMG2, "The ETSI UMTS Terrestrial Radio Access (UTRA) ITU-R RTT Candidate Submission", May/June 1998.
31. 3GPP TS 25.331 v5.2.0 (2002–09), "Radio Resource Control (RRC) Protocol Specification (Release 5)."
32. W. Stallings, "IPv6: The New Internet protocol", *IEEE Communications Magazine*, Vol. 34, pp. 96–108, July 1996.
33. R. Thomas, H. Gilbert, and G. Mazziotto, "Influence of the Mobile Station on the Performance of a Radio Mobile Cellular Network", in *Proc. 3rd Nordic Seminar, Paper 9.4*, Copenhagen, Denmark, 1988.
34. S. Mohan and R. Jain, "Two User Location Strategies for Personal Communications Services", *IEEE Personal Communications*, Vol. 1, pp. 42–50, 1st Qtr 1994.
35. S.N. Mukhi, "Design and Performance Analysis of a Feedback-based Handoff Scheme", Master's Thesis, University of Waterloo, Waterloo, Ontario, 1998.



Xin Liu received his B.E. and M.E. degrees in radio engineering from Harbin Institute of Technology (China), in 1990 and 1993, respectively, and his M.A.Sc degree in electrical engineering from the University of Waterloo in 2002. He joined Research In Motion in 2002 as a firmware developer. His current work involves GSM/GPRS and WCDMA firmware development.



Weihua Zhuang received the B.Sc. and M.Sc. degrees from Dalian Maritime University (China) and the Ph.D. degree from the University of New Brunswick (Canada), all in electrical engineering. Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, where she is a full professor. She is a co-author of the textbook *Wireless Communications and Networking* (Prentice Hall, 2003). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning. Dr. Zhuang received the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is an Editor of *IEEE Transactions on Wireless Communications*, an Associate Editor of *IEEE Transactions on Vehicular Technology*, and an Editor of *EURASIP Journal on Wireless Communications and Networking*.