Chapter 1

FAST SOFT HANDOFF SUPPORT AND DIFFSERV RESOURCE ALLOCATION IN WIRELESS MOBILE INTERNET

Yu Cheng, Xin Liu, and Weihua Zhuang
Centre for Wireless Communications
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
{ycheng, xinliu, wzhuang}@bbcr.uwaterloo.ca

Abstract The next-generation wideband CDMA (code-division multiple access) wireless networks are evolving toward a versatile IP (Internet Protocol) based network that can provide various real-time multimedia services to mobile users. Two major challenges in establishing such a mobile wireless Internet are support of fast soft handoff and provision of quality of service (QoS) over IP-based wireless access networks. Various domain-based schemes have been proposed to provide fast micromobility support, but limited only to hard handoff. In this chapter, we propose a new domain-based scheme, the Mobile Cellular IP (MCIP) scheme, for fast soft and hard handoffs. To provision QoS, the differentiated services (DiffServ) mechanism can be integrated naturally with the domain-based architecture. Such a DiffServ resource allocation scheme is proposed to guarantee the handoff call dropping probability by domain-based admission control. We also propose an adaptive assured service for the stream class of traffic, where bandwidth allocated to the traffic is adjusted according to the network condition. When congestion happens, the resources obtained by bandwidth reduction of the on-going calls are utilized to minimize handoff call dropping and new call blocking probabilities.

Keywords: Mobile wireless Internet, soft handoff, quality of service, differentiated services, resource allocation

1. Introduction

Provision of various real-time multimedia services to mobile users is the main objective of the next-generation wireless networks, which will implement Internet Protocol (IP) in the network layer and can interwork with the Internet backbone seamlessly [9, 28]. On the radio interface, the wideband code-division multiple access (CDMA) techniques are used, aiming to provide mobile users a reliable, high-speed, wireless Internet connection. The establishment of such wireless mobile Internet is technically very challenging. Two major tasks are the support of fast soft handoff and the provision of quality-of-service (QoS) guarantee over IP-based wireless access networks.

The next-generation, 3rd generation (3G) or 4th generation (4G), wireless networks will adopt micro/picocellular architectures for various advantages including higher data throughput, greater frequency reuse, and location information with finer granularity [30]. In this environment, the handoff rate grows rapidly and fast handoff support is essential. Especially for real-time traffic, the handoff call processing should be fast enough to avoid high loss of delay sensitive packets. A unique feature of CDMA is the use of soft handoff [22], which can effectively increase the capacity, reliability, and coverage range of the wireless system. Therefore, the implementation of fast soft handoff is critical for the efficient delivery of real-time services to mobile users.

To achieve fast handoff requires both a fast location/mobility update scheme and a fast resource allocation scheme. The popular scheme for fast location update is a *registration-domain-based* architecture, where the radio cells or the related base stations (BSs) in a geographic area are organized into a registration domain, and the domain connects to the Internet through a gateway or a domain root router. In such an environment, Mobile IP [38] is used to support *macromobility*, the inter-domain mobility. When a mobile node (MN)[1] moves into a registration domain for the first time, it will register the new care-of-address to its home agent (HA). While MNs migrate within the domain, various *micromobility* protocols [14, 41, 23, 19] have been proposed to complement Mobile IP by offering fast and seamless local handoff control. When MNs move across multiple BSs or multiple subnetworks within a domain, mobility updates messages will only be sent at farthest to the gateway, without interaction with the Mobile IP enabled Internet. In this way, the registration between the MNs and the home agent (which often locates far away) is eliminated, leading to an obviously reduced signaling overhead for location update, and a considerably less delay and packet loss during handoff. However, the previously proposed schemes focus

on enabling hard handoff and do not explicitly address IP soft handoff issues. So far, to our best knowledge, no protocol has been developed to support fast soft handoff in IP-based wireless networks. In this chapter, we propose a new domain-based architecture, called Mobile Cellular IP (MCIP), which not only inherits the ability of supporting fast, seamless, local hard handoff, but also integrates newly designed mechanisms to support fast soft handoff.

A handoff will be successful only when the new BS has enough resources to support the traffic. Therefore, fast handoffs require fast resource allocation. For this purpose, we establish a resource allocation scheme over a registration domain to achieve fast call admission control, QoS guarantee, and high utilization of the scarce wireless frequency spectrum.

The integrated services (IntServ) approach [10] and the differentiated services (DiffServ) approach [8] are the two main architectures for QoS provisioning in IP networks. The IntServ approach uses the Resource Reservation Protocol (RSVP) [11] to explicitly signal and dynamically allocate resources at each intermediate node along the path for each traffic flow. In this model, every change in an MN attachment point requires new RSVP signaling to reserve resources along the new path, which incurs latency in the call admission control and is not suitable for fast handoff. Also, the heavy signaling overhead reduces the utilization efficiency of the wireless bandwidth. On the other hand, the DiffServ approach uses a much coarser differentiation model to obviate the above disadvantages, where packets are classified into a small number of service classes at the network edge. The packets of each class are marked and traffic conditioned by the edge router, according to the resource commitment negotiated in the service level agreement (SLA). In each core router, QoS for different classes is differentiated by different *per-hop behaviors* [8]. Resource allocation is performed by the *bandwidth broker* [42] in a centralized manner, without dynamic resource reservation signaling and reservation status maintaining in the core routers. In this chapter, the registration domain will be modeled as a DiffServ administrative domain, with the gateway router as the edge router connecting to the Internet backbone and the BSs as the edge devices providing MNs wireless Internet access. A bandwidth broker will manage the resource allocation over the DiffServ registration domain.

The rest of this chapter is organized as follows. In Section 2, we give a brief review of the popular micromobility protocols and point out why they can not support soft handoff. Section 3 describes the system architecture for the soft handoff protocol and the DiffServ resource allocation. Details of the MCIP protocol are described in Section 4. In

Sections 5, 6 and 7, we discuss fast resource allocation in the DiffServ environment. Section 5 explains how to integrate the DiffServ QoS scheme with a registration domain. Section 6 investigates the call admission control over the DiffServ registration domain. After that, we present an adaptive assured service in Section 7, where multimedia applications experience bandwidth degradation and compensation, depending on the resource availability in wireless links. Section 8 presents numerical results to demonstrate the performance of the proposed MCIP protocol and the DiffServ resource allocation techniques. Finally, we conclude this research and discuss possible future research topics on IP-based soft handoff and wireless QoS in Section 9.

2. Previous Micromobility Protocols

Over the past several years a number of IP micromobility protocols [13] have been proposed that complement the Mobile IP by providing fast, seamless, and local handoff control. There have been a lot of studies [19, 15, 40, 12] of the protocol complexity, processing requirements, and the handoff performance of the micromobility protocols, especially of the two key protocols, Cellular IP [14] and HAWAII [41]. In this section, we review the micromobility protocols with the focus on their possibility to support soft handoff.

Soft handoff allows an MN to communicate with multiple BSs simultaneously during a handoff in CDMA systems. Therefore, the chance for successful data transfer can be increased, as the probability that all data copies from different BSs involved in the soft handoff are corrupted at the same time is significantly reduced. Soft handoff is typically very fast (on the order of 20 ms) [27]. There are three key requirements for realizing fast soft handoff:

- **Soft handoff signaling**: The serving BS (or the serving base station controller (BSC) in MCIP) needs to inform in time the new target BSs (or BSCs) and the MN to make resource preparations for the coming soft handoff, through a properly designed signaling protocol.

- **Data distribution and selection** [47]: Separate copies of the same data need to be sent via multiple BSs to the same MN in the downlink, or from an MN to multiple BSs in the uplink.

- **Data content synchronization** [47]: In the downlink, data segments arriving from multiple BSs to an MN at the same time should be copies of the same data content in order for the MN to correctly combine these copies into a single copy. In the uplink,

only one copy of the data sent by the MN to multiple BSs should be selected for delivery to the destination.

All the present micromobility protocols have been designed without explicitly considering the support of soft handoff. They can not fulfill all or some of the above requirements for fast soft handoff. Despite the apparent differences between various micromobility protocols, the operational principles that govern them are largely similar [12]. Therefore, we take Cellular IP and HAWAII as examples, and give detailed analyses of their deficiencies to support soft handoff. Other micromobility protocols suffer from the similar problems.

2.1 Cellular IP

2.1.1 Protocol Overview. As the name suggests, Cellular IP inherits cellular principles for mobility management such as passive connectivity, paging, and fast handoff control, but implements them in the IP paradigm. The universal component of a Cellular IP access network is the *base station* which serves as a wireless access point and router of IP packets while performing all mobility related functions. The BSs are built on a regular IP forwarding engine with the exception that IP routing is replaced by Cellular IP routing and location management. Cellular IP networks are connected to the Internet via *gateway* routers. MNs attached to an access network use the gateway IP address as their Mobile IP care-of-address. Assuming Mobile IPv4 [38] and no route optimization [39], packets are first routed to the host's home agent and then tunneled to the gateway. The gateway intercepts packets and forwards them toward the destination BS. Inside a Cellular IP network, MNs are identified by their home addresses, and data packets are routed to the host's actual location via Cellular IP routing protocol. Packets transmitted by MNs are first routed toward the gateway and from there onto the Internet.

An important concept in Cellular IP design is simplicity and minimal use of explicit signaling, which enable low-cost implementation of the protocol. In Cellular IP, location management and handoff support are integrated with routing. To minimize control messaging, regular data packets transmitted by MNs are used to refresh host location information. For example, *uplink* packets are routed from an MN with source IP address X to the gateway on a hop-by-hop basis. The path taken by these packets is cached by all intermediate BSs as soft-state mappings (X, BS_i), where BS_i is the neighbor from which packets enter the node. To route *downlink* packets addressed to MN X, the path used by recently transmitted packets from the MN is reversed. When the mobile node has

no data to transmit, it sends small, special *route-update* packets toward the gateway to maintain its downlink routing state. Otherwise, the route information will be deleted without refreshing in a system-specific time, called *route-timeout*. The same routing approach is used for *paging*. As an idle MN moves into a new *paging area*, it sends *paging-update* packets regularly to BSs which have better signal quality. As in the case of data and route-update packets, paging-update packets are routed toward the gateway on a hop-by-hop basis. The intermediate BSs maintain a *paging cache* in the same format and operation as the routing cache, with a longer timeout period called *paging-timeout*.

Cellular IP supports two types of handoff. Cellular IP *hard handoff* is based on a simple approach that trades off some packet loss for minimizing handoff signaling rather than trying to guarantee zero packet loss. Cellular IP *semisoft handoff* prepares handoff by proactively notifying the new access point before actual handoff. Semisoft handoff minimizes packet loss, providing improved TCP and UDP performance over hard handoff. However, semisoft handoff is in fact a kind of hard handoff, where the MN receive packets only from one BS at any time. In Cellular IP, both kinds of handoffs are mobile-controlled forward handoffs (MCFHOs) [5] [2]. MNs listen to beacons transmitted by BSs and initiate handoff based on signal strength measurements. To perform a handoff, a mobile tunes its radio to a new BS and sends a route-update packet. The route-update message creates routing cache mapping en route to the gateway, configuring the downlink route cache to point toward the new BS.

2.1.2 Soft Handoff Support. Cellular IP is targeted at providing high-speed packet radio access to the Internet with the design principle of lightweight nature. The development of the protocol is independent of the radio interface. Its mobility management functionalities, including the routing and handoff management, are implemented in the IP layer. However, without the participation of Layer 2, the Medium Access Control (MAC) and the Radio Link Control (RLC), at the radio interface, the space diversity combining (which is the fundamental of any soft handoff scheme) is impossible [47, 27, 25]. Specifically speaking, Cellular IP lacks the ability to support soft handoff because of the following reasons:

- There is no scheme for the BSs to communicate with each other in a Cellular IP access network. The BSs implement Cellular IP routing, where they only maintain mobile-specific routing entries. The packets with destination addresses other than an MN's IP address will be routed to the gateway. However, to support soft

handoff, different BSs must have the ability to exchange data with each other. The old serving BS has to exchange handoff signaling with the new target BS(s), and to forward the resource profile of the MN under consideration, all link layer and physical layer parameters to the new BS(s) so that the new air interface channel(s) between the MN and the new BS(s) can be established.

- The IP connection between the BS and the MN makes data content synchronization difficult. As an example, consider a downlink soft handoff of MN X involving two BSs. During the soft handoff, there are two traffic flows. One is from the old serving BS to MN X; the other is from the new BS to MN X, if we assume that the old BS can send the packets to the new BS without any delay. At both BSs, packets need to be buffered and scheduled to guarantee QoS in the IP layer. If the scalable DiffServ scheme is used, then at each BS traffic flow to MN X will be queued together with other flows subscribing to the same service class, in the same buffer [24, 18]. In this situation there is no way to guarantee that the two copies of a single packet destinated to MN X from the old BS and the new BS will be scheduled out simultaneously. With per-flow IntServ QoS model, the synchronization can be achieved by complicated per-flow scheduling algorithms [46], but it adds a large computation load to the already heavily burdened BSs. Furthermore, the scarce wireless bandwidth is not used in an efficient way when each IP packet is transmitted over the wireless link as it is. A detected error will lead to the retransmission of the whole packet, which will lead to a long delay and a waste of bandwidth when the packet is long.

- The MCFHO is not appropriate for seamless fast handoff in a non-random access system due to the associated large processing delay. In a CDMA network, the most common case is that the MN connects to the network via a dedicated channel to receive real-time services [25]. When the MN moves to the new BS's coverage area, it has to send the mobility update message via a Random Access Channel (RACH) [25]. An RACH usually has a rather low bit rate and uses a slotted-ALOHA like mechanism for bandwidth sharing among all the users in the cell. Thus the mobility update message is likely to experience an access delay due to collision and a long transmission delay due to the low bit rate. After the new BS receives the registration message, the BS performs a call admission control operation. If radio resources are available, the MN is assigned a dedicated channel by signaling through a downlink shared channel. This also takes time. During a hard handoff, if

the connection with the old BS turns off before the new BSs makes the call admission decision, and the last decision is "reject", the handoff call will be dropped. During a soft handoff, the long delay may also cause the connection to be dropped, if the handoff occurred due to bad radio link quality [37]. The work in [5] points out that the mobile assisted backward handoff should be a better choice for CDMA wireless networks.

2.2 HAWAII

HAWAII is similar in spirit to Cellular IP. HAWAII relies on Mobile IP to provide wide-area inter-domain mobility. A mobile entering a new foreign domain is assigned a collocated care-of address using Dynamic Host Configuration Protocol (DHCP). Packets are tunneled to the care-of address by the home agent in the home domain. The MN retains its care-of address unchanged while moving within the foreign domain, and the connectivity is maintained using dynamically established paths of HAWAII. Nodes in a HAWAII network are enhanced IP routers, which both execute the generic IP routing protocol and maintain mobility-specific routing information, where MN-based routing entries are added to the legacy routing table. Such MN location information is created, updated and modified by explicit signaling messages sent by MNs. HAWAII defines four alternative path setup schemes that control handoff between access points. The two forwarding schemes, the multiple stream forwarding (MSF) and the single stream forwarding (SSF), are for seamless hard handoff, where the old BS buffers packets and forwards them to the new BS. The multicast nonforwarding (MNF) scheme works in a way similar to the semi-soft handoff in cellular IP. The unicast nonforwarding (UNF) considers the situation that MNs can listen/transmit to multiple BSs in the CDMA network, but the ability is mainly used to continue data transmission between the MN and the old BS when the routing information and resource allocation are processed along the new path. HAWAII also uses IP multicasting to page idle MNs when incoming data packets arrive at an access network and no recent routing information is available.

HAWAII is different from Cellular IP mainly in two aspects. Firstly, the BSs and other nodes in the HAWAII network are enhanced IP routers. They have the ability to communicate with each other. It is attributed to this ability that HAWAII can build up its four path setup schemes to achieve seamless handoff. Actually, HAWAII nodes make soft handoff possible with the enhancement from the MCIP protocol. Secondly, Cellular IP relies on the gateway to act as a foreign

agent that decapsulates the packets before delivering them to the user. The presence of the gateway foreign agent can complicate per-flow QoS management, since the interior nodes cannot distinguish easily the packets destinated to different MNs but tunneled through the same gateway. However, if DiffServ QoS scheme is used, this would not be a problem for per-class QoS.

In HAWAII, BSs keep IP connections with MNs, and MCFHO schemes are used. As explained earlier, due to these two points, HAWAII can not be used directly for soft handoff.

2.3 Other Related Work

HAWAII and Cellular IP are local mobility protocols implemented in the IP layer. Protocols supporting local mobility in data link layer or by IP tunneling have also been proposed.

In [32] off-the-shelf Ethernet switches and wireless LAN cards are used to build wireless access networks. The learning feature of Ethernet switches is used for location management, which monitor (i.e., "snoop") mobile originated packets to update or fresh the bindings between host MAC address and the ports through which packets are received. Although this approach results in simple, low-cost, and efficient access networks, it is only suitable for small networks with the coverage radius limited to a few miles. Here, we consider outdoor cellular networks with a much larger coverage, where IP layer protocols are a better solution.

In [23, 21], Mobile IP is extended by arranging a hierarchy of foreign agents to locally handle Mobile IP registration. When necessary, MNs send Mobile IP registration messages to update their respective location information. Registration messages establish tunnels between neighboring foreign agents along the path from the MN to a gateway foreign agent at the top of the hierarchy. Packets addressed to the MN travel in this network of tunnels, which can be viewed as a separate routing network overlay on top of IP. Although the IP tunneling makes the protocols ready to be deployed over legacy IP routers, it brings complex operational and security issues [19]. The Hierarchy foreign agent architecture is not appropriate for soft handoff, where the signaling path (from the MN to the new BS, and then to a certain foreign agent) will be different from the data path (between the old BS and the new BS) in such an architecture. Also one BS has to tunnel copies of the packets to the other. The larger operational delay compared to that in Cellular IP and HAWAII systems makes the data content synchronization more difficult. The TeleMIP scheme [19] extends the protocol in [23] to a more comprehensive architectural framework for supporting intra-domain mobility

in a cellular wireless network. The *mobility* agent (MA) is introduced to the foreign agent hierarchy to achieve flexible address management, simple security management, and some traffic engineering functions [33]. But the mobility agent does not bring any enhancement regarding the soft handoff support.

The work in [27] also points out that the present micromobility protocols are not appropriate for soft handoff, from the perspective of the design of the CDMA radio access network (RAN). In the present CDMA RANs, soft handoff happens between BSs within the same RAN or in directly connected RANs. The extremely tight real time constraints (5ms to 80ms) on traffic in the RAN require that RAN traffic be quickly processed in the form of radio frames for efficient delivery to and from the radio medium by the BSs. In 3G wireless networks, when IP traffic is introduced, the IP layer processing should be terminated at the RAN gateway, also called BSC, which will be responsible for splitting packets into radio frames or combining radio frames into packets. However, the protocol to support soft handoff in such an "IP core, Layer 2 RAN" environment has not been designed yet. We propose such a protocol in Section 4.

3. System Architecture

We consider a domain-based architecture, as shown in Fig. 1.1. The radio cells and the related BSs in a geographic area are organized into a registration domain. Within the domain, the MCIP scheme is deployed for micromobility management and soft handoff support, and the DiffServ approach is used for QoS provision. In the following, we first focus on the MCIP protocol, and then discuss the DiffServ resource allocation techniques. The detailed structure of an MCIP domain is shown in Fig. 1.2, while how to implement DiffServ over such a domain is explained in Section 5.

As shown in Fig. 1.2, the proposed MCIP scheme adopts a 3-tier domain architecture to overcome the structural deficiency of present IP-based micromobility protocols for soft handoff. The first tier is a gateway router which connects the 3G wireless subnet [3] to the backbone Internet. All MNs within the gateway domain use the gateway's IP address as their care-of IP addresses. User traffic originating from or destined to the correspondent node (CN) is routed by Mobile IP outside the gateway domain and by MCIP routing inside the gateway domain. The gateway may collocate with a certain BSC (e.g., BSC 9 as in Fig. 1.2). The single gateway approach considerably simplifies the address management over the collocated care-of address assignment using DHCP. Although the

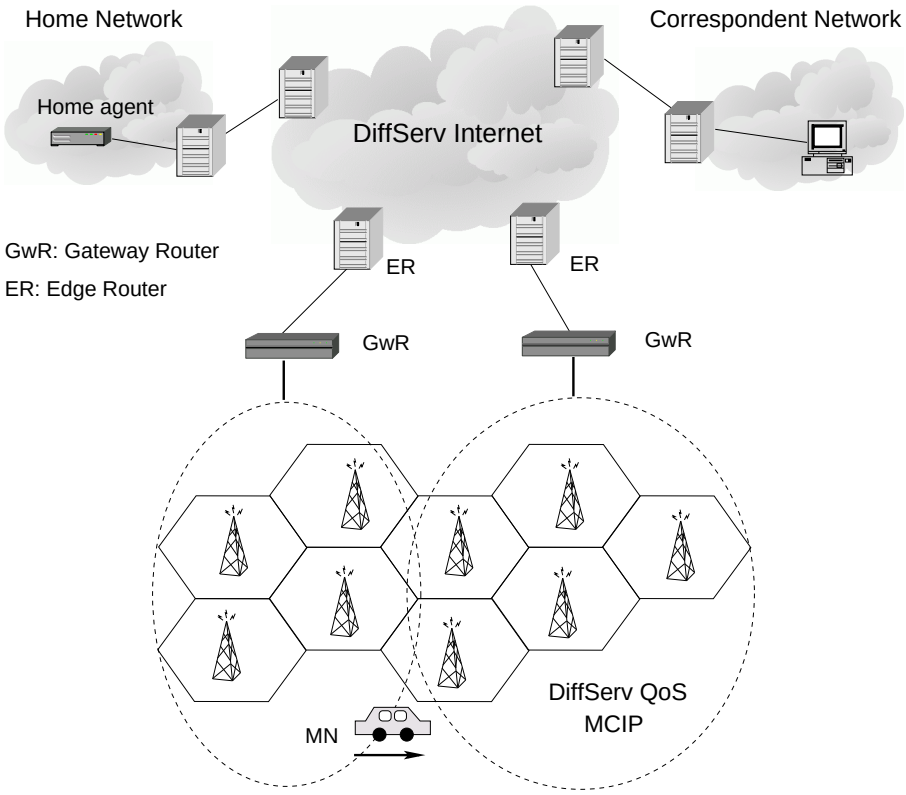


Figure 1.1. The domain-based system architecture.

single gateway may potentially impact reliability and complicate per-flow QoS management [41], the problems can be avoided by using backup devices and the DiffServ QoS approach, respectively.

The second tier consists of a mesh of BSCs, which are connected as a *core network*. A BSC has two roles: a router built on top of both regular IP and MCIP routing engines, and a radio network controller which is responsible for the radio resource management of all the cells, i.e., base transceiver stations (BTSs) [4], within its jurisdiction as defined in the 3G system proposals [25]. Each BSC and its associated BTSs constitute a cell cluster. Adjacent BSCs are connected by direct links to facilitate the inter-BSC (inter-cluster) soft handoff.

The third tier consists of hundreds of BTSs, which are organized into clusters; each cluster connects to a BSC. The BTS processes data only in the form of radio frames, and forwards them to its serving BSC for upper layer processing. A BTS has two roles: First, it takes part in the radio resource management within its cell under the control of the

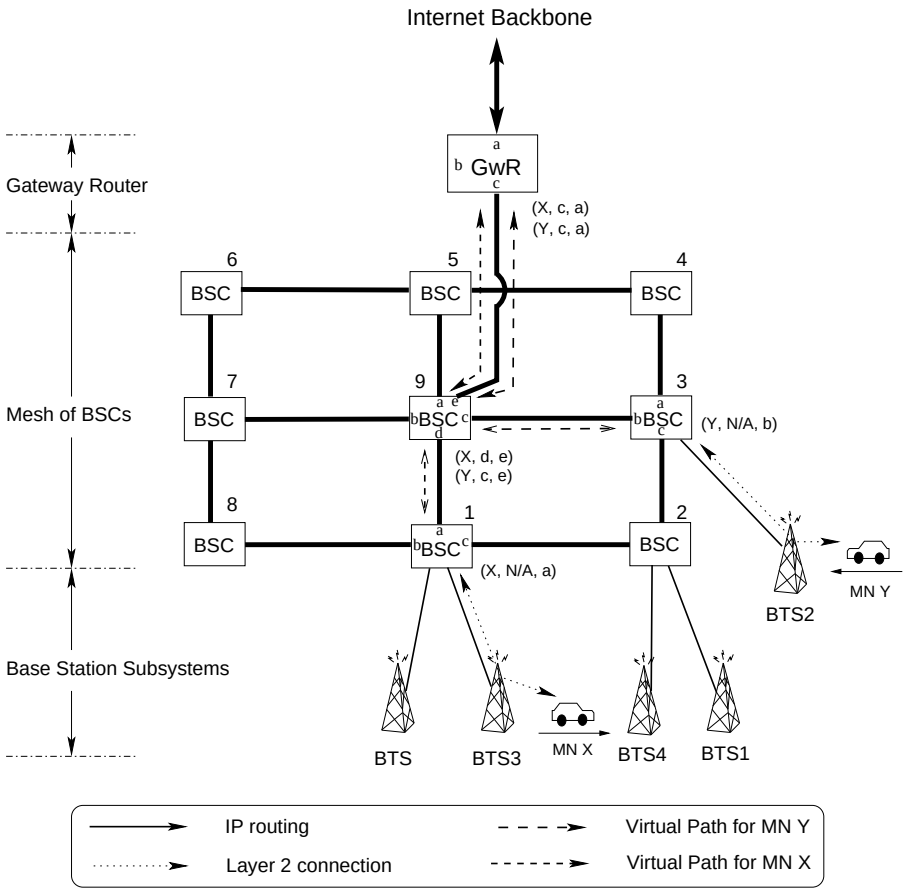


Figure 1.2. The 3-tier MCIP domain.

BSC, such as downlink close-loop fast power control, measurement of air interface traffic load, etc.. For this purpose, a BTS keeps an IP layer connection with its controlling BSC to obtain system parameters and to submit measurement reports; Second, a BTS works as a bridge in the radio interface between an MN and its serving BSC, delivering the basic MAC sublayer protocol data units (PDUs) which are received or to be transferred over the air. This function is implemented by the *bridge logic* at the BTS. Note that, in the system model, there is no direct IP layer connection between a BTS and an MN in the cell.

In summary, the 3-tier architecture is basically an "IP core, Layer 2 RAN" architecture, as suggested in [27]. Tier 1 and tier 2 have a structure similar to that of Cellular IP or HAWAII, and therefore, both the protocols can be applied directly to support the mobile controlled

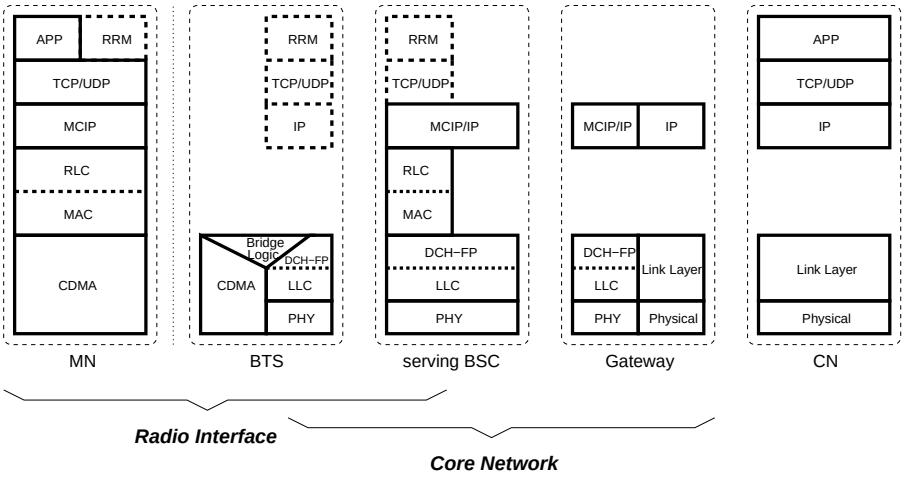


Figure 1.3. Protocol stacks in the MCIP access network.

forward hard handoff between BSCs. The introduction of tier 3 makes soft handoff achievable by processing data in Layer 2. In the next section, we design protocols for inter-cluster soft handoff and mobile assisted backward hard handoff.

4. Mobile Cellular IP (MCIP)

4.1 Protocol Architecture

The MCIP protocol architecture is shown in Fig. 1.3. Two interactive and overlapping protocol stacks are defined in MCIP: the radio interface protocol stack, and the core network protocol stack. They work in conjunction with each other to ensure fast and seamless microscopic mobility, with the capability to support the macroscopic diversity combining required by soft handoffs.

4.1.1 Radio Interface Protocol. The radio interface protocol defines the operations of the MN, its associated BTS and the serving BSC, to make the wireless link transparent to IP traffic and capable of providing bandwidth-on-demand services to the IP layer and above. The radio interface comprises the CDMA layer [25, 1], the Layer 2 consisting of the MAC [25, 2] and RLC sublayers [25, 3], and the Radio Resource Management (RRM) layer [25, 4]. Radio interface is actually the most complex part of a wireless network, and the detailed discussion is out of the scope of this chapter. However, some conceptual description is necessary for presenting the proposed soft handoff scheme.

The CDMA layer is the physical layer over a wireless channel, which provides digital data transmission over the wireless air interface. The MAC sublayer controls how much wireless bandwidth is allocated to a certain user, which can be considered as a logic channel. The logic channel is mapped to a *dedicated channel* in the CDMA layer, and the dedicated channel can be identified by the scrambling code and the spreading code used for the channel [25]. The RLC has the functions similar to those of its wireline counterpart, which implements error detection, error correction and the ARQ (Automatic Repeat Request) control upon the radio frames forwarded by the MAC sublayer. Basically, the wireless layers replace their wireline counterparts to achieve an IP-based radio interface. In Fig. 1.3, we can see that the protocol stack in the MN is designed exactly in this way. The complexity is brought by the other end of the wireless link, as the physical channel layer and logical channel layer reside in two different devices, the former in the BTS and the latter in the serving BSC. Therefore the BTS needs two interfaces: one is the wireline interface to the BSC, which receives the logic channel assignment and data radio frames from the BSC in the downlink and forwards data radio frames to the BSC in the uplink; the other is the wireless interface, which establishes the physical dedicated CDMA channel for data transfer. The bridge logic is necessary to understand the MAC resource allocation parameters received via the wireline interface and to map the logical channel to the physical dedicated channel.

In the MCIP scheme, the wireline interface of the BTS is enhanced so that it can establish an IP connection with its serving BSC to obtain system parameters and to submit measurement reports. This function also brings great flexibility for future IP-based network management. Hence, there are two data paths between the BTS and its serving BSC. One is for the radio resource management information and user data transfer:

$$\underbrace{\text{Bridge Logic} \leftrightarrow \text{LLC (Link Layer Control)} \leftrightarrow \text{PHY (Physical Layer)}}_{\text{BTS side}}$$
$$\Longleftrightarrow \underbrace{\text{PHY} \leftrightarrow \text{LLC} \leftrightarrow \text{MAC} \leftrightarrow \text{upper layers}}_{\text{BSC side}} \ ;$$

The other is for the IP connection between the BTS and its serving BSC:

$$\underbrace{\text{upper layers} \leftrightarrow \text{IP} \leftrightarrow \text{LLC} \leftrightarrow \text{PHY}}_{\text{BTS side}} \Longleftrightarrow \underbrace{\text{PHY} \leftrightarrow \text{LLC} \leftrightarrow \text{IP} \leftrightarrow \text{upper layers}}_{\text{BSC side}} \ .$$

In this two-path situation, when the LLC layer receives data from the PHY layer, it should determine which data path to follow. LLC itself

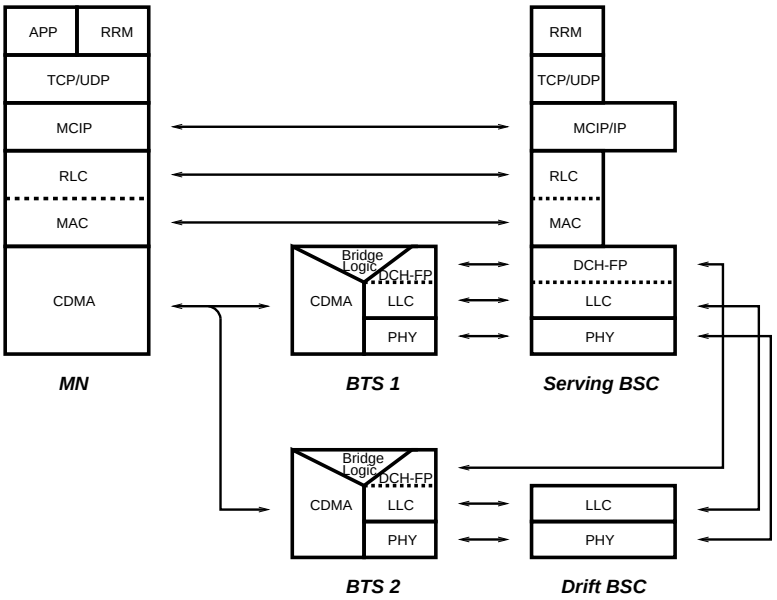


Figure 1.4. Space diversity in inter-cluster soft handoffs.

does not have this function, so we introduce the DCH-FP (Dedicated Channel - Frame Protocol) sublayer sitting on top of the LLC sublayer. The DCH-FP sublayer is to integrate the radio interface protocol stack and the core network protocol stack smoothly.

With the above protocol structure, all the radio resource management information, the user data, and other control messages can be transmitted as IP packets, which greatly simplify the network operation. The complex legacy network management model, where date plane and control plane are managed separately, is now unnecessary. Instead, we propose a *specific* application layer RRM protocol, which has control access to the RLC sublayer, the MAC sublayer and the CDMA layer (including receiving measurement reports from these layers), in addition to utilizing the transmission services offered by the TCP/UDP layer. On the other hand, a negative effect is that the MCIP air interface becomes a loosely layered protocol stack, due to the introduction of the RRM application layer.

Soft Handoff Support. The radio interface protocol stack is defined basically within a cell cluster, but can be extended to two or more clusters during handoffs. Fig. 1.4 shows the operation of the radio interface during an inter-cluster soft handoff. The *serving BSC* for an MN is the BSC that performs the Layer 2 (MAC and RLC sublayers) processing

of the IP packets to/from the *Transport Blocks* (TBs) [25]. The radio resource management operations, and outer loop power control, etc., are also executed in the serving BSC. Each MN has one and only one serving BSC. An MN is also associated with *drift BSC(s)*, which is the BSC(s) controlling those target BTSs involved in the soft handoff. In Fig. 1.4, the drift BSC is the controlling BSC of BTS 2. During an inter-cluster soft handoff, the communication between the MN and its serving BSC takes place via multiple air channels from multiple BTSs. Drift BSCs participate in the macrodiversity combining/splitting by forwarding the LLC PDUs, each of which encapsulates a DCH-FP PDU (containing a second copy of a TB) as its payload, between new target BTSs and the old serving BSC. The drift BSCs do not perform any Layer 2 processing of the IP packets. The combining/splitting procedures in an intra-BSC handoff are basically the same except that the drift BSC does not exist and the target BTSs have direct connections with the serving BSC. Here we only consider the general inter-cluster soft handoff. Although there is no limit for the number of BTSs participating in the soft handoff, it has been shown that no additional benefit will be achieved by involving more than two BTSs [5]. As a result, for presentation clarity, in the following we only describe the situation of two BTSs in the protocol design, even though the protocol structure is suitable for multiple (more than two) BTS soft handoff.

In the downlink direction, as illustrated in Fig. 1.4, 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* [29], the TB should be sent to both BTS 1 and BTS 2. The MAC sublayer at the serving BSC keeps the LLC addresses of both BTS 1 and BTS 2. The TB, the logic channel ID, and the LLC addresses of both BTS 1 and BTS 2 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 LLC PDUs each with a destination address of either BTS 1 or BTS 2. The two LLC PDUs are sent to BTS1 and BTS2, separately. In this way, the same TB together with its logical channel ID is delivered in Layer 2 to both the BTSs. At each of the two BTSs, the bridge logic receives the TB and the associated logical channel ID. The TB and the logical channel ID are passed down to the CDMA layer. After a series of physical layer processing procedures, two separate CDMA radio signals are sent to the MN by BTS 1 and BTS 2. At the MN, the two radio signals are combined by a *maximal ratio combining RAKE receiver* at the CDMA layer of the MN.

In the uplink direction, the radio frames sent from the MN are received by both the BTS 1 and BTS 2. The radio frames are combined into a TB and encapsulated with the CRC check result into a DCH-FP PDU at each BTS, and then routed to the DCH-FP sublayer at the serving BSC. The MAC layer of the BSC receives two versions of the same TB and CRC check result from the two BTSs, and selects the better one. The selected TB and CRC check result are submitted to the RLC layer for further processing.

Note that, during the soft handoff, IP layer processing is operated in the serving BSC. It is the LLC PDU copies of a packet that are sent to or collected from multiple BTSs. Therefore, the IP QoS processing has the same effect on all the paths involved in the soft handoff and does not affect the data content synchronization. Furthermore, implementing the LLC PDU transfer between the serving BSC and the drift BSC requires the BSCs to have the Layer 2 switching ability. It is not complicated to achieve in either hardware or software. We assume that the BSCs in the MCIP system indeed have the Layer 2 switching function.

4.1.2 Core Network Protocols. There exist two differences in the MCIP access system from the Cellular IP or the Internet. First, even though the network layer is based on the IP packet routing paradigm, two different routing algorithms are employed: the regular IP routing algorithm, and the MCIP routing algorithm. MCIP routing is used by user traffic and two signaling messages (*path-teardown*, *crossover-discovery*). All other signaling messages are routed by regular IP routing. Second, the data link layer comprises 2 sublayers: DCH-FP sublayer, and LLC sublayer which can be any regular link layer protocol supporting metropolitan area networking, such as ATM, Frame Relay, etc.. As shown in Fig. 1.3, the CDMA layer terminates at the MN and BTS, but the MAC and RLC sublayers terminate at the MN and the serving BSC. The DCH-FP sublayer is introduced to extend the transport channel, through which the CDMA layer provides services to the MAC sublayer, from the BTS to the serving BSC. Therefore, in addition to providing packet data transmission services to the network layer, the data link layer offers transparent data transmission services between the bridge logic at the BTS and the MAC sublayer at the serving BSC.

4.1.3 MCIP Routing. In MCIP, mobility management of active MNs is integrated with the routing and handoff schemes, as in the Cellular IP scheme, but explicit *path-setup* and *path-teardown* packets are employed to create and delete, respectively, an MN-specific *virtual path* [5] used by user traffic. A *path-setup* packet always originates from

the MN. It is first sent to the MN's serving BSC over the air via the random access channel and then routed towards the gateway by regular IP routing. The path taken is recorded by the intermediate nodes (BSCs) as an MN-specific routing entry. Such an entry is a triplet (*MN IP address, uplink incoming interface, uplink outgoing interface*). The downlink packet can find its destination MN by taking the reverse direction of the routing entry. Fig. 1.2 illustrates the MN-specific routing entries maintained for paths from gateway to MN X and MN Y. MN X and MN Y can exchange data easily by MCIP routing. After a virtual path is set up, all the user packets will follow it. A *path-teardown* packet may originate from any node on the virtual path. When the teardown decision is made by an intermediate BSC, it may send two *path-teardown* packets towards both directions respectively. Details about MCIP routing are given in [31].

4.1.4 Other Functions. In MCIP, *IP header compression* is used to improve the utilization efficiency of wireless resources. When transmitted over the air, the large headers of the IP packets degrade the spectrum efficiency and cause further delay to user traffic. As there exists a high degree of redundancy in the headers of consecutive packets that belong to the same packet stream, an IP header compression algorithm compacts the information by maintaining a context, and attaching each packet only a compressed header which carries information of changes to the context. During an inter-cluster handoff, the context is transmitted to the new BSC by handoff messages originating from the old BSC, and the IP address of the new BSC is sent to the MN over the old CDMA channel, so that the compression algorithm can work continuously. Details of IP header compression are given in [31].

MCIP also deploys a paging process, similar to that used in Cellular IP, except that the process is implemented at the IP layer. A gateway domain can be partitioned into one or several paging areas. After an MN is turned on or migrates to a new paging area, it receives the paging area ID number from the broadcast channel of the best cell, and then sends a *paging-update* packet to the gateway. A paging cache at the gateway maintains an entry for each MN. A Mobile IP binding update is needed if the MN is new to the gateway domain. When a call to an MN arrives at the gateway, a *paging* message is sent to the known paging area, and broadcast over the air via all the BTSs in the paging area.

4.2 Inter-Cluster Handoff

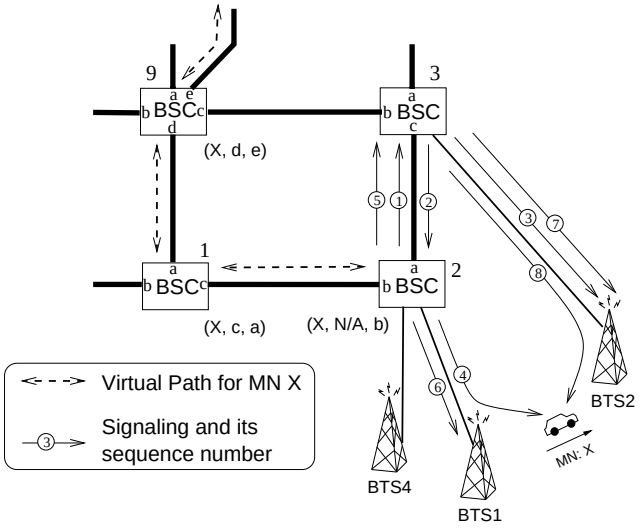
Both soft and hard handoffs are supported by the access system. Intra-cluster handoffs are handled by the BSC locally without inter-

vention of other system elements. Here, we focus on the inter-cluster soft and hard handoff.

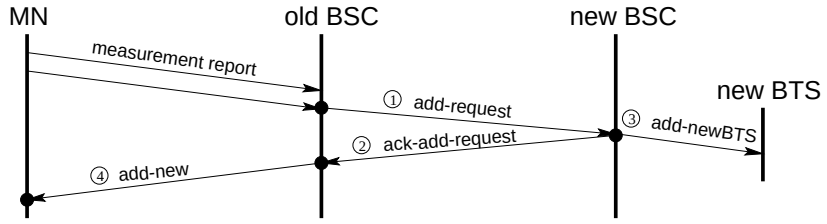
4.2.1 Soft Handoff. The inter-cluster soft handoff comprises two phases: *add* and *drop*. To highlight the handoff process, a segment of the access network (Fig. 1.2) is redrawn in Fig. 1.5(a). MN X connects to the network via BTS 1, and is moving toward BTS 2. MN X periodically performs the measurements of received signal-to-interference plus noise ratio over the pilot channels from its surrounding BTSs, averages the measurements over an averaging window, and reports the results to its serving BSC, BSC 1 [25]. After MN X moves further into the overlapping area, the new BTS (BTS 2) is added to the *active set* (Fig. 1.5(b)); when MN X leaves the overlapping area, the old BTS (BTS 1) is dropped from the active set as shown in Fig. 1.5(c). Both the add and drop decisions are made by BSC 2.

The add phase has 4 signaling messages: (1) *add-request* from BSC 2 to BSC 3: it carries all parameters of 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 executed. If available, the required radio resources at BSC 3 and BTS2 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 at 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 calculates 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 copy of the uplink TBs forwarded by BTS 2 via the DCH-FP sublayer and execute the uplink macrodiversity 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.

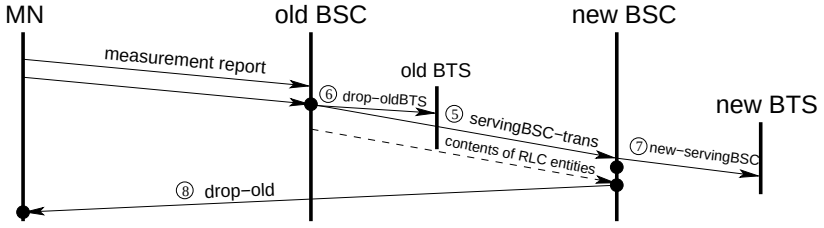
The drop phase also has 4 signaling messages: (1) *drop-oldBTS* from BSC 2 to BTS 1: it informs BTS 1 to release the old channel once the drop decision is made. BSC 2 also stops functioning as the serving BSC, stops popping downlink TBs into the DCH-FP sublayer, and releases



(a) MCIP soft handoff process

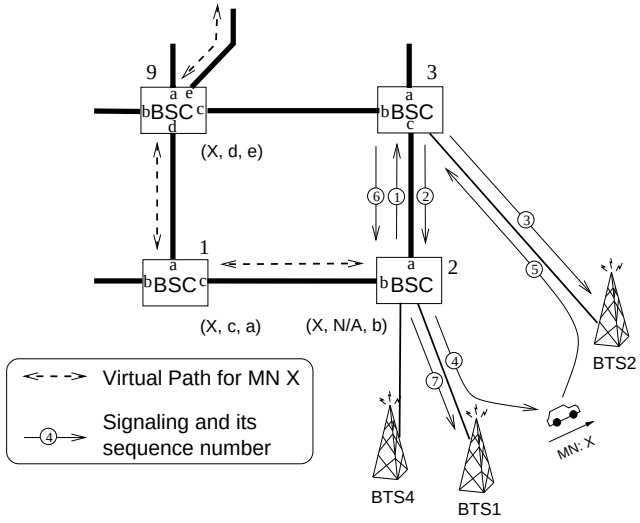


(b) The *add* phase signaling

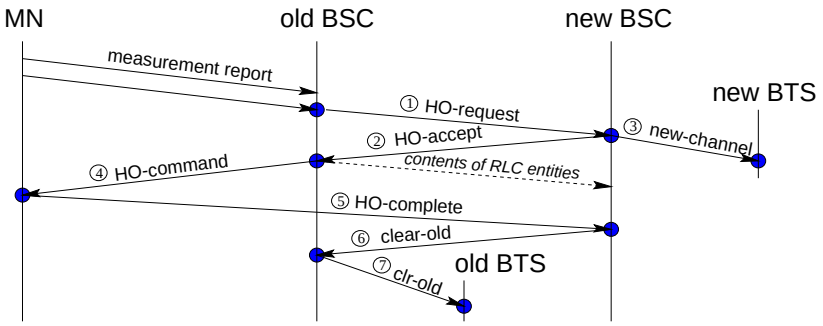


(c) The *drop* phase signaling

Figure 1.5. MCIP soft handoff.



(a) MCIP hard handoff process



(b) MCIP hard handoff signaling

Figure 1.6. MCIP hard handoff.

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. 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.

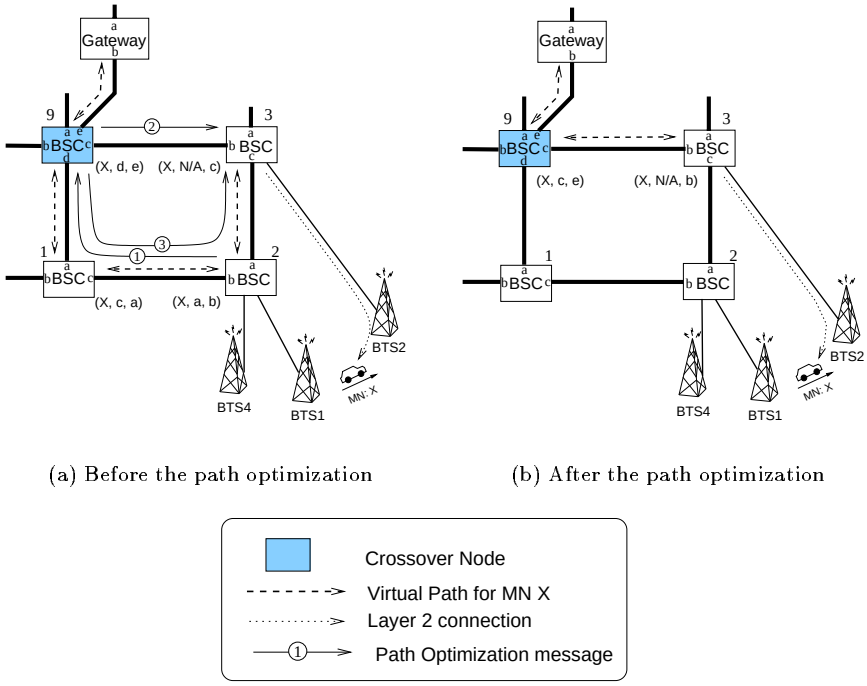
4.2.2 Hard Handoff. The hard handoff signaling procedure is given in Fig. 1.6. The details are omitted as the procedure is relatively straightforward and self-explanatory.

4.2.3 Path Optimization. For the case shown in Fig. 1.5, after MN X hands off from BTS 1 to BTS 2, its virtual path is extended from BSC 2 to BSC 3 as shown in Fig. 1.7(a). The path is 2 hops longer than the shortest path from BSC 3 to the gateway. A *path optimization* process is proposed to find the shortest path for a maximal path reuse efficiency. The *crossover node* is the common branch node of the two paths (the path from the old BSC to the gateway and the path from the new BSC to the gateway) that is closest to the old BSC. The handoffs can be classified into 3 cases in terms of the logical location of the crossover node: (1) the old BSC is the crossover node; (2) the new BSC is the crossover node; (3) neither the old nor the new BSC is the crossover node as shown in Fig. 1.5. In the first two cases, the path after the handoff is the shortest, and the path optimization is not necessary. The optimization signaling process for case (3) is shown in Fig. 1.7(a), and the path for MN X after the optimization is shown in Fig. 1.7(b). Details of the signaling are given in [31].

So far, we have presented the MCIP architecture, which inherits the advantages of existing micromobility protocols to support fast, seamless local handoff. Furthermore, in MCIP, the access network is properly structured to support soft handoff by utilizing space diversity in Layer 2. The protocol architecture is designed in such a way that the Layer 2 RAN and the IP core can cooperate well to provide wireless IP connection to any MN in the MCIP wireless domain. Next we will discuss the IP QoS provisioning over a wireless domain within the IP-based mobility management architecture.

5. DiffServ Registration Domain

The main objective of the next generation wireless networks is to provide real-time multimedia services to mobile users, which require strict QoS guarantee on packet loss, delay, or delay jitter [44]. Thus call admission control (CAC) is necessary to ensure the QoS satisfaction of the accepted connections. Also the associated resource allocation procedure should be fast enough to achieve seamless handoff.



(a) Before the path optimization

(b) After the path optimization

Figure 1.7. Path Optimization.

The IntServ approach uses per-flow resource reservation to achieve hard QoS guarantee and high efficient resource utilization compared to other QoS schemes [34]. Furthermore, in a wireless access network, the traffic load is not so heavy as that in an Internet backbone, and the per-flow processing is acceptable. But in a wireless domain, the IntServ has several obvious disadvantages as listed in the following.

- IntServ uses RSVP for resource reservation. RSVP involves hop by hop admission control and signaling processing. The RSVP signaling procedure may not be fast enough for a handoff, especially the soft handoff.

- In a wireless domain, normally the wireless links are the bottlenecks, while the bandwidths on the wireline links are sufficient. Using RSVP to reserve resources on those wireline links are not necessary.

- RSVP is receiver-oriented. In an "IntServ access, DiffServ backbone" environment [6], the gateway router of a wireless domain

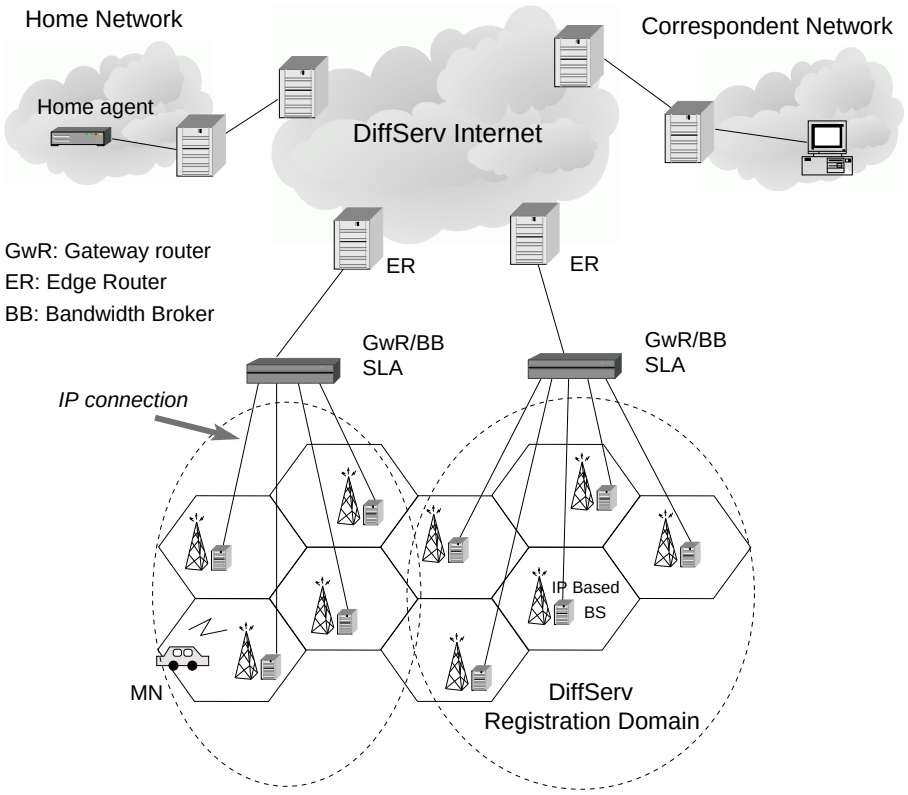


Figure 1.8. The DiffServ registration-domain-based wireless network architecture.

(see Fig. 1.2) acts as the receiver for all the MNs in the domain in resource reservation. The single hot spot leads to a long processing delay, even though we can use backup devices to avoid the potential reliability problem.

- When the IntServ domain interfaces with the DiffServ backbone, complex inter-operations are unavoidable to map the per-flow IntServ QoS specification to aggregate DiffServ QoS metrics [7].

Here we propose to implement the DiffServ QoS support over a wireless domain. The domain under consideration has a general architecture, and can deploy IP based micromobility protocols such as MCIP, Cellular IP, and HAWAII. The IP layer resource allocation is performed by a bandwidth broker in a centralized way. Therefore, the signaling overhead can be obviously decreased. Also, in the next section, we will propose a domain based call admission control algorithm to achieve fast resource allocation.

The DiffServ registration-domain-based wireless network architecture is shown in Fig. 1.8. In the system, all the registration domains are DiffServ administrative domains in which all the routers are DiffServ IP routers. The gateway and BSs are edge routers, and they are connected through core routers. Note that, in the system using MCIP, the BTS and the BSC cooperate to serve as one IP end of the wireless link. In discussing IP layer resource allocation, for simplicity, we can combine the two nodes into an *IP based base station*. The gateway is the interface to the DiffServ Internet backbone, where an SLA is negotiated to specify the resources allocated by the Internet service provider to serve the aggregate traffic flowing from/to the gateway. We consider wireless links as bottleneck links in the domain and the SLA is negotiated mainly based on the wireless resource availability. The gateway conditions the aggregate traffic for each class according to the SLA resource commitments. The BSs provide MNs the wireless access points to the Internet, and perform per-flow traffic conditioning and marking when data streams flow in the uplink direction. All the BSs in the same registration domain are connected to the same gateway router. All DiffServ routers use three separate queues to provide the premium service [26], the assured service [24], and the best-effort service, respectively. The three buffers are served under priority scheduling or weighted fair queue (WFQ) scheduling [46]. The traffic classes in the next-generation wireless networks can be mapped to these three DiffServ classes. For example, in a UMTS (Universal Mobile Telecommunications System) wireless network [20], the conversational class and the streaming class can be mapped to the premium service and the assured service, respectively, while the interactive class or the background traffic can be mapped to the best effort class. A bandwidth broker residing in the gateway router is responsible for the resource allocation and call admission control over the DiffServ registration domain.

6. Domain-Based Call Admission Control

For simplicity, we assume that an *effective bandwidth* can be used to characterize both the traffic characteristics and the QoS requirements, and the calls belonging to the same service class have homogeneous traffic characteristics, and thus have the same effective bandwidths. For example, for a premium service call the peak rate can be used as the effective bandwidth, and for an assured service call the effective bandwidth can be determined using the approach given in [16]. The resource commitments specified in the SLA can then be represented in terms of how many calls for each class are allowed in the registration domain. As

a result, the proposed admission control procedure is straightforward: whenever a new MN requests admission to a registration domain, the bandwidth broker determines whether to admit or reject the new call, based on the number of the calls currently in service and the SLA allocation for the service class to which the new call subscribes. The new call has to be dropped if all the SLA allocation has been occupied. This procedure requires very simple communications between the edge router (the BS) and the bandwidth broker (in the gateway router), and can be executed very fast. Furthermore, once an MN is admitted to a registration domain, it can hand off to other cells within the domain without the involvement of further call admission control in the bandwidth broker. During a soft handoff, a connection occupies bandwidth in two or more cells. However, as the handoff duration is very short compared to the connection's lifetime, the soft handoff's effect on bandwidth consumption can be ignored for steady-state analysis.

The above simple resource allocation scheme based on effective bandwidth in fact implies a very complicated design problem. The number of BSs in a domain, the resources allocated to each service class in each base station, and the resource commitments in the SLA should be determined carefully so that the new call blocking and handoff call dropping probabilities are reasonably low, while considering the traffic load in the registration domain, the mobility information and the call duration statistics. In [35], this design problem is solved for the situations that the interval between call arrivals, cell residence time and call duration are independently and exponentially distributed. An important conclusion is that a predetermined handoff call dropping probability can always be guaranteed by properly designing the admission controller. However, in the analysis, handoff calls and new calls are not differentiated. From users' point of view, it is better to be blocked at the beginning of a call than to be dropped in the middle of the call. As a result, handoff calls should be serviced with a higher priority than new calls. To further decrease the handoff call dropping probability, here we use the guard channel scheme [45] to reserve a fixed percentage of each BS's resources for handoff calls and extend the analysis given in [35] to include this situation.

Under the assumption that the complete partitioning scheduling mechanism among the service classes [36] is used, the admission control of each class can be considered separately. Consider a registration domain including M cells, where the new call arrivals of a certain service class are Poisson with a mean rate of λ calls per cell per unit time, and the call duration is exponentially distributed with mean $1/\mu$. Each cell can serve up to C calls of the class under consideration, and a percentage

(α) of the cell capacity, αC, is set as the guard channel to protect the handoff calls. The channel holding time in a BS (i.e., the time that a call spends with any BS before handing off to another BS) is exponentially distributed with mean $1/h$, i.e., the handoff rate is h. Similarly as in [35, 36], we assume that the handoff rate from any cell to any other cell is such that all cells experience the same rate of handoff call arrivals. The handoff call dropping can be taken into account by considering the fact that the effective (actual) call departure rate μ_e is higher than the "natural" call departure rate μ. Then the approximation technique in [35, 36] can be used to calculate the new call blocking and handoff call dropping probabilities.

A new call can be blocked by the admission controller if the total number of calls in the domain exceeds the SLA resource allocation N ($C < N < MC$), and/or if the serving cell is full and can not accept any additional connections. Let P_{Badm} denote the probability of a new call being blocked by the admission controller, and P_H denote the handoff call dropping probability. P_{Badm} and P_H can be calculated by solving the following set of nonlinear equations:

$$\mu_e = \mu + hP_H \tag{1.1}$$

$$\lambda_e = \lambda(1 - P_{Badm}) \tag{1.2}$$

$$P_{Badm} = E(M\lambda/\mu_e, N) \tag{1.3}$$

$$P_H = E_{gH}\left(\frac{\lambda_e}{\mu_e}, \frac{h}{\mu + h}\frac{\lambda_e}{\mu_e}, \alpha C, C\right) \tag{1.4}$$

where (1.1)-(1.3) are the same as those used in [35, 36], and (1.4) is different because of the guard channel introduced. In the above equations, $E(\rho, C)$ represents the Erlang loss formula defined as $(\rho^C/C!)/(\sum_{i=0}^{C}\rho^i/i!)$. $E_{gH}(\rho_{tot}, \rho_h, C_g, C)$ represents the handoff call dropping probability in a fixed guard channel system, where ρ_{tot} is the total Erlang load consisting of both the new calls and handoff calls, ρ_h is the Erlang load consisting of only the handoff calls, and C_g is the resources allocated for the guard channel. In the system, each call intends to hand off to a neighbor cell with a probability of $h/(\mu + h)$, so in the steady state the handoff Erlang load occupies $h/(\mu + h)$ of the total traffic load. Letting $E_{gB}(\rho_{tot}, \rho_h, C_g, C)$ denote the new call blocking probability in the

guard channel system, we have

$$E_{gH}(\rho_{tot}, \rho_h, C_g, C) = P_0 \frac{\rho_{tot}^C}{C!} \left(\frac{\rho_h}{\rho_{tot}} \right)^{C_g} \tag{1.5}$$

$$E_{gB}(\rho_{tot}, \rho_h, C_g, C) = 1 - P_0 \sum_{i=0}^{C-C_g-1} \frac{\rho_{tot}^i}{i!} \tag{1.6}$$

where P_0 is the probability that all channels are unoccupied and is given by

$$P_0 = \left[\sum_{i=0}^{C} \frac{\rho_{tot}^i}{i!} (\frac{\rho_h}{\rho_{tot}})^{\max(i-C+C_g,0)} \right]^{-1}. \tag{1.7}$$

Based on (1.5) and (1.7), equations (1.1)-(1.4) can be solved recursively [35, 36]. After obtaining P_{Badm} and P_H, we can calculate μ_e and λ_e from (1.1) and (1.2), respectively. Then based on μ_e and λ_e, we can compute the new call blocking probability in a cell from (1.6), given by $P_{Bcell} = E_{gB}(\frac{\lambda_e}{\mu_e}, \frac{h}{\mu+h}\frac{\lambda_e}{\mu_e}, \alpha C, C)$. Then, the total blocking probability of a new call is given by

$$P_B = 1 - (1 - P_{Badm})(1 - P_{Bcell}) \approx P_{Badm} + P_{Bcell}. \tag{1.8}$$

7. Adaptive Assured Service

The streaming class defined in the 3G wireless networks can be supported by the assured service in a DiffServ architecture. A streaming class traffic, such as a streaming video, normally does not require very strict timely delivery but requires a guaranteed minimum delivery rate. Adaptive coding can be applied to this type of traffic to improve sustain probability when the network congests. A good example is the MPEG video coding format where a base layer contains basic and extension layer additional information. The video quality and bandwidth consumption can be scaled down to the bottom by only transmitting the base layer information. In a wireless network, because resource availability fluctuates frequently due to user mobility and channel quality variations, this type of adaptive service is very important to improve resource utilization efficiency. The adaptive framework to be discussed in this section only takes mobility into consideration. That is, the bandwidth allocated to an adaptive video changes only when there is a new call arrival, call completion, or handoff.

Here we propose to use a partitioned buffer [16] with size B to serve a layered video at each DiffServ router. Assume that a video traffic is

Table 1.1. The buffer configurations used to provide the assured service

Buffer Configuration	Management Rule	QoS Level	Loss Probability	Effective Bandwidth
$[0, B]$	any packet accepted, when $0 \leq X < B$	`high`	ϵ_1 to L_1, L_2 and L_3	e_1
$[0, B_1, B]$ $(0<B_1<B)$	L_1 packets accepted, when $0 \leq X < B$; L_2 and L_3 packets accepted, when $0 \leq X < B_1$	`medium`	ϵ_1 to L_1, ϵ_2 to L_2 and L_3	e_2
$[0, B_1', B_2', B]$ $(0<B_1'<B_2'<B)$	L_1 packets accepted, when $0 \leq X < B$; L_2 packets accepted, when $0 \leq X < B_2'$; L_3 packets accepted, when $0 \leq X < B_1'$;	`low`	ϵ_1 to L_1, ϵ_2 to L_2, and ϵ_3 to L_3	e_3

coded to three layers, L_1, L_2, and L_3. The assured service buffer uses three configurations (1, 2, and 3) to provide three levels of QoS (`high, medium and low`) to the video traffic, as shown in Table 1.1, where X denotes the number of packets queued in the buffer, and ϵ_1, ϵ_2, and ϵ_3 ($\epsilon_1 < \epsilon_2 < \epsilon_3$) denote different levels of loss probability provided by the different buffer configurations, respectively. In a homogeneous environment, all the video traffic for the assured service in the domain has homogeneous statistical characteristics and the buffer for the assured service in a DiffServ router is a homogeneous multiplexing system. We model each video traffic flow as a multiclass Markov-modulated fluid source (MMFS), where an underline Markov chain determines the traffic generation rate at each time instant. At each state of the Markov chain, three layers of traffic, L_1, L_2 and L_3, are generated. In such a scenario, an *optimal effective bandwidth* can be calculated by using the technique developed in [16], which is the minimal channel capacity required to guarantee the loss requirements for all the layers of the MMFS traffic when the buffer partition thresholds are optimally selected. For each QoS level listed in Table 1.1, the associated effective bandwidth can be calculated and is denoted by e_1, e_2 and e_3 ($e_1 > e_2 > e_3$), respectively.

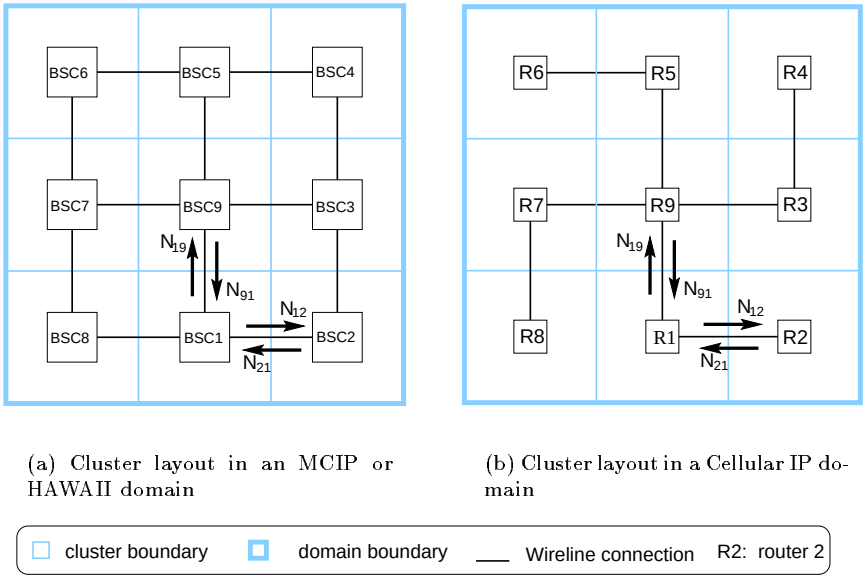
Under the assumption that the wireless link is always the bottleneck and the traffic will be served without loss in the wireline links, the adaptive mechanism is mainly used for the buffers in the MNs (for uplink traffic flows) or in the BSs (for downlink traffic flows). Each BS has a local admission/rate controller (LARC) to manage the admission control and the adaptive bandwidth allocation in the cell. In the downlink direction, the LARC can directly adjust the buffer configuration in the

BS when bandwidth adaptation happens; in the uplink direction, the LARC should send messages to MNs to control the buffer configuration adjustment for bandwidth adaptation. The adaptive algorithm based on the effective bandwidth is straightforward. If we begin with a light load situation when sufficient resources are available, a new or handoff call arrival to the cell is admitted with bandwidth allocation of e_1, meaning that the `high` level of QoS is provided to the traffic. As the traffic load increases to a level where a new or handoff call can not be admitted with bandwidth e_1, the LARC then reduces the bandwidth allocation to each already accepted video call from e_1 to e_2, and adjusts the assured service buffer in the BS or MNs from configuration 1 to configuration 2 to fit the adaptive bandwidth allocation. The new request is accepted also with bandwidth allocation of e_2. In this situation, the `medium` level of QoS is to be provisioned. If the traffic load further increases to a certain degree, the LARC reduces the bandwidth allocation of both existing traffic and new requests to e_3 and changes to buffer configuration 3, trying to accept as many calls as possible by degrading the QoS to the `low` level. On the other hand, when traffic load decreases, the buffer can shift back to the higher level of configuration and the LARC can allocate more bandwidth to the calls for better QoS. Consider that at the current level each call is allocated an effective bandwidth of e_i $(i = 2, 3)$. At the moment that a call completes or hands off to a neighbor cell, the LARC will check whether the available resources are enough to improve the bandwidth allocation of each call to e_{i-1}. If the available resources can be used to accept two more calls with allocation of e_{i-1} after the bandwidth increase,[6] the LARC will implement the adaptation and the buffer will shift up to the one-level higher configuration.

When the adaptation happens from one configuration to another providing higher QoS, the buffer is lightly loaded at that moment and has free queueing space left. The higher level configuration can use the space to allow more packets to be buffered to provide better QoS. The transition is always smooth. On the other hand, when the buffer needs to switch into a configuration providing lower QoS, more packets should be dropped. In [17], a simple and efficient mechanism is designed to achieve a smooth transition from the higher QoS level to the lower QoS level.

8. Performance Evaluation

In this section, we present some numerical examples to show the performance of the proposed MCIP micromobility management and the DiffServ resource allocation techniques.



(a) Cluster layout in an MCIP or HAWAII domain

(b) Cluster layout in a Cellular IP domain

Figure 1.9. Cluster layout.

8.1 MCIP Performance

In the following, we discuss the scalability of the MCIP access system through a numerical example, compare its handoff signaling cost with those of HAWAII and Cellular IP, and give a qualitative analysis why HAWAII and Cellular IP are not appropriate for 3G/IP interworking.

The HAWAII domain used in the comparison has the same topology and configuration as the MCIP domain shown in Fig. 1.2, except that the BSCs and BTSs are replaced by routers and BSs, respectively. The layout of clusters in an MCIP or HAWAII domain is shown in Fig. 1.9(a). Since Cellular IP does not support the topology shown in Fig. 1.2, the wireline connections between routers in its second tier are modified into a tree topology, as that shown in Fig. 1.9(b) while other configurations in the Cellular IP domain remain unchanged. Note that, in a HAWAII or Cellular IP access network, the radio interface is defined between an MN and its serving BS.

Scalability. The configuration parameters of an MCIP domain are listed in Table 1.2, which are similar to those used in [41]. The coverage area of a BTS is a square. The scalability of a HAWAII domain with similar configurations is demonstrated in [41]. For MCIP, the number of mappings at the gateway (which is the same as the number of ac-

tive users) is 39,438. It is well within the capability of modern routers [41]. Furthermore, a majority of these MNs are completely specified for a particular domain/subnet. In this case, perfect hashing is possible, resulting in $O(1)$ memory access for IP route lookup. Thus, route lookup for data forwarding can be done efficiently at the gateway [41].

Handoff Signaling Efficiency. Because the routing and handoff schemes used in MCIP, Cellular IP and HAWAII are different, their signaling overheads are also different. Here we focus on their mean handoff signaling costs. Since an MN in an MCIP domain keeps IP layer connection with its serving BSC, while an MN in a HAWAII or Cellular IP domain keeps IP connection with its serving BS, to make a fair comparison, we only compare the handoff signaling in the core network (at the second tier). Also, the metric we use in the comparison is "number of hops per second", but not the number of signaling messages per second. This is because a signaling message which travels several network nodes consumes more resources than a message which is delivered from a node to its next hop neighbor. In the following analysis, we use a fluid flow mobility model [43]. The model, also used in [41], assumes that MNs are moving at an average velocity of v=112 km/hr, and their direction of movement is uniformly distributed over [0, 2π]. Assuming that MNs are uniformly populated with a density of $\rho = 39$ users per km^2, the rate of cluster boundary crossing, R, is given by $R = \rho v L_R/3600\pi = 16.4$ (1/s). In terms of the cluster boundary crossing rate, letting N_{ij} denote the mean number of signalings from BSC i to BSC j per second, we calculate N_{12}, N_{21}, N_{19}, and N_{91} for the system shown in Fig. 1.9. The *path-setup* and *path-teardown* messages for the inter-domain handoffs are also included in the calculation, without considering the turn-ons and turn-offs inside each domain. With the

Table 1.2. Domain Configuration Parameters

Symbol	Description	Value
n_1	BTSs per Domain	144
n_2	BSCs per Domain	9
n_3	BTSs per cluster	16
L_c	Perimeter of a cell	10.6 km [41]
ρ	User density (active user)	39 per km^2 [41]
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 [$(L_D/4)^2$]	1,011.24 km^2
N	Number of active users in a Domain [$A\rho$]	39,438

Table 1.3. Comparison of Handoff Signaling Costs in the Core Network

	MCIP I	MCIP E	HAWAII	Cellular IP
N_{19}	20.5	18.4	10.2	98.3
N_{91}	10.2	6.1	4.1	49.1
N_{12}	7.2	6.1	4.1	24.6
N_{21}	12.3	10.2	6.1	49.1
N_{HO}	278.8	228.4	138.8	884.4

symmetrical characteristic of the network topology, the total signaling cost due to inter-cluster (and inter-domain) handoffs in an MCIP domain is given by $N_{HO} = 4(N_{19} + N_{91}) + 8(N_{12} + N_{21})$. The number is the same for both soft and hard handoff schemes. The signaling cost for HAWAII *MSF hard handoff* scheme [41] can be calculated in the same way, where intra-cluster handoffs do not cause extra signaling in the core network. For Cellular IP *indirect semi-soft handoff* scheme [14] (which is actually a hard handoff scheme), because its network topology is different from MCIP and HAWAII, its signaling cost is given by $N_{HO} = 4(N_{19} + N_{91} + N_{12} + N_{21})$, where both inter-cluster and intra-cluster handoffs require signaling exchanges in the core network. Table 1.3 lists the numbers of the signaling messages in MCIP, HAWAII, and Cellular IP, respectively. For MCIP, both the numbers including (MCIP I) and excluding (MCIP E) the path optimization signalings are given.

In MCIP, the signalings are not symmetric in the two directions of a link, i.e., $N_{19} \gg N_{91}$ and $N_{21} \gg N_{12}$. This is mainly due to the effect of the inter-domain handoffs (note that the gateway router is collocated with BSC 9). When an inter-domain handoff happens, for example, when an MN moves into the domain under consideration, a *path-setup* message is 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 signalings from the surrounding BSCs towards 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 three reasons. First, because HAWAII paths are soft-state, for inter-domain handoffs, when an MN moves out of the domain, it does not signal the intermediate nodes which cache entries for its soft-state path, but just 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 reduces the number of mappings cached by each BSC. This in turn reduces the route lookup time and improves the network scalability. Second, in HAWAII there is no path optimization

process. However, the path optimization is necessary, because the cost of each optimization process is 6-hop signaling, but without it each data packet has to travel 2 hops more than necessary. This degrades the network resource utilization efficiency and causes a longer delay to user traffic, which is not shown in Table 1.3. 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. MCFHO scheme may work well in random access radio networks, but it is not suitable for 3G/IP interworking.

Cellular IP is the least efficient in terms of the signaling cost for handoffs. Cellular IP handoff also suffers from a long handoff delay due to the MCFHO approach, similar to that used in HAWAII. Moreover, after a handoff, the old BS cannot quickly release the radio resources previously used by the MN, 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 resource utilization efficiency.

8.2 DiffServ Resource Allocation Performance

In the following, we present two numerical examples to show the performance of the proposed resource allocation techniques. The first example shows that handoff call dropping probability can be decreased by setting fixed guard channel in each cell for handoff calls. The second one shows that the proposed adaptive scheme can decrease the new call blocking and handoff call dropping probabilities as compared to the non-adaptive scheme.

8.2.1 Call Admission With Guard Channel. Here, we use the parameter configuration given in [35] and make comparisons between call admission with guard channel and that without guard channel in terms of the call blocking and dropping probabilities. With the complete partitioning scheduling mechanism, we focus on the admission control of assured service traffic. New calls arrive at each cell according to a Poisson process, and the exponentially distributed call duration has an average of $1/\mu = 0.5$ units of time. The channel holding time in each BS is also exponentially distributed with an average of $1/h = 0.1$ units of time. Each registration domain has $M = 20$ cells. Each cell can support up to $C = 20$ assured service calls. Bandwidth adaptation is not considered here. New call requests are rejected by the domain admission controller if there are $N = 320$ (80% of the total capacity of the domain)

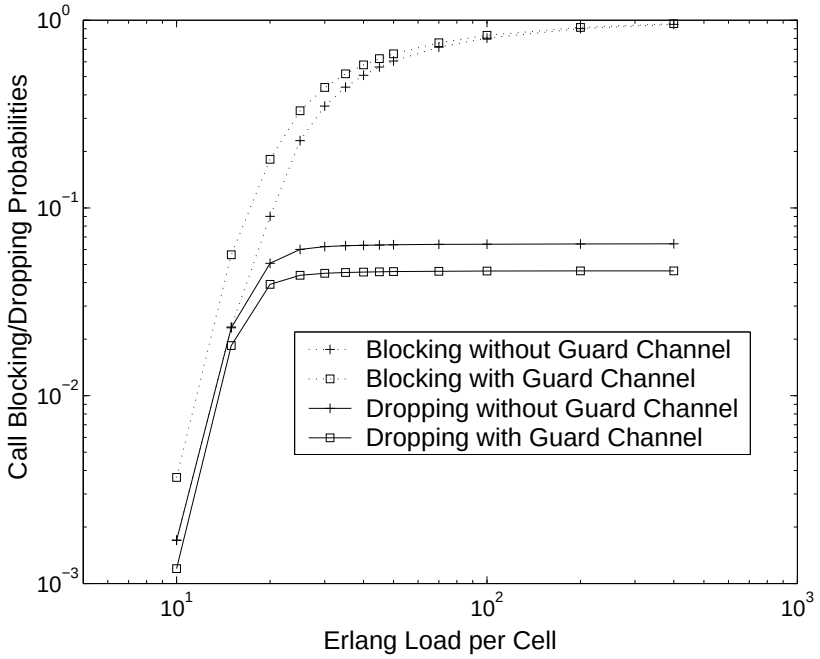


Figure 1.10. A comparison of the domain-based admission control systems with guard channel and without guard channel.

calls currently in service in the domain. When guard channel is used, $\alpha = 10\%$ of the cell capacity (2 calls) is reserved for the handoff calls.

Fig. 1.10 shows the new call blocking and handoff call dropping probabilities versus the Erlang load of new calls per cell, ρ. For comparison, the performance curves of the admission control without guard channel [35] are also included. From the figure, we can see that the introduction of guard channel can decrease the handoff call dropping probability at the cost of an increase in the new call blocking probability. In the light load situation, it is highly possible that the number of on-going calls in the registration domain is less than 320 and no call is rejected by the domain admission controller. The mechanism used in each cell determines the call blocking and dropping probabilities. The tradeoff between the decrease of handoff call dropping and the increase of new call blocking is clearly observed. As the Erlang load increases, both the call dropping and blocking probabilities increase. When the traffic load is extremely high, most of the new requests will be rejected by the domain admission controller which limits the Erlang load in each cell to approximately N/M [35], and therefore limits the handoff call dropping probability to 0.0644 without guard channel and to 0.0463 with guard

Table 1.4. The new call blocking and handoff call dropping probabilities achieved by the adaptive assured service.

Erlang Load	New Call Blocking Probability			Handoff Call Dropping Probability		
	`High` Config.	`Medium` Config.	`Low` Config.	`High` Config.	`Medium` Config.	`Low` Config.
15	0.0563	0.0246	0.0102	0.0185	0.0062	0.0026
50	0.6614	0.6286	0.5960	0.0458	0.0336	0.0303
400	0.9576	0.9535	0.9494	0.0462	0.0339	0.0307

channel. We can see that the handoff call dropping probability can be guaranteed to a predetermined level by properly selecting N when other parameters are fixed. This analysis should be helpful in determining the resource requirement in the SLA negotiation between the registration domain and the Internet service provider.

8.2.2 Adaptive Bandwidth Allocation. Consider a registration domain with the same configuration as that in the above numerical analysis example with the guard channel. To show how new call blocking and handoff call dropping probabilities can be reduced by implementing the adaptive assured service, the bandwidth allocation to each call can now be adjusted between e_1, e_2 and e_3. The capacity allocated to the assured service in each cell is Ce_1 packets/second. If at some time the LARC adjusts the QoS to `medium` level and reaches the steady state, then each cell can admit $\lfloor Ce_1/e_2 \rfloor$ calls, and the domain admission controller will allow $0.8M\lfloor Ce_1/e_2 \rfloor$ calls to be accepted. The guard channel in each cell will be set to $\lceil \alpha Ce_1/e_2 \rceil$. If the LARC adjusts the QoS to the `low` level, then the cell capacity, the admission controller capacity and the guard channel will be adjusted to $\lfloor Ce_1/e_3 \rfloor$, $0.8M\lfloor Ce_1/e_3 \rfloor$ and $\lceil \alpha Ce_1/e_2 \rceil$ calls, respectively.

For simplicity, we model a video flow as an on-off source as in [16] to illustrate the efficiency of the adaptive scheme. Each video source at the "on" state generates traffic with a total rate $R_p = 170.21$ packets/second, which can be coded into three layers L_1, L_2 and L_3 at rates of $R_p/8$, $R_p/8$ and $3R_p/4$, respectively. The buffer size in an MN or a BS is set to $B = 250$ packets. The three levels of loss probabilities, ϵ_1, ϵ_2 and ϵ_3, are set to 10^{-10}, 10^{-4} and 10^{-1}, respectively. For the `high` configuration listed in Table 1.1, a first-in-first-out buffer is used and the effective bandwidth (e_1) is 142.45 packets/second; for the `medium` configuration, the technique in [16] is used to calculate the effective bandwidth (e_2), equal to 119.82 packets/second, while the optimal partition threshold B_1 is 210 packets; for the `low` configuration, the effective bandwidth (e_3)

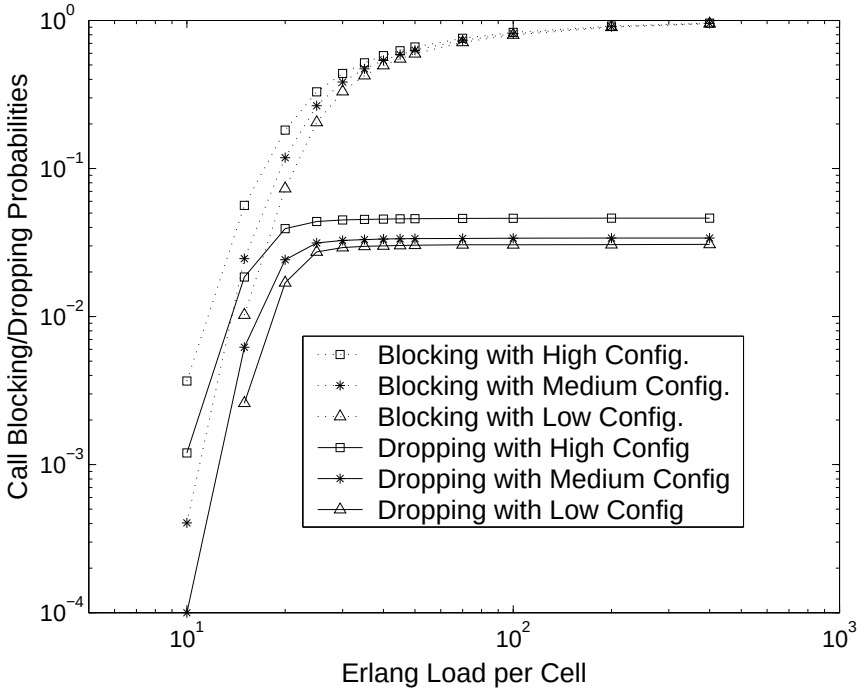


Figure 1.11. The performance improvement achieved by the adaptive framework.

is 112.65 packets/second and the optimal thresholds B_1' and B_2' are 64 and 171 packets respectively. By scaling down the QoS requirement, the bandwidth allocated to the on-going calls can be reduced. The resources accumulated by bandwidth reduction are then used to accept more new calls and handoff calls. Fig. 1.11 shows the new call blocking and handoff call dropping probabilities for the three configurations. It is observed that both the new call blocking and handoff call dropping probabilities decrease obviously in the adaptive scheme as the QoS level is reduced. For clarity, Table 1.4 gives the new call blocking and handoff call dropping probabilities for the medium load ($\lambda = 15$), high load ($\lambda = 50$), and extremely high load ($\lambda = 400$) conditions, respectively.

9. Conclusion and Future Research

This chapter addresses two important issues in establishing a mobile wireless Internet, i.e., support of fast soft handoff and provision of QoS over IP based wireless access networks.

First, the MCIP micromobility system is established. The MCIP architecture not only supports fast, seamless local hard handoff by inher-

iting the advantages of the existing micromobility protocols, but also supports soft handoff by properly selecting the access network structure and designing new handoff protocol. The scalability of the MCIP access system is illustrated by a numerical example, and the handoff signaling cost is compared with those of HAWAII and Cellular IP. The advantages of supporting soft handoff and QoS provisioning for real-time services are achieved at slightly increased system complexity.

Second, the DiffServ registration-domain-based resource allocation architecture is presented to support fast handoff with QoS guarantee. A novel adaptive assured service is proposed to improve wireless resource utilization efficiency. For the statistical environment with exponential distributions, the analysis demonstrates that (a) the handoff call dropping probability can be guaranteed to a predetermined level by properly allocating the resources to a certain class of traffic in a registration domain, (b) the guard channel scheme can be used to further reduce the handoff call dropping probability, and (c) the adaptive service can improve resource utilization while guaranteeing the call level QoS (i.e. the call blocking and dropping probabilities). This analysis can help to determine the resource requirement in the SLA negotiation between the registration domain and the Internet service provider.

The research results presented in this chapter is a first step towards developing the wireless mobile Internet. There are many open issues that need further investigation, some of which are described in the following.

MCIP supports fast soft handoff of mobile IP connections, mainly due to its adoption of the "Layer 2 RAN, IP core" architecture. However, there are several motivations for the use of IP transport within the RAN [20]: (a) IP is quickly becoming the widely accepted standard for packetization of voice, data, signaling, and operation, administration, and management (OAM) in the networking world; (b) IP as a network layer protocol is carefully designed to be independent of link/physical layers, so it allows easy cooperation of networks using different link/physical layer techniques; (c) the 3G core network is IP-based, and therefore an IP-based RAN will facilitate consistent backbone infrastructure, simple protocol stacks, operational efficiency, and industry standard OAM; (d) IP QoS management is approaching maturity.

With MNs' IP packets being routed directly in the RAN, IP over the air interface would be necessary. The current major impediment to IP over the air is the header size, but new work in header compression may eliminate this hurdle. Given that a spectrally efficient representation of IP on the radio medium is possible, IP packets can be sent out over the air by an MN, and the BS can handle the packets in the same way as it currently does with the specialized radio frames. However, there is still

a problem with the multipath nature of macrodiversity. In the downlink direction, the routing from a single source at the BS to multiple BSs is similar to multicast, but with extremely tight transmission delay and delay jitter constraints for real-time traffic. In the uplink direction, however, traffic flows need to be combined (in the selective diversity mode) in the wired network after the BSs. In order to accommodate macrodiversity, the RAN should be a routing domain using a multipath routing protocol, which performs frame selection in some specific routers and features real time routing table convergence (for soft handoff). These characteristics involve a considerable change from existing IP routing algorithms (which do not require real time convergence and do not involve multipath nor selection of particular packets in the diversity combining) [27]. Therefore, although the prospects for moving IP into the RAN for transport of RAN protocols and user traffic to/from the MN are promising, a lot of research work remains to be done.

While various micromobility protocols have been proposed to support fast and seamless handoff of real time multimedia IP connections, there is very little work on a suitable QoS model for micromobility. Due to the obvious disadvantages of the IntServ approach listed in Section 5, extending the DiffServ model to micromobility seems to be a better choice. The DiffServ resource allocation techniques proposed in this chapter are just a start point for this topic.

The domain based call admission control technique clearly illustrates that DiffServ QoS model can integrate micromobility naturally, and the resource allocation based on SLA and effective bandwidth should be fast enough for seamless handoff while guaranteeing QoS at both packet level and call level, but the model is over-simplified to facilitate mathematical analysis. First, the registration domain is assumed to be an exponentially distributed statistical environment, so that the new call blocking and handoff call dropping probabilities can be calculated by using the Erlang formula. However, the statistical characteristics of the arriving process, the connection lifetime, and the channel holding time of future wireless IP multimedia connections are unknown at this moment. To acquire accurate knowledge of the above random process/random variables, further investigation is necessary. Second, it is assumed that an effective bandwidth can be calculated for each connection of any service class. Although we present a technique to calculate effective bandwidth for the assured service provisioning packet loss guarantee, how to include the delay requirement (if any) in the effective bandwidth calculation is still unknown. Third, complete resource partition among different service classes is assumed; however, dynamic resource sharing among service classes is an important way to improve wireless bandwidth uti-

lization [37]. As a result, dynamic resource sharing techniques should be investigated, and the call admission control for a service class should consider the impacts of other service classes on the resource availability.

Notes

1. In different context, the mobile IP host may also be called mobile host, mobile station, or mobile user. Here we use the mobile IP terminology.

2. There are two mechanisms identified for handoff: backward and forward handoff. In general, backward handoff is initiated by the serving BS, whereas forward handoff is initiated via the target (new) BS.

3. As the 3G wireless wireless networks will be extensively deployed in the near future, we try to make our scheme compatible with the 3G standard as much as possible.

4. BTS is the 3G terminology for base station.

5. When the mobile-specific routing scheme is used, there is only one path between the domain root router and the MN. All packets belonging to the same connection will follow this path in the domain so as to keep the in-sequence packet order. Therefore, we call a mobile-specific path as a virtual path.

6. Here the capacity for two calls with allocation of e_{i-1} serves as a threshold to detect whether the traffic load really decreases. If the threshold is set to 0, the fluctuation of available capacity due to the new arrival, handoff or call completion will lead to too frequent adaptation of the buffer configuration, but with little improvement of the QoS. How to optimally set the threshold needs further investigation.

References

[1] 3GPP TS 25.211 v3.3.0 (2000-06). Physical Channels and Mapping of Transport Channels onto Physical Channels (FDD) (Release 1999).

[2] 3GPP TS 25.321 v3.8.0 (2001-06). MAC Protocol Specification (Release 1999).

[3] 3GPP TS 25.322 v3.7.0 (2001-06). RLC Protocol Specification (Release 1999).

[4] 3GPP TS 25.922 v4.0.0 (2001-03). Radio Resource Management Strategies (Release 4).

[5] A.H. Aghvami and P. Smyth. Forward or Backward Handover for W-CDMA? *First Int'l. Conf. 3G Mobile Communication Technologies*, (Conf. Publ. No. 471), pp. 235-239, 2000.

[6] Y. Bernet. The Complementary Roles of RSVP and Differentiated Services in the Full-Service QoS Network. *IEEE Commun. Mag.*, vol. 38, no. 2, pp. 154-162, Feb. 2000.

[7] Y. Bernet et al. A Framework for Integrated Services Operation over DiffServ Networks. IETF RFC 2998, Nov. 2000.

[8] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An Architecture for Differentiated Services. IETF RFC 2475, Dec. 1998.

[9] L. Bos and S. Leroy. Toward an All-IP-Based UMTS System Architecture. *IEEE Net.*, vol. 15, no. 1, pp. 36-45, Jan.-Feb. 2001.

[10] R. Braden, D. Clark, and S. Shenker. Integrated Services in the Internet Architecture: An Overview. Internet RFC 1633, June 1999.

[11] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin. Resource Reservation Protocol (RSVP)–Version 1 Functional Specification. Internet RFC 2205, Sept. 1997.

[12] A.T. Campbell et al. Comparison of IP Micromobility Protocols *IEEE Wireless Commun. Mag.*, vol. 9, no. 1, pp. 2-12, Feb. 2002.

[13] A.T. Campbell and J. Gomez. IP Micro-Mobility Protocols. *ACM SIGMOBILE, Mobile Comp. and Commun. Rev.*, vol. 4, no. 4, pp. 45-54, Oct. 2001.

[14] A.T. Campbell, J. Gomez, S. Kim, A.G. Valko, C-Y Wan, and Z.R. Turanyi. Design, Implementation and Evaluation of Cellular IP. *IEEE Pers. Commun.*, vol. 7, no. 4, pp. 42-49, Aug. 2000.

[15] A.T. Cambell, S. Kim, J. Gomez, and C-Y. Wan. Cellular IP Performance. Internet draft, draft-gomez-cellularip-perf-00.txt, work in progress, Oct. 1999.

[16] Y. Cheng and W. Zhuang. Optimal Buffer Partitioning for Multiclass Markovian Traffic Sources. *Proc. IEEE GLOBECOM'01*, vol. 3, 2001, pp. 1852-1856.

[17] Y. Cheng and W. Zhuang. DiffServ Resource Allocation for Fast Handoff in Wireless Mobile Internet. *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 130-136, May 2002.

[18] D.D. Clark and W. Fang. Explicit Allocation of Best-Effort Packet Delivery Service. *IEEE/ACM Trans. Networking*, Vol. 6, no. 4, pp. 362-373, Aug. 1998.

[19] S. Das, A. Misra, P. Agrawal, and S.K. Das. TeleMIP: Telecommunications-Enhanced Mobile IP Architecture for Fast Intradomain Mobility. *IEEE Pers. Commun.* , vol. 7, no. 4, pp. 50-58, Aug. 2000.

[20] S. Dixit, Y. Guo, and Z. Antoniou. Resource Management and Quality of Service in Third-Generation Wireless Networks. *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 125-133, Feb. 2001.

[21] S.F. Foo and K.C. Chua. Regional Aware Foreign Agent (RAFA) for Fast Local Handoffs. Internet draft, draft-chuafoo-mobileip-rafa-00.txt, work in progress, Nov. 1998.

[22] V.K. Garg. *IS-95 CDMA and CDMA 2000.* Prentice Hall, 2000.

[23] E. Gustafsson, A. Jonsson, and C. Perkins. Mobile IP Regional Registration. Internet draft, draft-ietf-mobileip-reg-tunnel-03, work in progress, July 2000.

[24] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski. Assured Forwarding PHB Group. IETF RFC 2597, June 1999.

[25] H. Homa and A. Toskala. *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications.* John Wiley & Sons, 2000.

[26] V. Jacobson, K. Nichols, and K. Poduri. An Expedited Forwarding PHB. IETF RFC 2598, June 1999.

[27] 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, work in progress, Sept. 2001.

[28] J. Kim and A. Jamalipour. Traffic Management and QoS Provisioning in Future Wireless IP Networks. *IEEE Pers. Commun.*, vol. 8, no. 5, pp. 46-55, Oct. 2001.

[29] J. Laiho-Steffens, M. Jasberg, K. Sipila, A. Wacker, and A. Kangas. Comparison of Three Diversity Handover Algorithms by Using Measured Propagation Data. *Proc. IEEE VTC*, vol. 2, pp. 1370-1374, 1999.

[30] 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 Pers. Commun.* , vol. 7, no. 6, pp. 42-47, Dec. 2000.

[31] X. Liu and W. Zhuang. Inter-Cluster Soft Handoff in 3G/IP Interworking. *Proc. 3Gwireless'2002*, pp. 778-783, San Francisco, May 2002.

[32] J.W. Lockwood. Implementation of Campus-wide Wireless Network Services Using ATM, Virtual LANs and Wireless Base Stations. *Proc. IEEE WCNC*, vol. 2, pp. 603-605, New Orleans, LA, Sept. 1999.

[33] A. Misra, S. Das, A. Mcauley, A. Dutta, and S.K. Das. Integrating QoS Support in TeleMIP's Mobility Architecture. *Proc. IEEE Int'l Conf. Personal Wireless Commun.*, pp. 57-64, 2000.

[34] B. Moon and H. Aghvami. RSVP Extensions for Real-Time Services in Wireless Mobile Networks. *IEEE Commun. Mag.*, vol. 39, no. 12, pp. 52-59, Dec. 2001.

[35] M. Naghshineh and A.S. Acampora. Design and Control of Micro-Cellular Networks with QoS Provisioning for Real-Time Traffic. *Proc. IEEE 3rd Int'l Conf. Universal Personal Commun.*, 1994, pp. 376-381.

[36] M. Naghshineh and A.S. Acampora. QoS Provisioning in Micro-Cellular Networks Supporting Multimedia Traffic. *Proc. IEEE INFOCOM'95*, 1995, pp. 1075-1084.

[37] D. Partain, G. Karagiannis, P. Wallentin, and L. Westberg. Resource Reservation Issues in Cellular Access Networks Internet draft, draft-partain-wireless-issues-00.txt, work in progress, April 2001.

[38] C. Perkins. IP Mobility Support IETF RFC 2002, Oct. 1996.

[39] C. Perkins and D.B. Johnson. Route Optimization in Mobile IP. Internet draft, draft-ietf-mobileip-optim-10.txt, work in progress, Nov. 2000.

[40] R. Ramjee et al. IP Micro-Mobility Support Using HAWAII. Internet draft, draft-ietf-mobileip-hawaii-00.txt, work in progress, June 1999.

[41] R. Ramjee, T. La Porta, S. Thuel, K. Varadhan, and S.Y. Wang. HAWAII: A Domain-based Approach for Supporting Mobility in Wide-area Wireless Networks. *IEEE/ACM Trans. Networking*, Vol. 10, No. 3, pp. 396-410, June 2002.

[42] A. Terzis, L. Wang, J. Ogawa, and L. Zhuang. A Two-Tier Resource Management Model for the Internet. *Proc. of IEEE GLOBECOM'99*, vol.3, pp. 1779-1791, 1999.

[43] R. Thomas, H. Gilbert, and G. Mazziotto. Influence of the Mobile Station On the Performance of a Radio Mobile Cellular Network. *Proc. 3rd Nordic Seminar, Paper 9.4*, Copenhagen, Denmark, 1988.

[44] X. Xiao and L.M. Ni. Internet QoS: A Big Picture. *IEEE Net.*, vol. 13, no. 2, pp. 8-18, March/April 1999.

[45] O.T.W. Yu and V.C.M. Leung. Adaptive Resource Allocation for Prioritized Call Admission over an ATM-Based Wireless PCN. *IEEE J. Select. Areas Commun.* , vol. 15, no. 7, pp. 1208-1225, Sept. 1997.

[46] H. Zhang. Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks. *Proc. IEEE*, vol. 83, no. 10, pp. 1374-1396, Oct. 1995.

[47] T. Zhang, P. Agrawal, and J-C Chen. IP-Based Base Stations and Soft Handoff in All-IP Wireless Networks *IEEE Pers. Commun.*, vol. 8, no. 5, pp. 24-30, Oct. 2001.