WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. 2005; **5**:713–732
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/wcm.338

Impact of mobility on mobile telecommunications networks

Yi-Bing Lin[1]*,†, Ai-Chun Pang[2] and Herman Chung-Hwa Rao[3]

[1] *Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu, Taiwan*
[2] *Graduate Institute of Networking and Multimedia, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan*
[3] *FET Laboratories, Far Eastone Telecommunication Co., Taipei, Taiwan*

Summary

This paper describes the mobility management mechanisms for mobile telecommunications networks. There are two major types of mobility: radio network mobility and core network mobility. Radio network mobility supports radio link switching of a mobile user during conversation, and core network mobility provides roaming and tunnel-related management for packet re-routing due to user movement. Impact of mobility on both the radio and the core networks is addressed in this paper. Also, potential research issues on these topics are discussed. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: core network; general packet radio service (GPRS); mobility management; radio network; universal mobile telecommunications system (UMTS)

1. Introduction

Mobile telecommunications services have become part of everyday life for most people. Such services allow one to communicate at anytime and in anywhere. In the United Kingdom, 31% of population considers the mobile phone as the most important invention in the modern world as compared with microwave (28%) and internet (10%) [11]. In Taiwan, the penetration rate of mobile subscription is now over 140%. Since 1980, mobile telecommunications technologies have been intensively studied. One of the major issues is mobility management for tracking locations of mobile users. There are two major types of mobility [30]. Radio network mobility supports handoff (i.e., radio link switching) of a mobile user during conversation. Core network mobility provides roaming mobility and tunnel-related management for packet re-routing in the core network due to user movement. In this paper, we focus on impact of mobility on both the radio and the core networks, and discuss potential research issues on these topics.

*Correspondence to: Yi-Bing Lin, Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan.
†E-mail: liny@csie.nctu.edu.tw

Contract/grant sponsor: NSC Excellence; contract/grant numbers: NSC93-2752-E-0090005-PAE; NSC 93-2213-E-009-100; NSC 93-2213-E-002-025; NSC94-2213-E-002-048.
Contract/grant sponsor: NTP VoIP Project; contract/grant number: NSC 92-2219-E-009-032.
Contract/grant sponsors: IIS/Academia Sinica; ITRI/NCTU Joint Research Center; Intel; CCL/ITRI.

1.1. Impact of Mobility on the Radio Network

We use universal mobile telecommunications system (UMTS) packet-switched (PS) domain [7,19] as an example to illustrate impact of mobility on the radio network. UMTS is a third generation system proposed by the 3rd generation partnership project (3GPP), which is designed to support higher data transmission rate for mobile users, and to provide streaming, interactive and background services with better quality of services [23]. As shown in Figure 1, the UMTS infrastructure includes the core network (CN) and the UMTS terrestrial radio access network (UTRAN). The CN is responsible for routing data connections to the external network, while the UTRAN handles all radio-related functionalities. In the CN, packet data services of a mobile station (MS; see Figure 1(f)) are provided by the serving GPRS support node (SGSN; see Figure 1(b)) and the gateway GPRS support node (GGSN; see Figure 1(c)). The SGSN connects the MS to external packet data networks (see Figure 1(a)) through the GGSN. The UTRAN consists of node Bs (the UMTS term for base stations; see Figure 1(d)) and radio network controllers (RNCs; see Figure 1(e)) connected by an asynchronous transfer mode (ATM) network. The connection between the UTRAN and the CN is achieved via the ATM links between the RNCs and the SGSNs. The user equipment (UE; the UMTS term for the MS) communicates with node Bs through the radio interface based on WCDMA (wideband CDMA) technology.

In UMTS networks, the data packets are routed between the UE and the GGSN. Through the packet data protocol (PDP) context activation procedure [7], a PDP context is created to establish the routing path for data packet delivery. The PDP context contains the UE's IP address, the QoS profiles and other parameters. Due to CDMA characteristics, multiple radio paths (for delivering the same data packet) may exist between a UE and more than one node Bs. An example of multiple routing paths is illustrated in Figure 2(a), where an IP-based GPRS tunneling protocol (GTP) connection is established between the GGSN and RNC1. The UE connects to two node Bs (B1 and B2). Node B1 is connected to RNC1, and node B2 is connected to RNC2. An Iur link between RNC1 and RNC2 is established so that the signal (i.e., data packets) sent from the UE to node B2 can be forwarded to RNC1 through RNC2. RNC1 then combines the signals from nodes B1 and B2, and forwards them to SGSN1. Similarly, the packets sent from the GGSN to RNC1 will be forwarded to both node B1 and RNC2 (and then node B2). In this example, RNC1 is called the serving RNC (SRNC). RNC2 is called the drift RNC (DRNC), which transparently routes the packets through the Iub interface (between the node B and the RNC) and Iur interface (between two RNCs). Suppose that the UE moves from node B1 toward node B2, and the radio link between the UE and node B1 is disconnected. In this case, the routing path will be ⟨UE↔node B2↔RNC2↔RNC1↔SGSN1↔GGSN⟩ as shown in Figure 2(b). In this scenario, it does not make sense to route packets between the UE and the CN through RNC1. Therefore, SRNC relocation may be performed to remove RNC1 from the routing path. After SRNC relocation, the packets are routed to the GGSN directly through RNC2 and SGSN2 (see Figure 2(c)), and RNC2 becomes the SRNC.

In the SRNC relocation procedure described above, the downlink packets may be lost or delayed during the SRNC switching period. For real-time multimedia applications, a long packet delay or a large amount of packet loss will result in service degradation. Therefore, providing an efficient real-time SRNC relocation

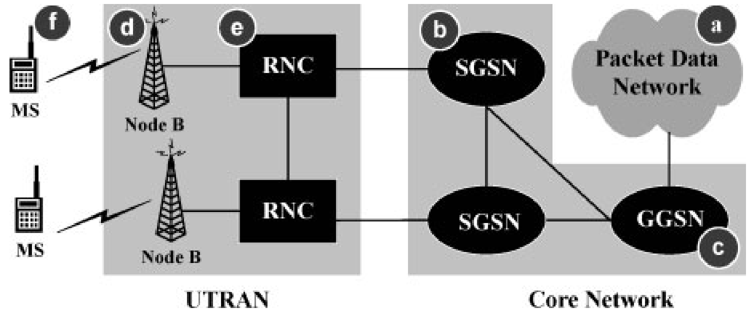

Fig. 1. A simplified network architecture for the universal mobile telecommunications system (UMTS) packet-switched (PS) domain. GGSN: gateway GPRS support node. MS: mobile station. Node B: base station. RNC: radio network controller. SGSN: serving GPRS support node. UTRAN: UMTS terrestrial radio access network.

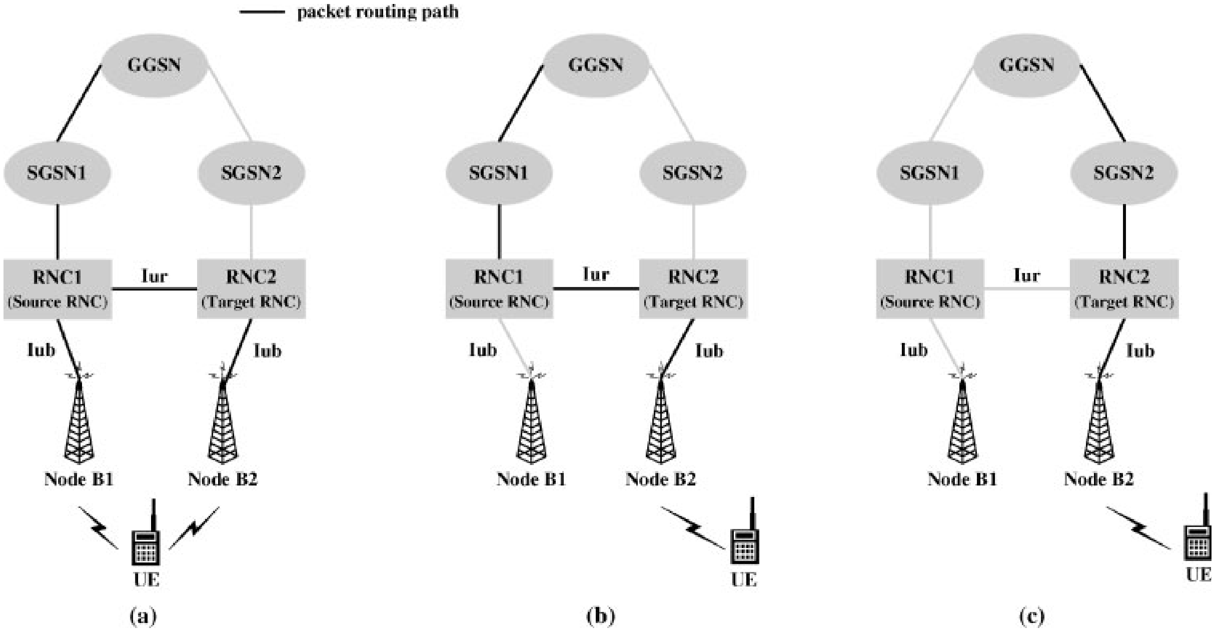

Fig. 2. Serving radio network controller (SRNC) relocation. (a) The UE connects to both B1 and B2. (b) The UE connects to B2 (before relocation). (c) The UE connects to B2 (after relocation).

has become one of the most important issues for UMTS network. This issue will be discussed in Subsection 2.1.

1.2. Impact of Mobility on the Core Network

Through mobility management, a mobile telecommunications network provides services to roaming users who move around the service areas covered by the network. Furthermore, with roaming agreement, mobile telecommunications networks belonging to different operators can interwork to offer services to users who move around various networks. For example, a mobile service subscriber of FarEasTone in Taiwan can use his/her MS (e.g., mobile phone) to make/receive phone calls in England through Vodafone/AirTouch, Cellnet (British Telecom) or other cellular operators. Mobile telecommunications networks use a distributed database architecture to support roaming of users. This architecture defines two types of databases: home location register (HLR) and visitor location register (VLR). When a user subscribes to the services of a cellular operator (called the home system of the user), a record is created in the operator's HLR. The record stores profiles of services (such as call waiting, call forwarding, voice mailbox, and so on) subscribed by the user. Furthermore, the location information of the user is kept in the record. Typical size of an HLR in Taiwan is around one million records.

When a mobile user visits a mobile telecommunications network other than the home system, a temporary record for the mobile user is created in the VLR of the visited system. The VLR temporarily stores subscription information (replicated from the HLR) for the visiting subscribers so that the visited system can provide services. In other words, the VLR is the location register other than the HLR used to retrieve information for handling calls to or from a visiting mobile user. The capacity of a typical VLR in Taiwan is around 250 000–500 000 records. To track the location of an MS, the MS automatically reports its location (to both the visited VLR and the HLR) when it moves to a new location. This procedure is called registration. To deliver a call to an MS, the network retrieves the location information stored in the HLR and the VLR, and the network sets up the trunk based on this location information. The registration procedure is illustrated in Figure 3 and is described in the following steps.

Step 1.1. Suppose that the home system of a mobile user is in Taiwan. When this mobile user moves from one visited system (e.g., Hong Kong) to another (e.g., London), the user's MS automatically registers in the VLR at London. Note that the

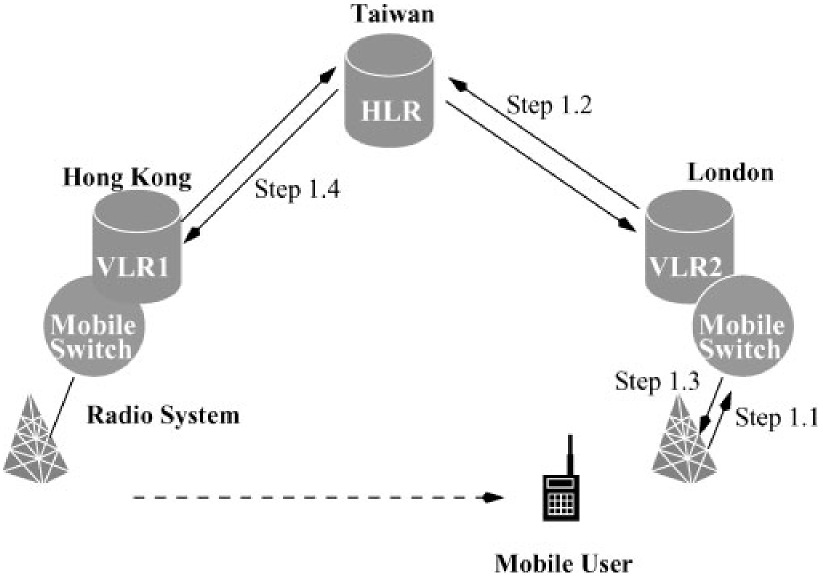

Fig. 3. Mobile station (MS) registration process.

radio base stations connected to the mobile switching center (MSC) are partitioned into several location areas (LA). To simplify our discussion, we assume that there is one location area per MSC. In registration, the addresses of the MSC and location area where the MS resides are sent to the VLR.

Step 1.2. The new VLR then informs the mobile user's HLR of its current location, i.e., the address of the new VLR. The HLR sends an acknowledgment, which includes the user's profile, to the new VLR.

Step 1.3. The new VLR then creates a record for the visiting user to store the profile received from the HLR. Then the VLR informs the MS of successful registration.

Step 1.4. After Step 1.2, the HLR also sends a deregistration message to cancel the obsolete record of the MS in the old VLR at Hong Kong. The old VLR acknowledges the deregistration.

To originate a call, the following steps shown in Figure 4 are executed:

Step 2.1. The MS first contacts the MSC in the visited mobile telecommunications network.

Step 2.2. The call request is forwarded to the VLR for approval. For example, the user profile may indicate that the user is not allowed to make international calls. Thus, any attempt to make international call will be rejected by the VLR.

Step 2.3. If the call is accepted, the MSC sets up the call to the called party following the standard PSTN (public switched telephone network) call setup procedure.

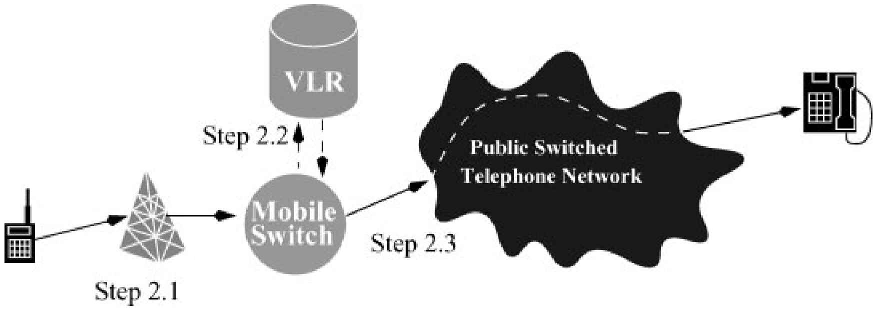

Fig. 4. MS call origination.

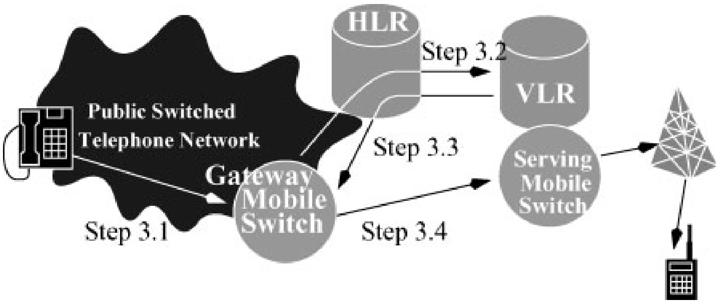

Fig. 5. MS call termination procedure.

The call delivery (or call termination) procedure to an MS is illustrated in Figure 5 and it is discussed in the following steps:

Step 3.1. If someone attempts to call a mobile subscriber, the call is forwarded to a gateway MSC.

Step 3.2. The gateway MSC queries the HLR to find the current VLR of the MS. Then the HLR queries the VLR of the MS to obtain a routable address.

Step 3.3. The VLR searches the record for the called mobile subscriber. Based on the location information, the VLR creates the routable address and returns it to the gateway MSC through the HLR.

Step 3.4. Based on the routable address, a trunk (voice circuit) is set up from the originating switch to the serving MSC. The serving MSC queries the VLR to find the location area of the MS. The radio base stations in the location area then page the MS and the call path to the MS is established.

Details of mobility management and call setup procedures can be found in Reference [15]. Many studies have focused on issues regarding normal mobile registration/call setup [14,33] and failure restoration [16,17]. We will describe some advanced core network mobility issues in Section 3.

2. Radio Network Mobility

This section discusses mobility issues on the radio network. Subsection 2.1 shows how the SRNC is relocated for a moving MS in communication session. Subsection 2.2 describes how high-speed downlink transmission can be maintained when a communicating MS moves across several cells.

2.1. SRNC Relocation

In 3GPP TS 23.060 [7], a lossless SRNC relocation procedure was proposed for non-real-time data services. In this approach, in the beginning of SRNC relocation, the source RNC (RNC1 in Figure 2(b)) first stops transmitting downlink data packets to the MS. Then it forwards the next packets to the target RNC (RNC2 in Figure 2(b)) via a GTP tunnel between the two RNCs. The target RNC stores all IP packets forwarded from the source RNC. After taking over the SRNC role, the target RNC restarts the downlink data transmission to the MS. In this approach, no packet is lost during the SRNC switching period. Unfortunately, this approach does not support real-time data transmission because the data traffic will be suspended for a long time (about 100 ms) during SRNC switching. In order to support real-time multimedia services, 3GPP TR 25.936 [8] proposed SRNC duplication (SD) and core network bi-casting (CNB). These two approaches duplicate data packets during SRNC relocation, which may not efficiently utilize system resources. In Reference [39], we proposed an approach called fast SRNC relocation (FSR) to provide real-time SRNC switching without packet duplication. The detailed procedure for FSR is described below.

As shown in Figure 2(b), the UE is connected to the source RNC and SGSN1 before SRNC relocation. After relocation, data packets for the UE are directly routed through the target RNC and SGSN2 as shown in Figure 2(c). Figure 6 illustrates the four stages of the FSR procedure.

Stage I (Figure 6(a)) initiates SRNC relocation, where the routing path of downlink packets is ⟨GGSN→SGSN1→source RNC→target RNC→UE⟩. The following steps are executed in Stage I.

Steps 1 and 2. When the node B of the source RNC no longer connects to the UE, the source RNC initiates SRNC relocation and sends the ID of the target RNC to SGSN1 through the Relocation_Required message.

Step 3. Based on the ID of the target RNC, SGSN1 determines that it is inter-SGSN SRNC relocation.

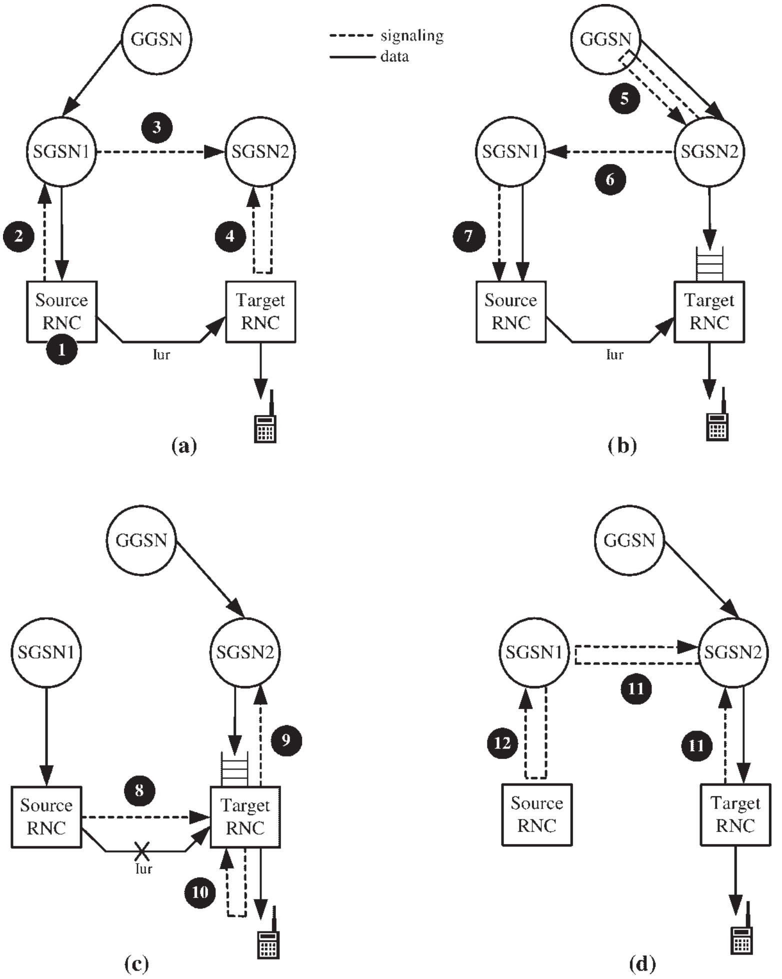

Fig. 6. The fast SRNC relocation (FSR) approach. (a) Stage I. (b) Stage II. (c) Stage III. (d) Stage IV.

SGSN1 requests SGSN2 to allocate the resources for the UE through the Forward_Relocation_Request message.

Step 4. SGSN2 and the target RNC exchange the Relocation_Request and Relocation_Request_Acknowledge message pair to allocate the necessary resources for the UE.

In stage II (Figure 6(b)), the GGSN routes the downlink packets to the old path before receiving the Update_PDP_Context_Request message (Step 5 in Figure 6(b)). The packets delivered through the old path are called 'old' packets. After the GGSN has received the Update_PDP_Context_Request message, the downlink packets are routed to the new path ⟨GGSN→SGSN2→target RNC⟩. The packets delivered by the new path are called 'new' packets. The new packets arriving at the target RNC are buffered until the target RNC takes over the SRNC role. The following steps are executed in stage II.

Step 5. Upon receipt of the Relocation_Request_Acknowledge message, SGSN2 sends a Update_PDP_Context_Request message to the GGSN. Based on the received message, the GGSN updates the corresponding PDP context fields and returns a Update_PDP_Context_Response message to SGSN2. Then the downlink packet routing path is switched from the old path to the new path. At this stage, the new downlink packets arriving at the target RNC are buffered.

Steps 6–7. SGSN2 sends a Forward_Relocation_Response message to SGSN1 to indicate that all resources for the UE are allocated. SGSN1 forwards this information to the source RNC through the Relocation_Command message.

In stage III (Figure 6(c)), the Iur link between the source RNC and the target RNC is disconnected. The old packets arriving at the source RNC later than the Relocation_Command message (Step 7 in Figure 6(b)) are dropped. In this stage, Steps 8–10 switch the SRNC role from the source RNC to the target RNC.

Step 8. With the Relocation_Commit message, the SRNC context of the UE is transferred from the source RNC to the target RNC.

Steps 9 and 10. The target RNC sends a Relocation_Detect message to SGSN2. At the same time, the target RNC sends a RAN_Mobility_Information message to the UE, which triggers the UE to send the uplink IP packets through the new path ⟨UE→target RNC→SGSN2→GGSN⟩.

By executing Steps 11 and 12 at stage IV (Figure 6(d)), the target RNC informs the source RNC that SRNC relocation is successfully performed. Then the source RNC releases the system resources for the UE.

Step 11. The target RNC sends the Relocation_Complete message to SGSN2, which indicates that SRNC relocation is successfully performed. Then SGSN2 exchanges this information with SGSN1 through the Forward_Relocation_Complete and Forward_Relocation_Complete_Acknowledge message pair.

Step 12. Finally, SGSN1 and the source RNC exchange the Iu_Release_Command and Iu_Release_Complete message pair to release the Iu connection in the old path.

Based on the above discussion, Table I compares FSR with SD and CNB. The following issues are addressed.

Packet Duplication. During SRNC relocation, IP packets are duplicated at the source RNC in SD. Similarly, IP packets are duplicated at the GGSN in CNB. Packet duplication will significantly consume system resources. On the other hand, packet duplication is not needed in the FSR approach.

Table I. Comparing FSR with SD and CNB

	FSR	SD	CNB
Packet duplication	No	Yes	Yes
Packet loss at source RNC	Yes	Yes	No
Packet loss at target RNC	No	Yes	Yes
Packet buffering	Yes	No	No
Out-of-order delivery	No	Yes	No
Extra signaling	No	No	Yes

Packet Loss. Packet loss may occur in these three approaches either at the source RNC or at the target RNC. For SD and FSR, data packets arriving at the source RNC may be lost. In SD, the old packets are dropped at the source RNC when the data-forwarding timer expires. In FSR, the old packets are dropped if they arrive at the source RNC later than the Relocation_Command message (see Step 7 in Figure 6(b)) does.

For SD and CNB, data packets may be lost at the target RNC. In SD, the target RNC discards the forwarded packets from the source RNC if these packets arrive at the target RNC earlier than the Relocation_Commit message does. In CNB, duplicated packets may be lost at the target RNC because the packets from the new path are dropped before the target RNC becomes the SRNC. On the other hand, since the packet buffering mechanism is implemented in FSR, the packets are not lost at the target RNC.

Packet Buffering. To avoid packet loss at the target RNC, a packet buffering mechanism is implemented in FSR, which is not found in both SD and CNB approaches.

Out-of-order Delivery. In SD, two paths (i.e., the forwarding and new paths) are utilized to simultaneously transmit the downlink packets. Since the transmission delays for these two paths are not the same, the packets arriving at the target RNC may not be in sequence, which results in out-of-order delivery. On the other hand, this problem does not exist in FSR and CNB because the target RNC in these two approaches only processes the packets from one path (either the old path or the new path) at any time, and the out-of-order packets are discarded.

Extra Signaling. The SD approach follows the standard SRNC relocation procedure proposed in 3G 23.060 [7]. The FSR approach reorders the steps of the 3G 23.060 SRNC relocation procedure. Both approaches do not introduce any extra signaling cost. On the other hand, CNB exchanges additional Update_PDP_Context_Request and Update_PDP_Context_Response message pair between the GGSN and SGSN2, which incurs extra signaling cost. Note that all three approaches can be implemented in the GGSN, SGSN, and RNC without introducing new message types defined in the existing 3GPP specifications.

In conclusion, SD and CNB require packet duplication that will double the network traffic during SRNC relocation. For the SD approach, it is not clear if the Iu link in the forwarding path can be directly established between two RNCs. If not, an indirect path $\langle$source RNC→SGSN1→SGSN2→target RNC$\rangle$ is required. Also, it is not clear if the target RNC will be informed to stop receiving the forwarded packets when the data-forwarding timer expires. Packet duplication is avoided in FSR. We note that packets may be lost during SRNC relocation for these three approaches. Packet loss cannot be avoided in SRNC relocation if we want to support real-time applications. Our performance study indicated that packet loss at the source RNC can be ignored in FSR. Furthermore, the expected number of buffered packets at the target RNC is small, which does not result in long packet delay [39].

2.2. Packet Re-Routing for High-Speed Downlink Packet Access

In UMTS, a UE communicates with cells or the coverage area of Node Bs in an active set through the air interface Uu [1]. If the quality of the wireless link between the UE and a cell is above some threshold, then this cell is included in the active set. When the quality of the wireless link of a cell in the active set is below the threshold, the cell is removed from the active set. Change of cells in the active set is generally due to movement of the UE. In standard UTRAN [6], multiple paths exist between the UE and all node Bs in the active set. This mechanism does not support high-speed downlink transmission because multiple links for a UE may increase the overall interference within an UTRAN, and thus the data transmission rate decreases.

3GPP TR 25.950 [5] proposes a mechanism to support high-speed downlink packet access (HSDPA) [3–5], where a UE only communicates with one cell (called the serving cell) in the active set. This 'serving cell' is selected by the fast cell selection mechanism [4] based on the common pilot channel received signal code power measurements of the cells in the active set. Two physical channels, high-speed-physical downlink shared channel (HS-PDSCH) and dedicated physical control channel (DPCCH) are used for downlink packet frame transmission and uplink/downlink signaling, respectively. While multiple cells may be members of the active set, only one of them transmits at any time in the HSDPA mode. Therefore, the interference within a cell is potentially decreased, and the system capacity is increased. The HSPDA topics include adaptive modulation and coding, hybrid automatic repeat request, packet scheduler, and fast cell selection. This section focuses on the HSDPA buffer overflow control issue due to mobility.

Figure 7 illustrates the network architecture of HSDPA with the active set $\{\mathrm{Cell}_1, \mathrm{Cell}_2, \mathrm{Cell}_3\}$ and the serving cell Cell_1. In HSDPA, the RNC sends the packet frames to all cells in the active set. For the serving cell, the packet frames are forwarded to the UE. For each non-serving cell, the packet frames are queued in a buffer. The Stop-And-Wait Hybrid ARQ (SAW-Hybrid ARQ) [35] algorithm is exercised between the UE and the serving cell for wireless link flow control. If the link quality for high-speed downlink transmission degrades below some threshold, the UE gives up the current serving cell, and the network selects the best cell in the active set as the serving cell. Then the next packet frames are transmitted from the new serving cell to the UE. In HSDPA, the buffer in a non-serving cell may be full, and a mechanism is required to avoid buffer overflow at that non-serving cell. Furthermore, when the UE switches to a new serving cell for downlink packet access, the new serving cell should be informed the status of the buffer (i.e., the number of packet frames received by the UE) in the old serving cell. This action is referred to as frame synchronization [5]. Since the non-serving cells do not send packet frames to the UE, their buffers may overflow. The buffer overflow issue is not addressed in 3GPP TR 25.950. In Reference [26], we proposed four overflow control schemes basic frame synchronization (BFS), network frame synchronization (NFS), and combined BFS and NFS (CFS).

In BFS, the information needed for frame synchronization is carried by the uplink DPCCH. When the size of frame synchronization information exceeds the capacity of an uplink DPCCH, this information must be carried through multiple uplink DPCCH transmissions. To avoid multiple HSDPA transmissions, NFS guarantees one uplink DPCCH transmission by exchanging frame synchronization information between the old and new serving cells without involving the UE. The CFS takes advantage of both BFS and NFS, where the old serving cell decides whether to transmit frame synchronization information through the network or uplink DPCCH. In Reference [24], we constructed analytical and simulation models to investigate the delays of frame synchronization for the

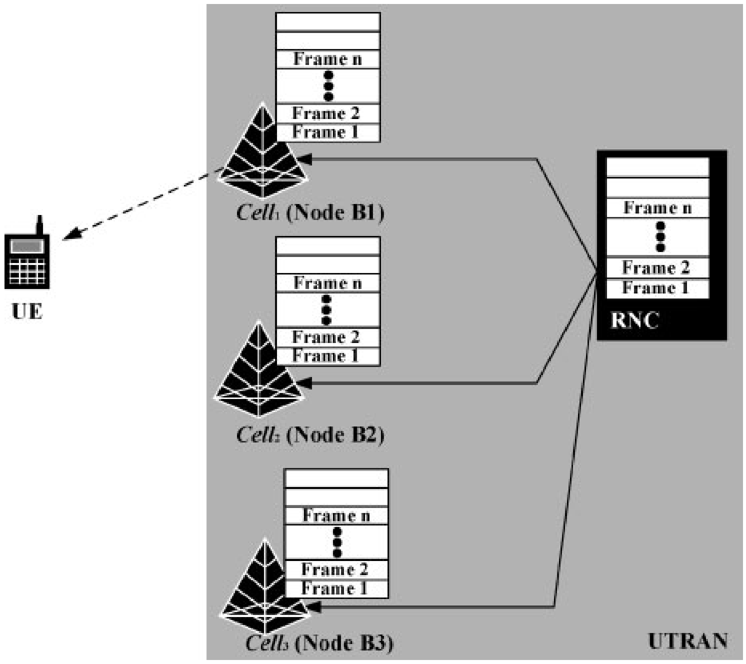

Fig. 7. High-speed downlink packet access (HSDPA) network architecture. RNC: radio network controller. UE: user equipment. UTRAN: UMTS terrestrial radio access network.

four schemes. We showed how traffic patterns (the session time and the packet frame arrival rate) and the serving cell residence time (the period in which the UE resides in a serving cell and receives downlink packet frames from that cell) affect the performance of BFS, NFS, and CFS. Specifically, when the UE switches from the jth cell to the $j+1$st cell during a communication session, we compute the number $n_{r,j}$ of messages delivered in the radio interface (between the UE and the cells) and the number $n_{n,j}$ of messages delivered in the network (among the new serving cell, the old serving cell, and the RNC). Let d_r and d_n be the expected transmission delays in the radio interface and the network, respectively. Then the net cost $d_{s,j}$ for frame synchronization and cell switching to the $j+1$st cell can be expressed as

$$d_{s,j} = n_{r,j}d_r + n_{n,j}d_n \qquad (1)$$

Let packet arrivals to the UE in a downlink transmission session be a Poisson stream with rate λ. Let the serving cell residence time of UE be a random variable with the expected value $1/\mu$. Figure 8(a) plots $d_{s,j}$ as functions of λ for the BFS, NFS, and CFS algorithms, where $d_r = 5d_n$, $N^* = 1024$, and $j = 3$. In this figure, both the packet inter-arrival times and the cell residence times are exponentially distributed. We observe that CFS outperforms BFS and NFS in terms of $d_{s,j}$ delay. When λ/μ is small (i.e., $\lambda/\mu < 500$), $d_{s,j}(\text{CFS}) \approx d_{s,j}(\text{BFS}) \ll d_{s,j}(\text{NFS})$. When $\lambda/\mu = 100$, CFS has 17% improvement over NFS in terms of the $d_{s,j}$ performance. As λ/μ becomes large (i.e., $\lambda/\mu \geq 500$), $d_{s,j}(\text{CFS}) \approx d_{s,j}(\text{NFS}) \ll d_{s,j}(\text{BFS})$. When $\lambda/\mu = 1100$, CFS has 30% improvement over BFS in terms of the $d_{s,j}$ performance. Figure 8(b) show the impact of standard derivation σ of the cell residence times on the CFS $d_{s,j}$ performance. The figure indicates that when λ/μ is small, smaller σ values yield better performance. On the other hand, when λ/μ is large, the result reverses. Therefore, the moving pattern of a UE has significant impact on the high-speed downlink transmission. Several improvements have been made to further improve the performance of the frame synchronization algorithms (see References [24–26]).

3. Core Network Mobility

Several studies have devoted to roaming management (see References [9,10,36] and the references therein). This section describes advanced issues on core network related mobility. Specifically, Subsection 3.1 shows how to select the sizes of a location or a routing area so that core network mobility management can

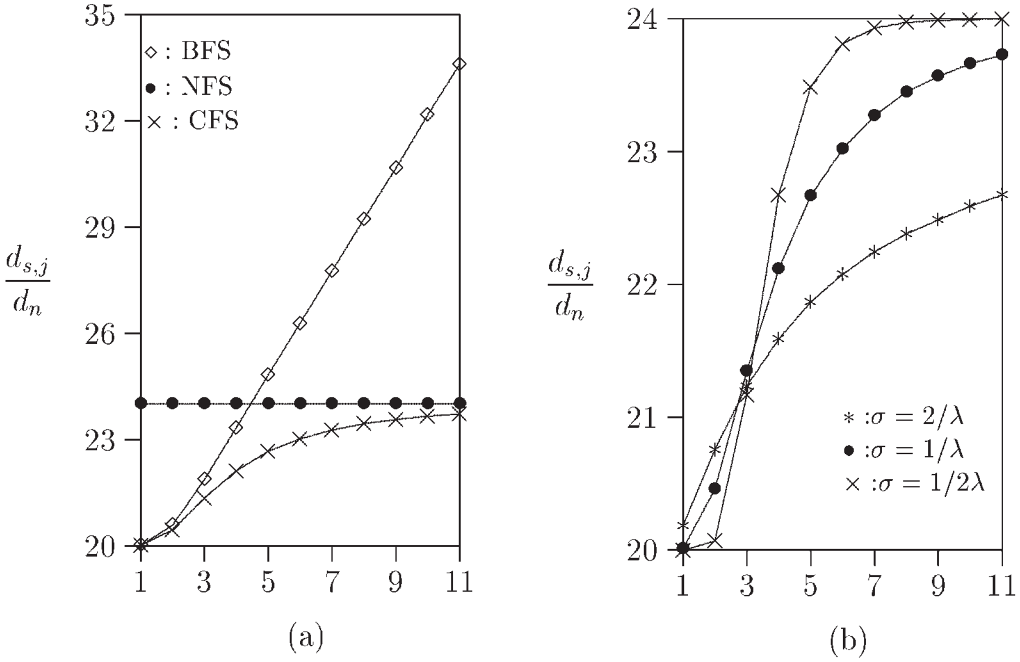

Fig. 8. The $d_{s,j}$ performance for BFS, NFS, and CFS ($d_r = 5d_n$; $N^* = 1024$; $j = 3$). (a) Comparison for CFS BFS and NFS. (b) Effects of cell residence time variance on CFS.

be effectively exercised. Subsection 3.2 demonstrates that mobility of users allows detection of fraudulent usage. Subsection 3.3 presents a VLR overflow technique to allow operator to provide services to new coming users when the VLR is full. The same technique applies to SGSN overflow for data services. Subsection 3.4 describes checkpointing techniques for HLR backup. Backup operation is required to recover user mobility information when an HLR fails. Other issues related to HLR and VLR failure are discussed in Reference [30] and will not be elaborated here.

3.1. Routing Area Selection

General packet radio service (GPRS) provides packet switched data services for existing mobile telecommunications networks such as GSM and Digital AMPS [30]. Most GSM-based mobile operators are deploying GPRS for wireless internet services. The network architecture of GSM/GPRS is shown in Figure 9. In this figure, the dashed lines represent signaling links, and the solid lines represent data and signaling links. The core network consists of two service domains, a circuit-switched (CS) service domain (i.e., PSTN/ISDN) and a packet-switched (PS) service domain (i.e., IP). GPRS introduces two new core network nodes: serving GPRS support node (SGSN) and gateway GPRS support node (GGSN). Existing GSM nodes including base station subsystem (BSS), VLR, and HLR are upgraded. GPRS BSS consists of base transceiver stations (BTSs), and base station controller (BSC) where the BSC is connected to the SGSN through frame relay link. The BTS communicates with the MS through the radio interface *Um* based on the TDMA technology.

The cells (i.e., radio coverages of BTSs) in a GPRS service area are partitioned into several groups. To deliver services to an MS, the cells in the group covering the MS will page the MS to establish the radio link. In the CS domain, cells are partitioned into LAs. The LA of an MS is tracked by the VLR. In the PS domain, the cells are partitioned into routing areas (RAs). An RA is typically a subset of an LA. The RA of an MS is tracked by the SGSN. The SGSN also tracks the cell of an MS when packets are delivered between the MS and the SGSN.

In GPRS, mobility management activities for an MS are characterized by an mobility management (MM) finite state machine exercised in both the SGSN and the MS. There are three states in the machine. In the IDLE state, the MS is not known (i.e., not attached) to GPRS. In the STANDBY state, the MS is attached to GPRS and the MS is tracked by the SGSN at the RA level. In the READY state, the SGSN tracks the MS at the cell level. Packet data units can only be delivered in this state. Descriptions of

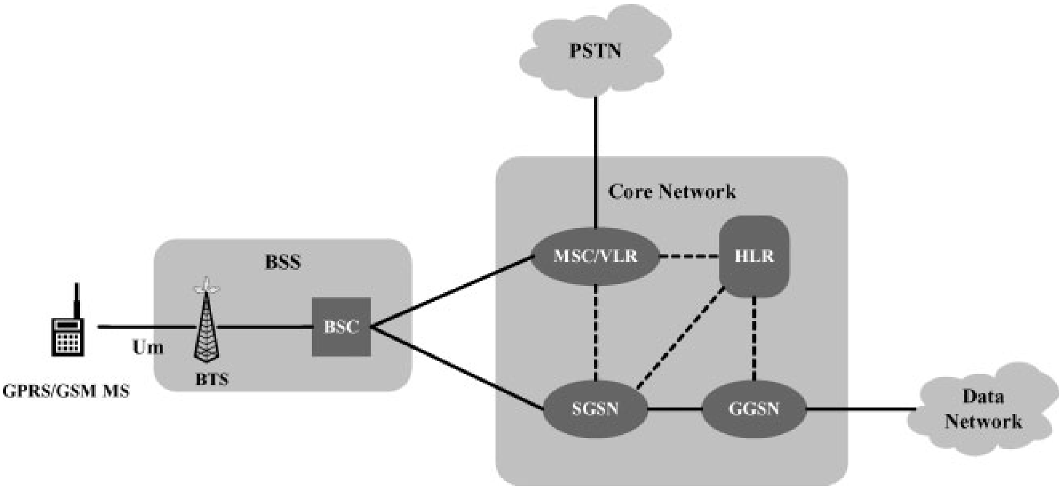

Fig. 9. The network architecture of GSM/GPRS. BSS: base station subsystem. HLR: home location register. Ms: mobile station. VLR: visitor location register. BTS: base transceiver station. GGSN: gateway GPRS support node. MSC: Mobile switching cneter. PSTN: public switched telephone network. SGSN: serving GPRS support node.

transitions among the MM states can be found in Reference [30] and are briefly described as follows.

T1. IDLE → READY. This transition is triggered by an MS when the MS performs GPRS attach.

T2. READY → IDLE. This transition is triggered by the MS or the SGSN when the MS is detached from the GPRS network.

T3. STANDBY → READY. This transition occurs when the MS sends a packet data unit to the SGSN, possibly in response to paging from the SGSN.

T4. READY → STANDBY. This transition is triggered by either the SGSN or the MS. In GPRS, a READY timer is maintained in the MS and the SGSN. If no packet data unit is transmitted before the timer expires, then this MM transition occurs. The length of the READY timer can only be changed by the SGSN. The MS is informed of the READY timer value change through messages such as Attach Accept and Routing Area Update Accept. This MM transition may also occur when the SGSN forces to do so, or when abnormal condition is detected during radio transmission.

T5. STANDBY → IDLE. This transition is triggered by the SGSN when tracking of MS is lost. This transition may also be triggered by SGSN when the SGSN receives a Cancel Location message from the HLR, which implies that the MS has moved to the service area of another SGSN.

Transition T4 merits further discussion. In the READY state, the MS expects to receive packets in short intervals. Therefore, when the MS moves to a new cell, it should inform the SGSN of the movement immediately. In this way, the SGSN can deliver the next packet to the destination cell without paging the whole RA. On the other hand, if the communication session between the MS and the SGSN completes, the SGSN may not send the next packet (the first packet of the next session) to the MS in a long period. In this case, tracking the MS at the cell level is too expensive. Thus the MM state should be switched to STANDBY and the MS is tracked at the RA level. To conclude, in the READY state, no paging is required (the packets are sent directly to the MS) while the location update cost is high (location update is performed for every cell movement). In the STANDBY state, the paging cost is high (all cells in the RA are paged) while the location update cost is low (location update is performed for every RA movement). The T4 transition can be implemented by two approaches. In the READY Timer (RT) approach [7], an RT threshold T is defined. At the end of a packet transmission, the RT timer is set to the T value, and is decremented as time elapses. Transition T4 occurs if the MS does not receive the next packet before the RT timer expires.

The RT approach has a major fallacy that the RT timers in both the MS and the SGSN may lose synchronization (i.e., when the SGSN moves to STANDBY, the MS may be still in READY). To resolve this problem, we consider the READY Counter (RC) approach. In the RC approach, an RC counter counts the number of cell movements in the packet idle period between two packet transmissions to an MS. If the number of movements reaches a threshold K, then the MM state switches from READY to

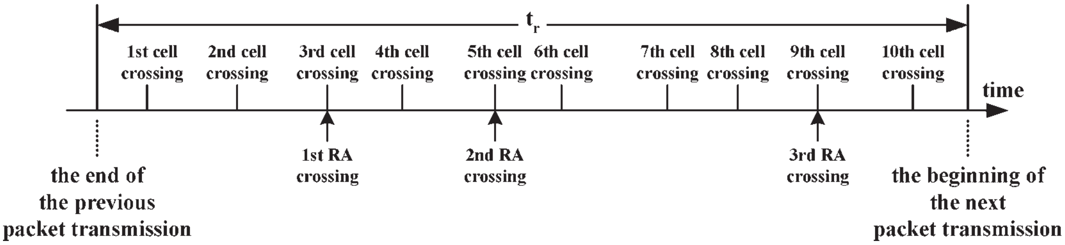

Fig. 10. Cell and RA crossings in an idle period.

STANDBY. To capture user mobility more accurately, one may dynamically adjust the K value to further reduce the net cost of cell/RA updates and paging. Consider the interval t_r between the end of a packet transmission and the beginning of the next packet transmission. If we know the number of cell crossings in t_r and the distribution of RA crossings among these cell crossings, then we can find the optimal K value such that the net cost is minimized. Let N_c be the number of cell crossings during t_r. Let $N_r(j)$ be the number of RA crossings occurring between the $j+1$st cell crossing and the N_cth cell crossing, where $j < N_c$. By convention, $N_r(K) = 0$ for $K \geq N_c$. Figure 10 illustrates the cell and RA crossings in an idle period. In this example, $N_c = 10$. If $K = 2$, then $N_r(2) = 3$. If $K = 4$, then $N_r(4) = 2$. Let U be the cost for a cell/RA update, and V be the cost for paging in a cell. Let S be the number of cells in an RA. Consider the RC algorithm with threshold K. The net cost $C_T(K)$ in an idle period can be expressed as

$$C_T(K) = \begin{cases} UN_c, & \text{for } K > N_c \\ U[K + N_r(K)] + SV, & \text{for } K \leq N_c \end{cases} \tag{2}$$

With the following theorem, we show how to find the optimal threshold value K^* for RC in an idle period such that the net cost is minimized.

Theorem. Consider an idle period where no packet is delivered. Let N_c be the number of cell crossings in this period. In the RC algorithm, let K^* be the optimal threshold value that minimizes the net cost $C_T^* = C_T(K^*)$ in the idle period. Then $K^* = 0$ or $K^* = N_c + 1$.

The proof of this theorem can be found in Reference [31]. Based on this theorem, we devised an algorithm to select K as follows. Let $t_r(i)$ be the interval between the end of the $i-1$st packet transmission and the beginning of the ith packet transmission. Let $K(i)$ be the optimal K value for $t_r(i)$. The K value can be dynamically adjusted using the following algorithm.

Dynamic READY Counter (DRC) Algorithm

Initialization: Assign an arbitrary value to $K(0)$. Exercise the RC approach with threshold $K(0)$ before the first packet arrives.

When the ith Packet Transmission is Completed: Compute the optimal $K(i)$ that minimizes the net cost for the period $t_r(i)$. Based on the above Theorem, $K(i)$ is either 0 or $N_c + 1$ in $t_r(i)$, which can be quickly computed with very low cost. Exercise the RC approach with threshold $\bar{K}$ during $t_r(i+1)$, where

$$\bar{K} = \begin{cases} \left\lceil \sum_{j=i-M+1}^{i} \frac{K(j)}{M} \right\rceil, & \text{for } i \geq M \\ \left\lceil \sum_{j=1}^{i} \frac{K(j)}{i} \right\rceil, & \text{for } i < M \end{cases} \tag{3}$$

In other words, the threshold $\bar{K}$ between the ith and the $i+1$st packet transmissions is selected as the average of the previous M optimal K values. Simulation experiments show that $M \geq 5$ is appropriate when using Equation (3) to compute $\bar{K}$.

Since the MS has complete information on the numbers of cell and RA crossings in t_r, the Dynamic RC (DRC) algorithm is implemented in MS. In DRC, the state transition of SGSN (i.e., from READY to STANDBY) is triggered by the MS. When the MS crosses the $\bar{K}$th cell boundary during t_r, it sends an RA update message to the SGSN instead of the cell update message. Once the SGSN receives the first RA update message from the MS, it switches the MM state from READY to STANDBY. No extra message is introduced to switch the MM states in DRC. Thus, network signaling cost is the same as that of the static RC mechanism. The only cost incurred by DRC is the computation of $\bar{K}$ in the MS. As shown in Equation (3), the computation can be done in micro seconds, which is not significant and can be ignored.

We have investigated how the RA size and the location update/paging costs affect the performance of these approaches [31]. Specifically, we showed that RC is better than RT, and that DRC can automatically adjust the K value to minimize the net cost.

In UMTS, three-layer structure is employed for location tracking, including cell-level tracking, UTRAN RA (URA) tacking, and RA tracking. The technique described in this section can be extended for UMTS three-layer tracking [38].

3.2. Fraudulent Usage Due to Mobility

As the penetration rate of mobile services significantly increases, fraudulent usage has become a serious issue. At the mobile subscriber side, modern mobile services such as GSM and UMTS are secured by the subscriber identity module (SIM). Without the SIM card, an MS cannot access mobile services (except for emergency calls). The SIM can be a smart card that typically has the size of a credit card, a smaller sized 'plug-in SIM,' or a smart card that can be performed, which contains a plug-in SIM that can be broken out of it. A SIM contains the subscriber-related information including its identity, authentication key, and encryption key. The subscriber identity and authentication key are used to authorize mobile service access, and to avoid fraudulent access of a cloned MS. Through interaction between the SIM and the authentication center (AuC), GSM provides one-way authentication. In UMTS, mutual authentication is achieved by sharing a secret key between the SIM and the AuC [2]. This procedure follows a challenge/response protocol identical to the GSM subscriber authentication and key establishment protocol [30] combined with a sequence-number based one-pass protocol for network authentication derived from ISO/IEC 9798–4 [21]. The above authentication procedures assume that the SIM is securely kept by the corresponding legal mobile user. Unfortunately, we have observed the existence of cloned SIMs that also have the same identities and authentication keys as the legal SIMs. These cloned SIMs will result in fraudulent usage as well as mis-routing of incoming calls to the legal users. In this subsection, we show how mobility of users allows detection of fraudulent usage.

In GSM/UMTS, authentication is typically performed in the following events: registration (MS attach or location update), call origination (i.e., an outgoing call), and call termination (i.e., an incoming call). An illegal user gains access to the mobile telecommunications network through registration and then enjoys 'free' services through call originations. Such fraudulent usage may be detected by the GSM/UMTS network when the legal user performs the next registration or call origination.

Suppose that the SIM card of a legal user u has been cloned by an illegal user u^*. In Figure 11, u enters a location area LA$_1$ at time τ_0, and moves to another location area LA$_2$ at time τ_3. Following the standard GSM/UMTS registration procedure (See Subsection 1.2), u informs the network that it has moved into LA$_1$ at τ_0, and informs that it has left LA$_1$ and moved into LA$_2$ at τ_3. In this procedure, authentication is exercised between u and the AuC (through VLR1 and the HLR not shown in Figure 3). Assume that authentication is successful, and Steps 1.1–1.4 in Subsection 1.2 are executed. After execution, VLR1 maintains a record for u where the location field indicates LA$_1$. Then the MS registers to VLR1 and the HLR.

After registration, the HLR record of u indicates that the MS is in VLR1, and the record in VLR1 indicates that the MS is in LA$_1$. Suppose that u^* attempts to illegally access the network at time τ_1 in the interval $[\tau_0, \tau_3]$. This illegal user u^* either issues an attach or a normal location area update action at τ_1. If u^* and u are in the same location area LA$_1$, then this abnormal action will be detected immediately; that is, the network notices that the same MS registers to the same location area twice. The network may suspend the services for u until the fraudulent usage issue is cleared. If u^* registers in a different location area LA$_3$ (controlled by VLR2 in our example), then the network considers that u has moved from LA$_1$ to LA$_3$. As illustrated in Figure 12, after the registration, the HLR record indicates that u is in VLR2, and the VLR2 record indicates that u is in LA$_3$. The VLR1 record for

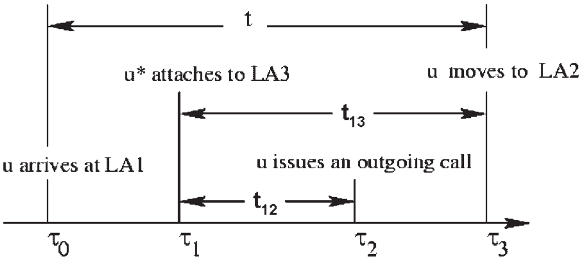

Fig. 11. The timing diagram for fraudulent usage.

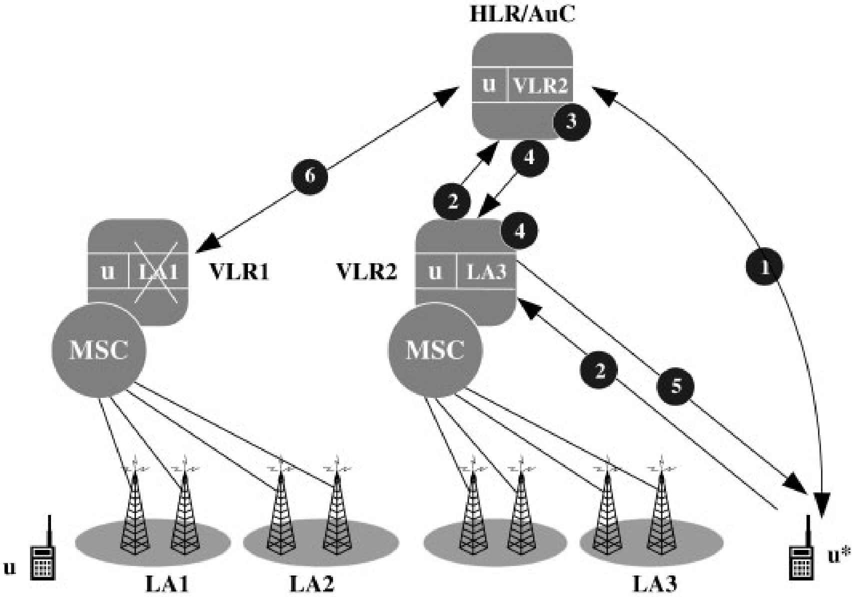

Fig. 12. Registration of the illegal user u^* at time τ_1. (1) Authentication is exercised between u^* and the AuC (through VLR2 and the HLR). Since u^* has the same SIM card as u, authentication is successfully performed, and Steps 2–5 are executed. (2) VLR2 forwards the registration request from u^* to the HLR. (3) The HLR updates the HLR record of u. (4) The HLR informs VLR2 that the registration is successful. VLR2 updates the record content of u. (5) VLR2 informs u^* that the registration operation is successful. (6) The HLR performs location cancellation to delete the VLR record of u in VLR1.

u is deleted through location cancellation. At this point, u^* gains mobile service access, and fraudulent usage may occur. Such fraudulent usage can continue until either there is an outgoing call originated by u at time τ_2 or when u moves to another location area LA_3 and issues location update at time τ_3. Therefore two cases may occur.

Case I. Suppose that after τ_1, the first outgoing call for u occurs before u moves to LA_2 (i.e., $\tau_2 < \tau_3$). Then at τ_2, u issues a call origination request to VLR1 as illustrated in Figure 13. Since u's record at VLR1 has been erased or modified (in Figure 13, the record has been erased), the network will detect that u has registered with location area LA_3 but is issuing an outgoing call from location area LA_1. At this point, the mobile service of u will be suspended until the network has resolved the fraudulent usage issue.

Case II. If $\tau_3 < \tau_2$, then after τ_1, the first event of u is a registration request due to movement from LA_1 to LA_2 (assume that both location areas are covered by VLR1). Similar to the situation in Case I, when u issues the registration request to VLR1, VLR1 cannot identify the record for u, and the potential fraudulent usage situation is detected.

In Figure 11, let $t_{13} = \tau_3 - \tau_1$ and $t_{12} = \tau_2 - \tau_1$. Then $T = \min(t_{13}, t_{12})$ is the potential fraudulent usage period. During this period, u^* can illegally originate calls. Also, all incoming calls to u will be directed to u^* during T; i.e., these calls are mis-routed as illustrated in Figure 14. The calling parties of these mis-routed calls may misleadingly think that the mobile operator accidentally routes the calls to wrong places.

In Reference [34], we investigated the fraudulent usage of mobile services due to cloned SIM. The output measures considered are the expected potential fraudulent usage period and the probability of mis-routed calls in this period. Our study indicated that

- If the legal users exhibit more irregular movement patterns (i.e., the LA residence times with large variance), then fraudulent usage is more likely to occur.
- If the legal users often make outgoing phone calls, then the network has better opportunity to detect the fraudulent usage.
- Incoming calls to the legal user are mis-routed during the potential fraudulent usage period. The probability of mis-routed calls increases as the movement pattern of a legal user becomes more irregular.

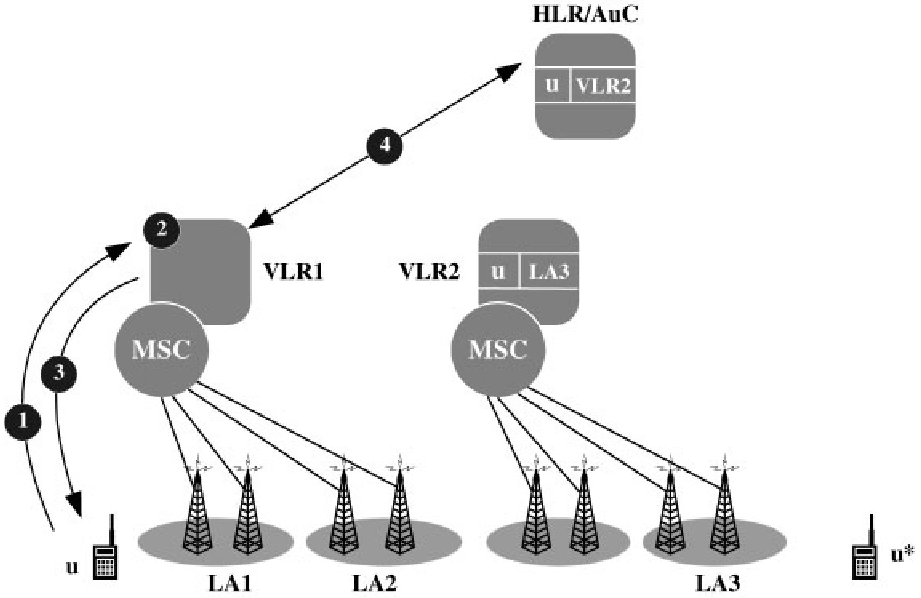

Fig. 13. An outgoing call of u at time τ_2. (1) u issues call origination request to VLR1. (2) VLR1 attempts to initiate the authentication procedure, but cannot locate u's VLR record (because the record has been erased at time τ_1). (3) VLR1 rejects the call origination request. (4) VLR1 informs the HLR that fraudulent usage may occur. The HLR takes necessary actions to resolve the issue.

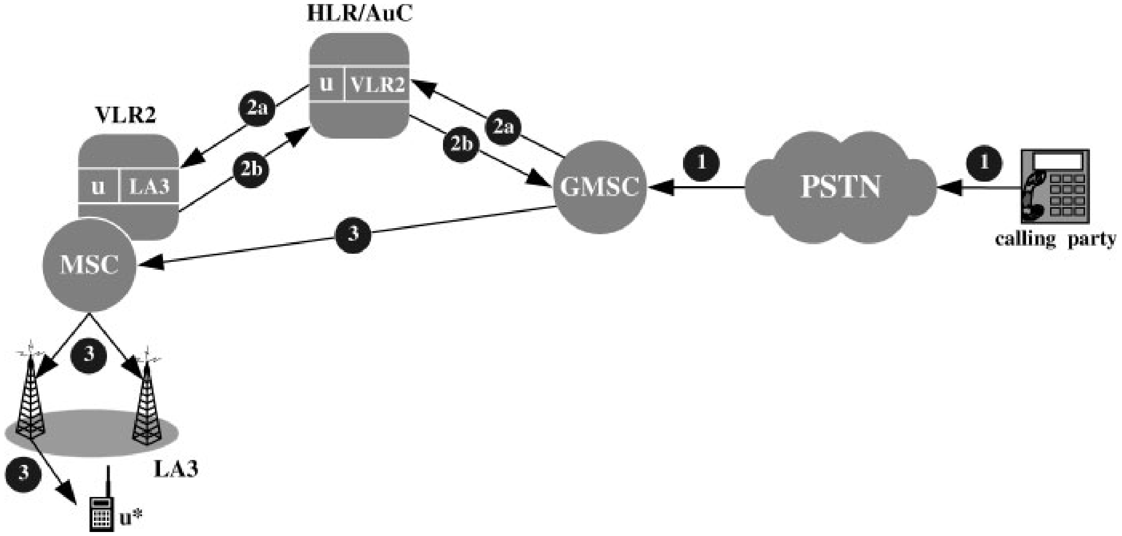

Fig. 14. An incoming call to u in period $[\tau_1, \tau_3]$. (1) The calling party dialed u's phone number. The call is connected to the GMSC (gateway mobile switching center). (2) The GMSC queries the HLR and then VLR2, and determines that u is connected to the MSC of VLR2. (3) The call is directed to u^* through the MSC of VLR2. This call is mis-routed.

- Periodic LA update (PLAU) can effectively speed up the fraudulent usage detection, and also reduce the probability of mis-routed calls. Therefore by exercising PLAU, we can enhance security at the cost of increasing the signaling overhead.

As a final remark, the HLR/AuC can determine whether an abnormal event is due to fraudulent access of an illegal user or not. When the HLR/AuC receives a report of potential fraudulent usage from the VLR, several approaches can assist to determine if the reported case is actually caused by fraudulent usage. For example, in the rare event approach, the MS is requested to include statistics of some rare events (such as the timestamp of last registration, the number of calls made in the last location area, and so on) when it issues location update or call origination requests. The network can compares these statistics received from the MS with that obtained from network OA&M (Operations, Administrations, and Management).

3.3. Mobility Database Overflow

In a mobile telecommunications network, the VLRs are used to temporarily hold the subscription information of the roaming users who visit the service area of the VLR. When the users leave the VLR area, the corresponding records in the VLR are deleted. Due to user mobility, the capacity of the VLR may not be large enough to hold information for all visitors in the VLR area at some time periods. This issue is called VLR overflow. A VLR database overflows if the number of visiting customers exceeds the capacity of the VLR database. In this case, the incoming visitors cannot register using the standard registration procedure described in Steps 1.1–1.4 of Subsection 1.2, and thus cannot receive mobile services. Note that the HLR does not have the database overflow problem. The number of subscriber records in the HLR is known for an operator, which is the number of customers subscribing to the services of that specific operator. Thus the HLR database capacity can be scaled under control, and database overflow never occurs. On the other hand, the number of records in a VLR changes dynamically. This size increases when registrations occur and decreases when deregistrations occur. It is possible that many users enter a VLR in a short period. If the number of users in a VLR area is larger than the capacity of that VLR database, then the VLR database overflows, and the incoming users cannot successfully perform registration. In this case, these users will not be able to receive services, and are referred to as overflow users.

In References [20,27,28], we proposed an approach to resolve VLR overflow issue. In this approach, when a VLR overflows, the visited system still can provide services to incoming users. Our approach takes advantage of the distributed database structure of mobile telecommunications network where the subscription information of a user is duplicated in both HLR and VLR. For overflow users (i.e., the users who do not have records in the VLR), call setup can be complete by using the information stored in the HLR. However, extra cost is required in the call setup procedure of an overflow user. The VLR overflow resolution modifies mobile registration, call origination and call delivery procedures as follows.

Overflow Registration. Suppose that the MS of user u initiates the registration procedure. At Step 1.3 in Subsection 1.2, if the VLR is full, then a record is selected for replacement. That is, an existing record in the VLR is deleted and the reclaimed storage is used to hold the record of u. In this case, the user of the replaced record becomes an overflow user.

Alternatively, user u may be considered as the overflow user, and no record replacement occurs. In this case, Steps 1.2–1.4 in Subsection 1.2 are executed as before except that in Step 1.3 in Subsection 1.2, no record for u is created.

At Step 1.4 in Subsection 1.2, if u is an overflow user at the old VLR, then no record cancellation occurs at that VLR.

Overflow MS Call Origination. When an overflow user u attempts to make an outgoing call, the VLR notices that no record exists for u at Step 2.2 in Subsection 1.2. The VLR will request u to perform an overflow registration operation to create a record for u. (In this registration, u cannot be selected as the overflow user.) Then normal call origination procedure is executed to set up the call.

Overflow MS Call Delivery. For an incoming call to an overflow user u, the VLR cannot find the record for u and thus cannot generate a routable address at Step 3.3 in Subsection 1.2. In this case, the HLR will generate a routable address based on its knowledge of u's location [15]. Through a replacement at Step 3.3 in Subsection 1.2, the VLR creates a record for u to store subscription data as well as location information.

For an overflow user, the costs of executing Steps 2.2 and 3.3 in Subsection 1.2 are higher than that for the standard GSM/UMTS procedure (that is, an extra registration operation is required). Therefore, it is desirable to reduce these extra overheads. One possibility is to select an 'inactive' record for replacement so that the corresponding overflow user does not have any call activity before the user leaves the VLR area. In Reference [27], we consider the random replacement policy where every record in the VLR is selected for replacement with the same probability. The performance of the random replacement policy is acceptable in a homogeneous environment where the mobile users have relatively low call activities. In reality, call activities of visiting users in a VLR area may vary significantly. If a record with low call activity is selected for replacement at Step 1.3 in Subsection 1.2, then it is more likely that the overflow user leaves the VLR without creating any call activity. To achieve this goal, we proposed the inactive replacement policy [28] that attempts to select records with low call activities for replacement. A period called inactive threshold is utilized to determine if a user is not active. If a user does not have any call activity during the inactive threshold, then he/she is considered inactive and the VLR record can be selected for replacement. With record replacement, we can provide mobile services to more users than that can be accommodated in a VLR It is important to select 'appropriate' VLR records for replacement to reduce the possibility of overflow operations in the future. In Reference [28], we compared the inactive replacement policy with the random replacement policy. Our study indicated that the inactive replacement policy significantly outperforms the random replacement policy (by reducing over 90% of the overflow call setups).

3.4. HLR Checkpointing for Failure Restoration

In roaming management, all permanent subscriber data are stored in the HLR. An HLR record consists of three types of information: MS Information such as the telephone number and the International Mobile Subscriber Identity (used by the MS to access the network); Service Information such as service subscription, service restrictions, and supplementary services; Location Information such as the addresses of the VLR and the SGSN where the MS resides. The location information in the HLR is updated whenever the MS moves to a new SGSN. To access the MS, the HLR is queried to find the current SGSN location of the MS. Note that both the MS and service information items are only occasionally updated. On the other hand, an MS may move frequently and the location information is often modified. Details of HLR operations due to call delivery are described in Subsection 1.2.

If the HLR fails, one will not be able to access the MSs. To guarantee service availability to the MSs, database recovery is required after an HLR failure. In UMTS/GPRS [7], the HLR recovery procedure works as follows: The HLR database is periodically checkpointed. After an HLR failure, the database is restored by re-loading the backup information. There are several approaches to checkpointing the HLR database. In the all-record checkpoint approach, all HLR records are saved into the backup at the same times [18]. The checkpoint overhead for this approach is very high and is typically performed at midnight when the HLR activities are infrequent. Alternatively, checkpointing can be exercised for individual mobile users, which is referred to as per-user checkpointing [12,22,32,37]. We describe two algorithms for the per-user checkpoint approach. The first algorithm (referred to as Algorithm I) is the same as all-record checkpointing except that the checkpoint frequencies for individual MSs may be different. The second algorithm (referred to as Algorithm II) is a new approach we proposed in Reference [29].

Algorithm I (the Basic Algorithm). For every MS, we define a timeout period t_p. In Figure 15 the t_p timeouts occur at time t_0, t_1, t_2, t_4, and t_9. When this timer expires, checkpoint is performed to save the HLR record of the MS. Therefore, the checkpoint interval t_c is equal to the timeout period t_p. After a failure (see t_6 in Figure 15), the HLR record in the backup database is restored to the HLR. The backup copy is obsolete if the HLR record is updated between the last checkpoint and when the

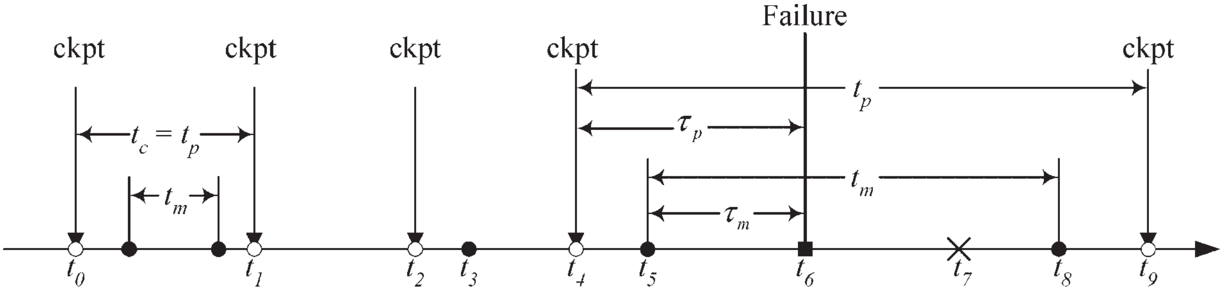

Fig. 15. The timing diagram for Algorithm I. •, represents a registration; ckpt (↓) represents a checkpointing; ×, represents an incoming call; ○ represents a t_p timeout.

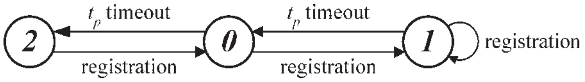

Fig. 16. The state diagram for Algorithm II.

failure occurs (i.e., a registration occurs in $[t_4, t_6]$ in Figure 15). After the HLR record is restored, one of the following two events may occur next:

- The record may be updated again if the MS issues a registration (i.e., $t_8 < t_7$ in Figure 15), or
- the record may be accessed due to an incoming call to the MS (i.e., $t_7 < t_8$ in Figure 15).

After a failure, if the backup record is obsolete and the next event to the MS is an incoming call ($t_7 < t_8$ in Figure 15), then the call is lost. On the other hand, if the next event is a registration, then the location information of the HLR record is modified and the record is up to date again.

Algorithm II (Lin's Algorithm). The intuition behind our algorithm is simple: If registration activities are very frequent (i.e., a registration always occurs before the t_p timer expires), then Algorithm II behaves exactly the same as Algorithm I. On the other hand, if no registration has occurred before the t_p timer expires, then there is no need to checkpoint the record (because the backup copy is still valid). In this case, checkpoint is performed when the next registration occurs.

In Algorithm II, a three-state finite state machine (FSM) is implemented for an HLR record. The state diagram for the FSM is shown in Figure 16. Initially, the FSM is in state 0, and the t_p timer starts to decrement. If a registration event occurs before the t_p timer expires, the FSM moves to state 1, and remains in state 1 until the t_p timeout event occurs. Then the FSM moves back to state 0, the HLR record is checkpointed into the backup, and the t_p timer is re-started. If the timeout event occurs at state 0, then the FSM moves to state 2, and the t_p timer is stopped. If a registration event occurs at state 2, the FSM moves to state 0, a checkpoint is performed, and the t_p timer is re-started.

Consider the timing diagram in Figure 17. At time t_0, the FSM is at state 0 (when a registration occurs). At time t_1, the next registration occurs, and the FSM moves from state 0 to state 1 (where $t_\mathrm{m} = t_1 - t_0$ is the inter registration interval). At time t_2, the t_p timer expires, and the FSM moves from state 1 to state 0 (where $t_\mathrm{c} = t_\mathrm{p} = t_2 - t_0$). At time t_3, the t_p timer expires again, and the FSM moves from state 0 to state 2. At time t_4, a registration occurs. The FSM moves from state 2 to state 0, and $t_\mathrm{c} = \tau_\mathrm{m}^* = t_4 - t_2$ where τ_m^* is the excess life or residual time of t_m.

By utilizing per-user checkpoint, an HLR record is saved into a backup database from time to time. When a failure occurs, the backup record is restored to the HLR. In Reference [29], an analytic model was developed to compare these two algorithms in terms of the checkpoint cost and the probability of obsolete HLR backup record when the record is accessed. Our study indicates that Algorithm II can save more than 50% of the checkpoint cost over Algorithm I. For the performance of α, Algorithm II demonstrates 20–55% improvement over Algorithm I. As a final remark, we note that failure restoration for a SGSN (or a VLR in the circuit switched service domain) is very different from HLR failure restoration described in this paper. No checkpointing is performed for a SGSN because all MS records in the SGSN are temporary, and it is useless to store these temporary records into backup. Details of SGSN failure restoration can be found in [13,16].

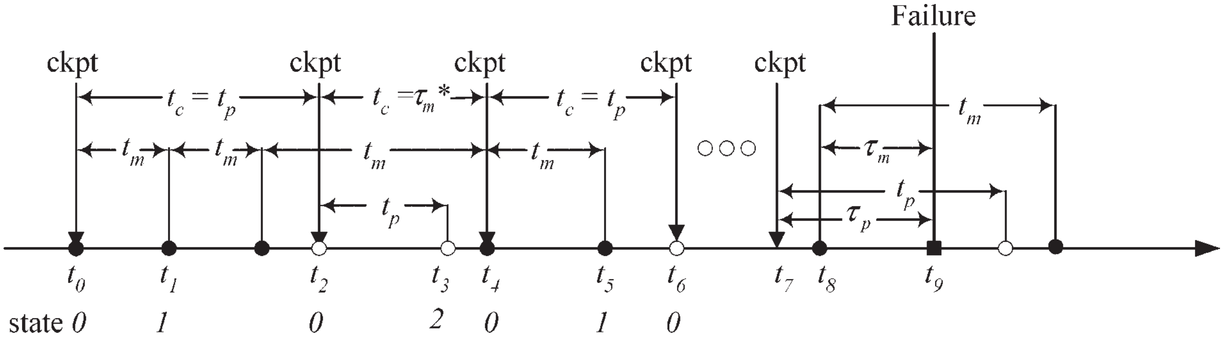

Fig. 17. The timing diagram for Algorithm II. •, represents a registration; ckpt (↓), represents a checkpointing; ×, represents an incoming call; ○ represents a t_p timeout.

4. Summary

This paper described the mobility management mechanisms for mobile telecommunications systems, and discussed impact of mobility on both the radio and the core networks. Radio network mobility supports radio link switching of a mobile user during conversation. We investigated the SRNC relocation, where the downlink packets may be lost or delayed during the SRNC switching period. For real-time multimedia applications, a long packet delay or a large amount of packet loss results in service degradation. We proposed a fast SRNC relocation approach to provide real-time SRNC switching without increasing the network traffic load. Furthermore, we addressed the HSDPA buffer overflow control issue due to mobility, and proposed four overflow control schemes to avoid buffer overflow in the non-serving cells.

Core network mobility provides roaming mobility and tunnel-related management for packet re-routing due to user movement. One of the important issues on core network related mobility is how to select the sizes of the location or the routing area so that mobility management for the core network can be effectively exercised. In this paper, we proposed ready counter (RC) approach, and compared RC with 3GPP TS 23.060 ready timer (RT) approach. Numerical examples showed that RC is better than RT. Then, we investigated the fraudulent usage of mobile services due to cloned subscriber identity module, and demonstrated that mobility of users allows detection of the fraudulent usage. We also presented a VLR/SGSN overflow technique to allow mobile operators to provide services to new coming users when the VLR/SGSN is full. Finally, the checkpointing techniques for HLR failure restoration were presented to recover user mobility information when an HLR fails.

References

1. 3GPP (3rd Generation Partnership Project). Technical Specification Group Radio Access Network; Working Group 2; Radio Interface Protocol Architecture. Technical Specification 3G TS 25.301 version 3.4.0 (2000–03), 2000.
2. 3GPP (3rd Generation Partnership Project). Technical Specification Group Services and Systems Aspects; 3G Security; Security Architecture. Technical Specification 3G TS 33.102 V3.7.0 (2000–12), 2000.
3. 3GPP (3rd Generation Partnership Project). Technical Specification Group Radio Access Network; High Speed Downlink Packet Access; Overall UTRAN Description; Release 5. Technical Report 3G TR 25.855 version 5.0.0 (2001–09), 2001.
4. 3GPP (3rd Generation Partnership Project). Technical Specification Group Radio Access Network; Physical Layer Aspects of UTRA High Speed Downlink Packet Access; Release 4. Technical Report 3G TR 25.848 version 4.0.0 (2001–03), 2001.
5. 3GPP (3rd Generation Partnership Project). Technical Specification Group Radio Access Network; UTRA High Speed Downlink Packet Access; Release 4. Technical Report 3G TR 25.950 version 4.0.0 (2001–03), 2001.
6. 3GPP (3rd Generation Partnership Project). Technical Specification Group Radio Access Network; UTRAN Iub Interface: General Aspects and Principles; Release 4. Technical Specification 3G TS 25.430 version 4.1.0 (2001–06), 2001.
7. 3GPP (3rd Generation Partnership Project). Technical Specification Group Services and Systems Aspects; General Packet Radio Service (GPRS); Service Description; Stage 2. Technical Specification 3G TS 23.060 version 4.1.0 (2001–06), 2001.
8. 3GPP (3rd Generation Partnership Project). Technical Specification Group RAN 3; Handovers for real-time services from PS domain. Technical Report 3G TR 25.936 version 4.0.1, 3GPP, 2001.
9. Akyildiz IF, McNair J, Ho JSM, Uzunalioglu H, Wang W. Mobility management in next generation wireless systems. *IEEE Proceedings* 1999; **87**(8): 1347–1385.
10. Wang W, Akyildiz IF. A new signaling protocol for intersystem roaming in next-generation wireless systems. *IEEE Journal on Selected Areas in Communications* 2001; **19**(10): 2040–2052.
11. Barton P. I love mobile phone. *Reader's Digest (Asia Version)*, April 2004.
12. Cao G. Proactive power-aware cache management for mobile computing systems. *IEEE Transactions on Computers* 2002; **51**(6): 608–621.
13. Chang M-F, Lin Y-B, Su S-C. Improving fault tolerance of GSM network. *IEEE Network* 1998; **1**(12): 58–63.
14. Chlamtac I, Liu T, Carruthers J. Location management for efficient bandwidth allocation and call admission control. *IEEE WCNC*, New Orleans, September 1999.
15. ETSI/TC. Mobile application part (MAP) specification, version 7.3.0. Technical Report Recommendation GSM 09.02, ETSI, 2000.
16. Fang Y, Chlamtac I, Fei H. Analytical results for optimal choice of location update interval for mobility database failure restoration in PCS networks. *IEEE Transactions on Parallel and Distributed Systems* 2000; **11**(6): 615–624.
17. Fang Y, Chlamtac I, Fei H. Failure recovery of HLR mobility databases and parameter optimization for PCS networks. *Journal on Parallel and Distributed Computing* 2000; **60**: 431–450.
18. Haas Z, Lin Y-B. On Optimizing the location update costs in the presence of database failures. *ACM/Baltzer Wireless Networks Journal* 1998; **4**(5): 419–426.
19. Holma H, Toskala A. (eds). *WCDMA for UMTS*. John Wiley & Sons: Chichester, 2000.
20. Hung H-N, Lin Y-B, Peng N-F, Yang S-R. Resolving mobile database overflow with most-idle replacement. *IEEE Journal on Selected Areas in Communications* 2001; **19**(10): 1953–1961.
21. ISO/IEC. Information Technology-Security Techniques—Entity Authentication—Part 4: Mechanisms Using a Cryptographic Check Function. Technical Report ISO/IEC 9798-4, ISO/IEC, 1999.
22. Kahol A, Khurana S, Gupta S, Srimani P. An efficient cache management scheme for mobile environment. *IEEE ICDCS*, 2000.
23. Li B, Yin Y, Wong KYM, Wu S. An efficient and adaptive bandwidth allocation scheme for mobile wireless networks based on on-line local parameter estimations. *ACM/Kluwer Journal of Wireless Networks* 2001; **7**(2): 107–116.
24. Lin P, Lin Y-B, Chlamtac I. Modeling frame synchronization for UMTS high-speed downlink packet access. *IEEE Transactions on Vehicular Technology* 2003; **50**(1): 132–141.
25. Lin P, Lin Y-B, Chlamtac I. Module count-based overflow control scheme for UMTS high speed downlink packet access. *IEEE Transactions on Vehicular Technology* 2004; **53**(2): 425–432.

26. Lin P, Lin Y-B, Chlamtac I. Overflow control for UMTS high-speed downlink packet access. *IEEE Transactions on Wireless Communications* 2004; **3**(2): 524–532.
27. Lin Y-B. Overflow control for cellular mobility database. *IEEE Transactions on Vehicular Technology* 2000; **49**(2): 520–530.
28. Lin Y-B. Eliminating overflow for large-scale mobility databases in cellular telephone networks. *IEEE Transactions on Computer* 2001; **50**(4): 356–370.
29. Lin Y-B. Per-user checkpointing for mobility database failure restoration. *IEEE Transactions on Mobile Computing* 2005; **4**(2): 189–194.
30. Lin Y-B, Chlamtac I. *Wireless and Mobile Network Architectures*. John Wiley & Sons, 2001.
31. Lin Y-B, Yang S-R. A mobility management strategy for GPRS. *IEEE Transactions on Wireless Communications* 2003; **2**(6): 1178–1188.
32. Lin Y-B, Lai W-R, Chen J-J. Effects of cache mechanism on wireless data access. *IEEE Transactions on Wireless Communications* 2003; **2**(6): 1240–1246.
33. Lin Y-B, Lai WR, Chen RJ. Performance analysis for dual band PCS networks. *IEEE Transactions on Computers* 2000; **49**(2): 148–159.
34. Lin Y-B, Chen M-F, Rao HC-H. Potential fraudulent usage in mobile telecommunications networks. *IEEE Transactions on Mobile Computing* 2002; **1**(2): 123–131.
35. Lucent. ARQ Technique for HSDPA. Technical Report R2A010021, Lucent.
36. Roos A, Hartman M, Dutnall S. Critical issues for roaming in 3G. *IEEE Wireless Communications Magazine* 2003; **10**(1): 29–35.
37. Shim J, Scheuermann P, Vingralek R. Proxy cache algorithms: design, implementation, and performance. *IEEE Transactions on Knowledge and Data Engineering* 2000; **11**(4): 549–562.
38. Yang S-R, Lin Y-B. Performance evaluation of location management in UMTS. *IEEE Transactions on Vehicular Technologies* 2003; **52**(6): 1603–1615.
39. Pang A-C, Lin Y-B, Tsai H-M, Agrawal P. Serving radio network controller relocation for UMTS All-IP network. *IEEE Journal on Selected Areas in Communications* 2004; **22**(4): 617–629.

Authors' Biographies

Yi-Bing Lin is chair professor and vice president of Research and Development, National Chiao Tung University. His current research interests include wireless communications and mobile computing. Dr Lin has published over 190 journal articles and more than 200 conference papers. Lin is the co-author of the book *Wireless and Mobile Network Architecture* (with Imrich Chlamtac; published by John Wiley & Sons). Lin is an IEEE fellow, an ACM fellow, an AAAS fellow, and an IEE fellow.

Ai-Chun Pang was born in Hsinchu, Taiwan, in 1973. She received the B.S., and M.S. degrees and her Ph.D. in Computer Science and Information Engineering from National Chiao Tung University in 1996, 1998, and 2002, respectively. She joined the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, as an assistant professor in 2002. From August 2004, Dr Pang also serves as an assistant professor in Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan. Her research interests include design and analysis of personal communications services network, mobile computing, voice over IP and performance modeling.

Herman Chung-Hwa Rao is vice president of Technology Development, Far Eastone Telecommunication Co., Ltd. He leads a team for Product Development, Service Network, and Platform planning and development, and Emerging Technology Assessment and Development. He also is an adjunct professor in Oriental Institute of Technology, Taiwan. Dr Rao received his B.S. in Mechanical Engineering from The National Taiwan University, and M.S. degree and his Ph.D. in Computer Science from The University of Arizona. Before joined Far Eastone, he has been a senior researcher in AT&T Bell Labs for 10 years. He has published more than 40 technical pagers, and received five USA Patents, and two Taiwan Patents.