Serving Radio Network Controller Relocation for UMTS All-IP Network

Ai-Chun Pang, *Member, IEEE*, Yi-Bing Lin, *Fellow, IEEE*, Hsien-Ming Tsai, *Member, IEEE*, and Prathima Agrawal, *Fellow, IEEE*

***Abstract*—To support real-time multimedia services in UMTS all-IP network, Third-Generation Partnership Project TR 25.936 proposed two approaches to support real-time serving radio network controller (SRNC) switching, which require packet duplication during SRNC relocation. These approaches significantly consume extra system resources. This paper proposes the fast SRNC relocation (FSR) approach that does not duplicate packets. In FSR, a packet buffering mechanism is implemented to avoid packet loss at the target RNC. We propose an analytic model to investigate the performance of FSR. The numerical results show that packet loss at the source RNC can be ignored. Furthermore, the expected number of packets buffered at the target RNC is small, which does not prolong packet delay.**

***Index Terms*—All-IP network, real-time multimedia services, serving radio network controller relocation, Universal Mobile Telecommunications System (UMTS).**

NOMENCLATURE

- D_1 Transmission delay between the GGSN and SGSN1.
- D_2 Transmission delay between the GGSN and SGSN2.
- D_3 Transmission delay between SGSN2 and the target RNC.
- δ_i Random variable to indicate if the previous ith packet is lost.
- λ_a Interpacket arrival rate.
- N_B Number of the buffered packets at the target RNC.
- N_L Number of the lost packets.
- θ Random variable of $x + y + z$.
- τ_0 Time when the Update_PDP_Context_Request message arrives at the GGSN.
- $\bar{\tau}_i$ Time when the previous ith packet is sent from the GGSN (tracking back from τ_0).
- τ_i Time when the ith packet is sent from the GGSN.
- $\bar{T}_i$ Time interval between τ_0 and $\bar{\tau}_i$.
- T_i Time interval between τ_0 and τ_i.
- w_i $= T_i + x_i$.
- x Transmission delay for the signaling messages Update_PDP_Context_Response and Relocation_Command.
- x_i $= D_1 + D_3$ or $D_2 + D_4$.
- y Transmission delay between SGSN1 and SGSN2.
- Y_i $= \bar{T}_i + x + y$.
- z Transmission delay between the source RNC and the target RNC.

Manuscript received January 28, 2003; revised September 25, 2003. This work was supported in part by the MOE Program for Promoting Academic Excellence of Universities under Grant 89-E-FA04-1-4, in part by the Chair Professorship of Providence University, IIS/Academia Sinica, in part by FarEastone, in part by CCL/ITRI, in part by the Lee and MTI Center for Networking Research/NCTU, and in part by the National Science Council under Contract NSC 92-2213-E-002-049.

A.-C. Pang is with the Department of Computer Science and Information Engineering, Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 106, Taiwan, R.O.C. (e-mail: acpang@csie.ntu.edu.tw; liny@csie.nctu.edu.tw).

Y.-B. Lin is with the Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu 30050, Taiwan, R.O.C. (e-mail: liny@csie.nctu.edu.tw).

H.-M. Tsai is with Quanta Research Institute, Quanta Computer, Inc., Taoyuan 333, Taiwan, R.O.C. (e-mail: samuel.tsai@quantatw.com).

P. Agrawal is with Internet Architecture Research Laboratory, Telcordia Technologies, Morristown, NJ 07960-6438 USA (e-mail: pagrawal@research.telcordia.com).

Digital Object Identifier 10.1109/JSAC.2004.825962

I. INTRODUCTION

MOBILITY, privacy, and immediacy offered by wireless access commonly create new opportunities for Internet business, and mobile networks are becoming a platform that provides leading-edge Internet services. Through integration of the Internet and the third-generation (3G) wireless communication, next-generation telecommunications networks will provide global information access for mobile users [11]. Third-Generation Partnership Project (3GPP) [1], [5], [6] proposed the Universal Mobile Telecommunications System (UMTS) all-IP architecture to integrate the IP and wireless technologies, which has evolved from the Global System for Mobile Communication (GSM), General Packet Radio Service (GPRS), and UMTS Release 1999.

Fig. 1 shows a UMTS all-IP network architecture (another UMTS all-IP option can be found in [1] and [5]). In this figure, the dashed lines represent signaling links and the solid lines represent data and signaling links. The UMTS all-IP network connects to the packet data network (PDN) [see Fig. 1(a)] or the IP multimedia core network subsystem [see Fig. 1(b)] through the serving GPRS support node (SGSN) [see Fig. 1(c)] and the gateway GPRS support node (GGSN) [see Fig. 1(d)]. The SGSN connects to the radio access network. The GGSN provides interworking with the external PDN, and is connected with SGSNs via an IP-based GPRS backbone network. Both the GGSN and SGSN communicate with the *home subscriber server* [see Fig. 1(e)] to obtain mobility and session management information of subscribers. The UMTS Terrestrial Radio Access Network (UTRAN) consists of Node Bs [the UMTS term for base stations; see Fig. 1(f)] and radio network controllers (RNCs) [see Fig. 1(g)] connected by an ATM network. A user equipment (UE) [see Fig. 1(h)] communicates with one or more Node Bs through the radio interface Uu based on the

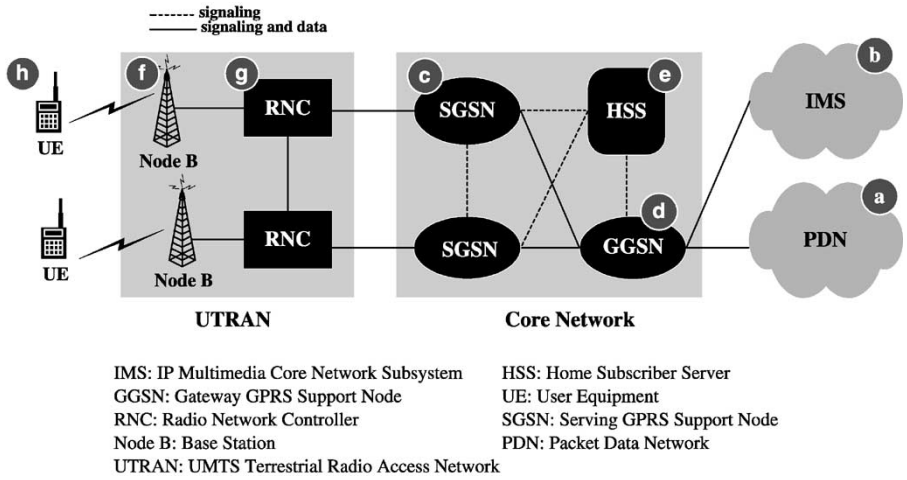

Fig. 1. UMTS all-IP network architecture.

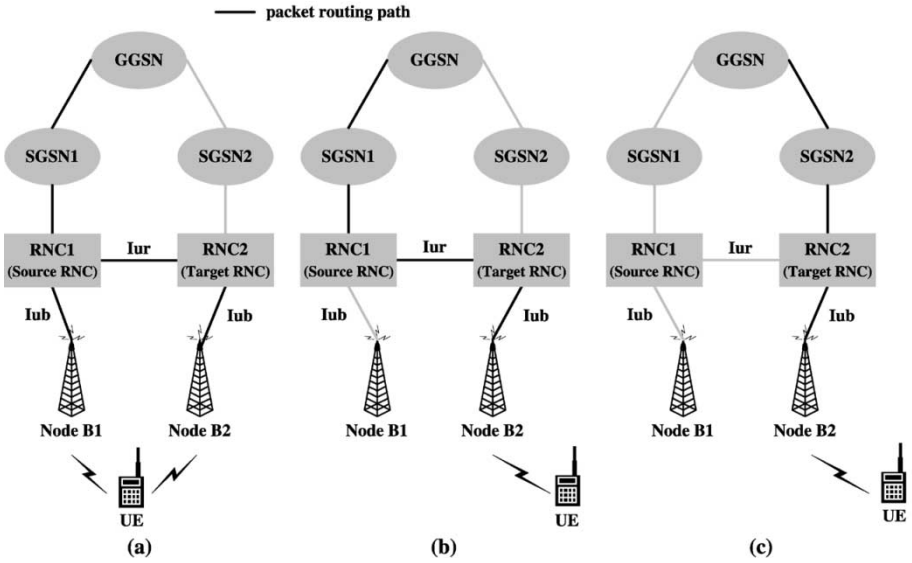

Fig. 2. SRNC relocation. (a) UE connects to both B1 and B2. (b) UE connects to B2 (before relocation). (c) UE connects to B2 (after relocation).

wideband code-division multiple-access (WCDMA) radio technology [8].

In the UMTS all-IP network, the IP packets are routed between the UE and the GGSN. By using the packet data protocol (PDP) context activation procedure [4], a PDP context is created to establish the routing path for IP packet delivery. Besides the packet routing information (e.g., the UEs IP address), the PDP context also contains the quality-of-service (QoS) profiles and other parameters. Due to the CDMA characteristics, multiple radio paths (for delivering the same IP packets) may exist between the UE and more than one Node Bs. An example of multiple routing paths is illustrated in Fig. 2(a). In this figure, 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., IP packets) sent from the UE to Node B2 can be forwarded to RNC1 through RNC2. RNC1 then combines the signals from Node B1 and B2, and forward 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 (between the Node B and the RNC) and Iur (between two RNCs) interfaces. 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 $\langle \mathrm{UE} \leftrightarrow \mathrm{Node\ B2} \leftrightarrow \mathrm{RNC2} \leftrightarrow \mathrm{RNC1} \leftrightarrow \mathrm{SGSN1} \leftrightarrow \mathrm{GGSN} \rangle$ as shown in Fig. 2(b). In this scenario, it does not make sense to route packets between the UE and the core network 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 Fig. 2(c)], and RNC2 becomes the SRNC.

In 3GPP TS 23.060 [4], a lossless SRNC relocation procedure was proposed for nonreal-time data services. In this approach, in the beginning of SRNC relocation, the source RNC [RNC1 in Fig. 2(b)] first stops transmitting downlink IP packets to the UE. Then, it forwards the next packets to the target RNC

[RNC2 in Fig. 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 UE. In this approach, no packet is lost during the SRNC switching period. Unfortunately, this approach does not support real-time data transmission because the IP 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 [3] proposes SRNC duplication (SD) and core network bicasting (CNB). These two approaches duplicate data packets during SRNC relocation, which may not efficiently utilize system resources. In this paper, we propose a new approach called fast SRNC relocation (FSR) to provide real-time SRNC switching without packet duplication. An analytic model is proposed to investigate the performance of FSR.

II. Related Work

This section describes the previously proposed SRNC relocation procedures for real-time multimedia services; that is, SD and CNB proposed in 3GPP TR 25.936 [3].

A. SRNC Duplication (SD)

Consider Fig. 2(b). Suppose that the UE is connected to the source RNC and SGSN1 before performing SRNC relocation. The target RNC is the drift RNC, which is connected to the source RNC via the Iur interface. After SRNC relocation, the SRNC role is moved from the source RNC to the target RNC, and the IP packets for the UE are directly routed through SGSN2 and the target RNC [see Fig. 2(c)]. Fig. 3 shows the four stages of the SD procedure. Stage I [Fig. 3(a)] initiates SRNC relocation. In this stage, the user IP packets are delivered through the old path $\langle \mathrm{GGSN} \leftrightarrow \mathrm{SGSN1} \leftrightarrow \text{source } \mathrm{RNC} \leftrightarrow \text{target } \mathrm{RNC} \leftrightarrow \mathrm{UE} \rangle$. The following steps are executed.

Steps 1 and 2: When the Node B of the source RNC no longer connects to the UE, the source RNC initiates SRNC relocation. Specifically, the source RNC sends a Relocation_Required message (including the ID of the target RNC) to SGSN1.

Step 3: Based on the ID of the target RNC, SGSN1 determines if the SRNC relocation is intra-SGSN SRNC relocation or inter-SGSN SRNC relocation. Assume that it is inter-SGSN SRNC relocation. By sending a Forward_Relocation_Request message, SGSN1 requests SGSN2 to allocate the resources (to be described in Step 4) for the UE.

Step 4: SGSN2 sends a Relocation_Request message with the radio access bearer (RAB) parameters to the target RNC. The RAB parameters include the traffic class (e.g., conversational, streaming, interactive, or background), traffic handling priority, maximum and guaranteed bit rates, and so on [2]. After all necessary resources for the RAB are successfully allocated, the target RNC sends a Relocation_Request_Acknowledge message to SGSN2.

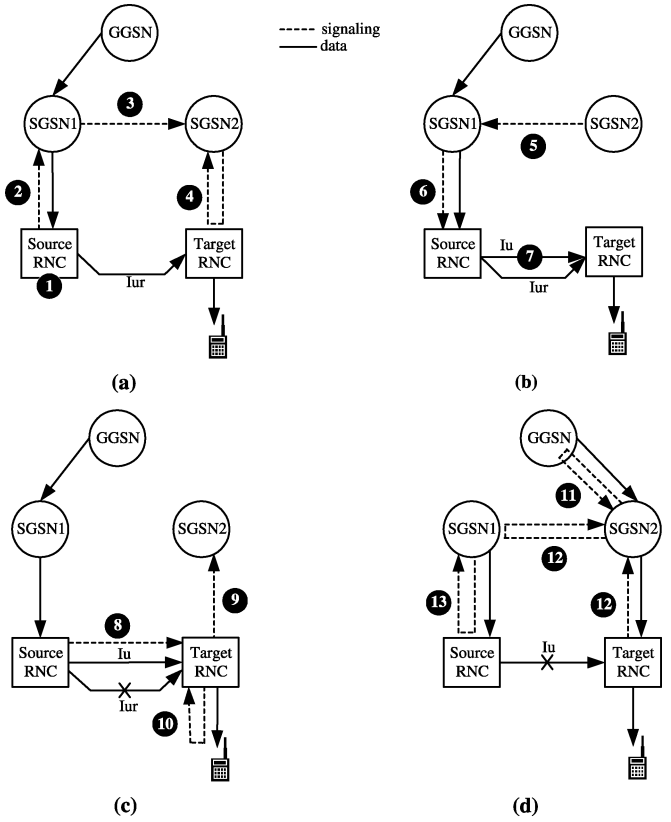

Fig. 3. SD approach. (a) Stage I. (b) Stage II. (c) Stage III. (d) Stage IV.

In Stage II [Fig. 3(b)], a forwarding path $\langle \text{source } \mathrm{RNC} \rightarrow \text{target } \mathrm{RNC} \rightarrow \mathrm{UE} \rangle$ for downlink packet delivery is created between the source and the target RNCs through the Iu interface. The source RNC duplicates the packets and forwards these packets to the target RNC. Thus, the downlink packets are simultaneously transmitted through both the old path (via the Iur interface) and the forwarding path (via the Iu interface) between the source RNC and the target RNC. Note that 3G TR25.936 [3] did not clearly describe if an Iu link can be directly established between two RNCs. If not, an indirect path $\langle \text{source } \mathrm{RNC} \rightarrow \mathrm{SGSN1} \rightarrow \mathrm{SGSN2} \rightarrow \text{target } \mathrm{RNC} \rangle$ is required. To favor the SD approach, we assume a direct link between the source and target RNCs. The following steps are executed in Stage II.

Steps 5 and 6: SGSN2 sends a Forward_Relocation_Response message to SGSN1, which indicates that all resources (e.g., RAB) are allocated. SGSN1 forwards this information to the source RNC through a Relocation_Command message.

Step 7: Upon receipt of the Relocation_Command message, the source RNC duplicates the downlink packets and transmits the duplicated packets to the target RNC through the forwarding path (via the Iu interface at the IP layer). The forwarded packets are discarded at the target RNC before it becomes the SRNC (i.e., before the target RNC receives the Relocation_Commit message at Step 8).

In Stage III [Fig. 3(c)], the Iur link between the source RNC and the target RNC (i.e., the old path) is disconnected. The downlink packets arriving at the source RNC are forwarded to the target RNC through the Iu link (i.e., the forwarding path). A

data-forwarding timer is maintained in the source RNC. When the timer expires, the forwarding operation at the source RNC is stopped. The following steps are executed in Stage III.

Step 8: With a Relocation_Commit message, the source RNC transfers serving radio network subsystem (SRNS) context (e.g., QoS profile for the RAB) to the target RNC.
Step 9: Upon receipt of the Relocation_Commit message, the target RNC sends a Relocation_Detect message to SGSN2, which indicates that the target RNC will become the SRNC.
Step 10: At the same time, the target RNC sends a RAN_Mobility_Information message to the UE. This message triggers the UE to send the uplink IP packets to the target RNC. After the UE has reconfigured itself, it replies the RAN_Mobility_Information_Confirm message to the target RNC.

In Stage IV [Fig. 3(d)], the packet routing path is switched from the old path to the new path $\langle \mathrm{GGSN} \leftrightarrow \mathrm{SGSN2} \leftrightarrow \mathrm{target\ RNC} \leftrightarrow \mathrm{UE} \rangle$. At this stage, the target RNC becomes the SRNC. The source RNC forwards the downlink packets to the target RNC until the data-forwarding timer expires. The following steps are executed in Stage IV.

Step 11: SGSN2 sends a Update_PDP_Context_Request message to the GGSN. Based on the received message, the GGSN updates the corresponding PDP context 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 moment, the target RNC receives the downlink packets from two paths (i.e., the forwarding and new paths), and transmits them to the UE. 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.
Step 12: By sending the Relocation_Complete message to SGSN2, the target RNC indicates the completion of the relocation procedure. Then, SGSN2 exchanges this information with SGSN1 using the Forward_Relocation_Complete and Forward_Relocation_Complete_Acknowledge message pair.
Step 13: Finally, SGSN1 sends an Iu_Release_Command message to request the source RNC to release the Iu connection in the forwarding path. When the data-forwarding timer expires, the source RNC replies an Iu_Release_Complete message.

B. Core Network Bicasting (CNB)

Fig. 4 shows the four stages of the CNB procedure when the communicating UE moves from the source RNC to the target RNC. Stage I [Steps 1–4, Fig. 4(a)] is the same as Stage I in SD, which requests the target RNC to allocate the necessary resources for relocation.

In Stage II [Fig. 4(b)], the downlink packets are duplicated at the GGSN, and are sent to the target RNC through both the old path $\langle \mathrm{GGSN} \rightarrow \mathrm{SGSN1} \rightarrow \mathrm{source\ RNC} \rightarrow \mathrm{target\ RNC} \rangle$ and the new path $\langle \mathrm{GGSN} \rightarrow \mathrm{SGSN2} \rightarrow \mathrm{target\ RNC} \rangle$. The following steps are executed.

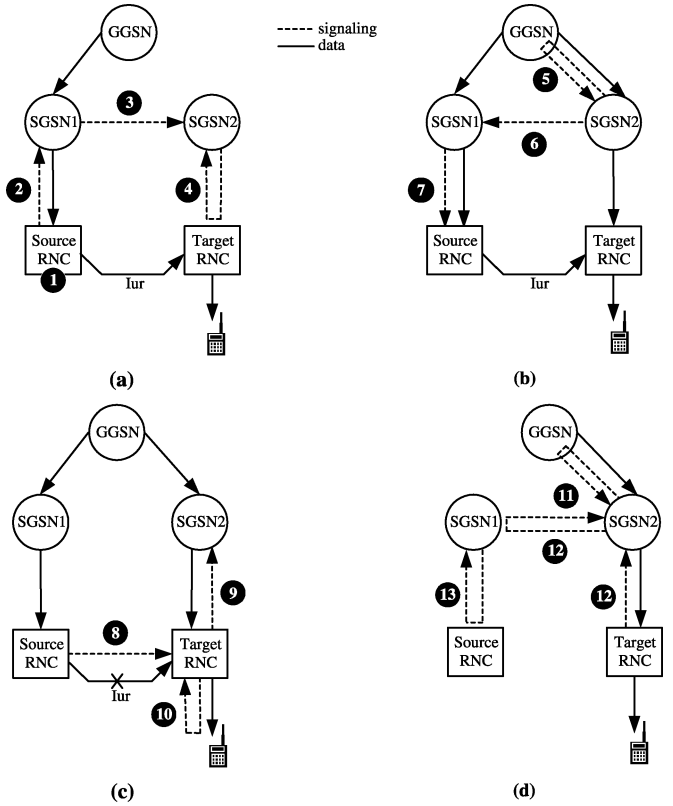

Fig. 4. CNB approach. (a) Stage I. (b) Stage II. (c) Stage III. (d) Stage IV.

Step 5: Upon receipt of the Relocation_Request_Acknowledge message at Step 4, SGSN2 sends a Update_PDP_Context_Request message that requests the GGSN to bicast the downlink packets. The GGSN starts to perform bicasting and replies SGSN2 a message Update_PDP_Context_Response. At this moment, the downlink packets are simultaneously transmitted to the target RNC through the old and the new paths. Since the target RNC has not taken the SRNC role (i.e., the target RNC has not received the Relocation_Commit message), the packets routed through the new path are discarded at the target RNC.
Steps 6 and 7: These steps are used to inform the source RNC that all necessary resources are allocated, which are similar to Steps 5 and 6 in the SD approach.

In Stage III [Fig. 4(c)], the Iur link between the source RNC and the target RNC is disconnected, and the downlink packets arriving at the source RNC are discarded.

Steps 8–10: These steps are used to move the SRNC role from the source RNC to the target RNC, which are similar to Steps 8–10 in the SD approach.

In Stage IV [Fig. 4(d)], the GGSN is informed to stop downlink packet bicasting. The target RNC takes the SRNC role to transmit the downlink packets to the UE.

Step 11: Through the Update_PDP_Context_Request message, SGSN2 informs the GGSN to stop downlink packet bicasting. Then, the GGSN removes the GTP tunnel between the GGSN and SGSN1, and replies SGSN2 the Update_PDP_Context_Response message.

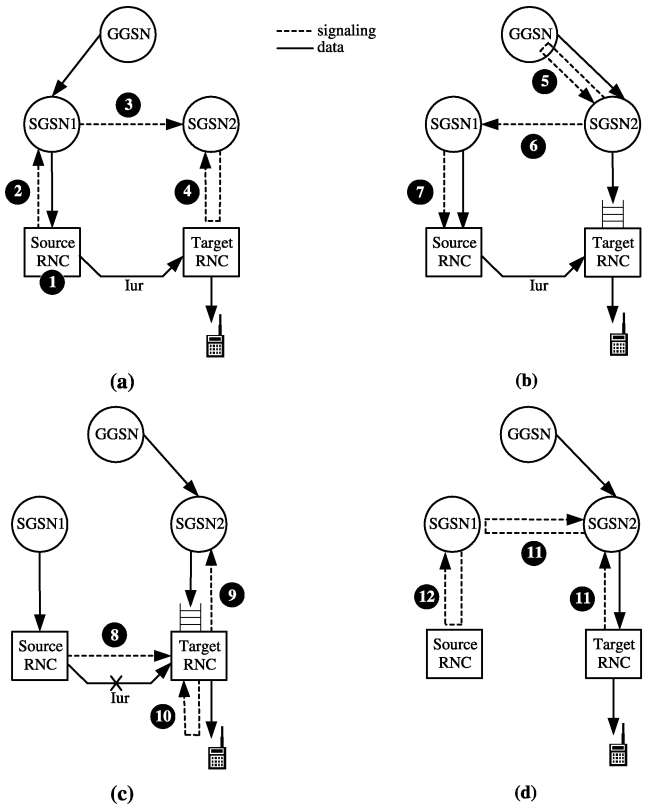

Fig. 5. FSR approach. (a) Stage I. (b) Stage II. (c) Stage III. (d) Stage IV.

Step 12: With the Relocation_Complete message, the target RNC informs SGSN2 that the relocation procedure is successfully performed. Then, SGSN2 exchanges this information with SGSN1 using the Forward_Relocation_Complete and Forward_Relocation_Complete_Acknowledge message pair.

Step 13: 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.

III. Fast SRNC Relocation (FSR)

This section describes the FSR approach and compares this approach with SD and CNB. As shown in Fig. 2(b), the UE is connected to the source RNC and SGSN1 before SRNC relocation. After relocation, the data packets for the UE are directly routed through the target RNC and SGSN2, as shown in Fig. 2(c). Fig. 5 illustrates the four stages of the FSR procedure.

Stage I [Fig. 5(a)] initiates SRNC relocation. In this stage, the routing path of downlink packets is $\langle \mathrm{GGSN} \rightarrow \mathrm{SGSN1} \rightarrow \mathrm{source\ RNC} \rightarrow \mathrm{target\ RNC} \rightarrow \mathrm{UE} \rangle$. 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. 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 [Fig. 5(b)], the GGSN routes the downlink packets to the old path before receiving the Update_PDP_Context_Request message [Step 5 in Fig. 5(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 $\langle \mathrm{GGSN} \rightarrow \mathrm{SGSN2} \rightarrow \mathrm{target\ RNC} \rangle$. 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 and 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 [Fig. 5(c)], the Iur link between the source RNC and the target RNC is disconnected. The "old" downlink packets arriving at the source RNC later than the Relocation_Command message [Step 7 in Fig. 5(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 $\langle \mathrm{UE} \rightarrow \mathrm{target\ RNC} \rightarrow \mathrm{SGSN2} \rightarrow \mathrm{GGSN} \rangle$.

By executing Steps 11 and 12 at Stage IV [Fig. 5(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 exchanges the Iu_Release_Command and Iu_Release_Complete message pair to release the Iu connection in the old path.

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

TABLE I
COMPARING FSR WITH SD AND CNB

Issues \ Approaches	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

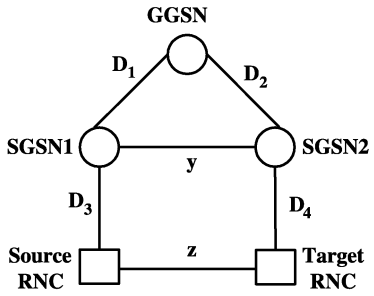

Fig. 6. Transmission delays.

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.

Packet Loss. Packet loss may occur in these three approaches either at the source RNC or at the target RNC. For SD and FSR, the 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 [Step 13 in Fig. 3(d)]. In FSR, the "old" packets are dropped if they arrive at the source RNC later than the Relocation_Command message [see Step 7 in Fig. 5(b)] does.

For SD and CNB, the 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 [Step 7 in Fig. 3(b)]. In CNB, the 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 [see Step 5 in Fig. 4(b)]. 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, the 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 [see Step 11 in Fig. 3(d)]. 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 [see Step 5 in Fig. 4(b)].

Extra Signaling. The SD approach follows the standard SRNC relocation procedure proposed in 3G 23.060 [4]. 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 [see Step 5 in Fig. 4] 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 to the existing 3GPP specifications.

In conclusion, SD and CNB require packet duplication that will double the network traffic load 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 $\langle \text{source RNC} \rightarrow \text{SGSN1} \rightarrow \text{SGSN2} \rightarrow \text{target RNC} \rangle$ 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. We will show that packet loss for FSR is not a serious problem in the subsequent sections.

IV. ANALYTIC MODELING

Since it is not clear how SD correctly functions, we will not conduct performance analysis for SD. Also, the performance of CNB is similar to that of FSR, which will be treated in a separate paper. In this paper, we only model the FSR approach. As described in the previous section, the routing path of the downlink packets for the UE is switched from the old path $\langle \text{GGSN} \rightarrow \text{SGSN1} \rightarrow \text{source RNC} \rightarrow \text{target RNC} \rangle$ to the new path $\langle \text{GGSN} \rightarrow \text{SGSN2} \rightarrow \text{target RNC} \rangle$ after the GGSN receives the Update_PDP_Context_Request message [Step 5 in Fig. 5(b)]. The packets delivered through the old path are lost if these packets arrive at the source RNC later than the Relocation_Command message does [Step 7 in Fig. 5(b)]. Therefore, an important performance measure is the expected number of lost packets $E[N_L]$ during SRNC relocation. Furthermore, the packets transmitted through the new path are buffered at the target RNC if they arrive at the target RNC earlier than the Relocation_Commit message does [Step 8 in Fig. 5(c)]. Hence, another important performance measure is the expected number of buffered packets $E[N_B]$ during SRNC relocation.

Fig. 6 denotes the transmission delays among the network nodes, which are represented by the random variables described as follows.

- D_1 The transmission delay between the GGSN and SGSN1.
- D_2 The transmission delay between the GGSN and SGSN2. Without loss of generality, we assume that D_1 and D_2 have the same distribution.
- D_3 The transmission delay between SGSN1 and the source RNC.

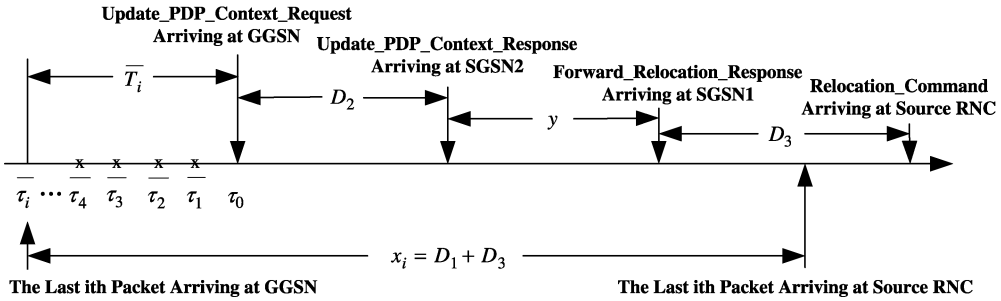

Fig. 7. Timing diagram for computing the number of lost packets.

D_4 The transmission delay between SGSN2 and the target RNC. Without loss of generality, we assume that D_3 and D_4 have the same distribution.

y The transmission delay between SGSN1 and SGSN2.

z The transmission delay between the source RNC and the target RNC.

Based on the above random variables, we develop an analytic model to derive the expected numbers of lost and buffered packets for FSR. The notations used in this paper are listed in the front.

A. Expected Number of Lost Packets

Consider the timing diagram in Fig. 7. Suppose that the GGSN receives the Update_PDP_Context_Request message from SGSN2 at time τ_0 [Step 5 in Fig. 5(b)]. Tracing back from τ_0, the previous ith packet was sent from the GGSN to SGSN1 at time $\bar{\tau}_i$. Following the assumption widely used in the literature, we assume that the interpacket arrivals are a Poisson stream, and the interpacket arrival times $\bar{\tau}_{i-1} - \bar{\tau}_i$ are exponentially distributed with the arrival rate λ_a. If the arrival of the Update_PDP_Context_Request message from SGSN2 to the GGSN at time τ_0 is a random observer, then from the residual life theorem [12] and memoryless property of the exponential distribution, $\tau_0 - \bar{\tau}_1$ has the exponential distribution with mean $(1/\lambda_a)$. Therefore, $\bar{T}_i = \tau_0 - \bar{\tau}_i$ has an Erlang distribution with the density function

$$f_{\bar{T}_i}(t) = \left[\frac{(\lambda_a t)^{i-1}}{(i-1)!}\right] \lambda_a e^{-\lambda_a t}. \tag{1}$$

For $i \geq 1$, the transmission delay for the previous ith packet through the path $\langle \text{GGSN} \rightarrow \text{SGSN1} \rightarrow \text{source RNC} \rangle$ can be represented by the random variable $x_i = D_1 + D_3$. The transmission delay for the signaling messages Update_PDP_Context_Response [Step 5 in Fig. 5(b)], Forward_Relocation_Response [Step 6 in Fig. 5(b)] and Relocation_Command [Step 7 in Fig. 5(b)] through the path $\langle \text{GGSN} \rightarrow \text{SGSN2} \rightarrow \text{SGSN1} \rightarrow \text{source RNC} \rangle$ can be represented by the random variable $x + y$, where $x = D_2 + D_3$ is identical to the random variable x_i. The intervals x and y have general distributions determined by the layout and transmission property of the UMTS all-IP core network. We assume that both x and y have mixed-Erlang density functions

$$f_x(t) = \sum_{j=1}^{J} \alpha_{x,j} \left[\frac{(\lambda_{x,j} t)^{m_{x,j}-1}}{(m_{x,j}-1)!}\right] \lambda_{x,j} e^{-\lambda_{x,j} t} \tag{2}$$

where $\sum_{j=1}^{J} \alpha_{x,j} = 1$, and

$$f_y(t) = \sum_{l=1}^{L} \alpha_{y,l} \left[\frac{(\lambda_{y,l} t)^{m_{y,l}-1}}{(m_{y,l}-1)!}\right] \lambda_{y,l} e^{-\lambda_{y,l} t} \tag{3}$$

where $\sum_{l=1}^{L} \alpha_{y,l} = 1$. In (2) and (3), J, L, $\alpha_{x,j}$ and $\alpha_{y,l}$ determine the shapes and scales of the distributions. The mixed-Erlang distribution is selected because this distribution has been proven as a good approximation to many other distributions as well as measured data [7], [9].

The previous ith packet is lost if it arrives at the source RNC later than the Relocation_Command message does. Let N_L be the number of the lost packets, and define

$$\delta_i = \begin{cases} 1, & \text{if the previous } i\text{th packet is lost} \\ & \text{(i.e., } \bar{T}_i + x + y < x_i\text{)} \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

Then, $N_L = \sum_{i=1}^{\infty} \delta_i$, and

$$\begin{aligned} E[N_L] &= E\left[\sum_{i=1}^{\infty} \delta_i\right] \\ &= \sum_{i=1}^{\infty} \Pr[\text{the previous } i\text{th packet is lost}] \\ &= \sum_{i=1}^{\infty} \Pr[\bar{T}_i + x + y < x_i] \end{aligned} \tag{5}$$

The Laplace transforms for x, y, and $\bar{T}_i$ respectively, are

$$f_x^*(s) = \sum_{j=1}^{J} \alpha_{x,j} \left(\frac{\lambda_{x,j}}{s + \lambda_{x,j}}\right)^{m_{x,j}} \tag{6}$$

$$f_y^*(s) = \sum_{l=1}^{L} \alpha_{y,l} \left(\frac{\lambda_{y,l}}{s + \lambda_{y,l}}\right)^{m_{y,l}} \tag{7}$$

$$f_{\bar{T}_i}^*(s) = \left(\frac{\lambda_a}{s + \lambda_a}\right)^{i} \tag{8}$$

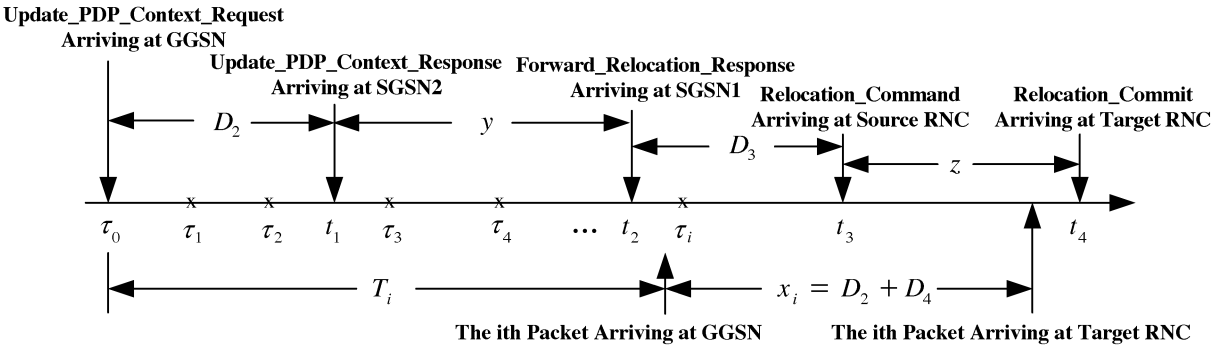

Fig. 8. Timing diagram for computing the number of buffered packets.

Let $Y_i = \bar{T}_i + x + y$ have the density function $f_{Y_i}(t)$ and Laplace transform $f^*_{Y_i}(s)$. Then, from (6)–(8) and the convolution property of distributions

$$\begin{aligned} f^*_{Y_i}(s) &= f^*_{\bar{T}_i} f^*_x(s) f^*_y(s) \\ &= \left(\frac{\lambda_a}{s+\lambda_a}\right)^i \left[\sum_{j=1}^{J} \alpha_{x,j} \left(\frac{\lambda_{x,j}}{s+\lambda_{x,j}}\right)^{m_{x,j}}\right] \\ &\quad \times \left[\sum_{l=1}^{L} \alpha_{y,l} \left(\frac{\lambda_{y,l}}{s+\lambda_{y,l}}\right)^{m_{y,l}}\right]. \end{aligned} \tag{9}$$

In (5), $\Pr[\bar{T}_i + x + y < x_i]$ is rewritten as

$$\begin{aligned} &\Pr[Y_i < x_i] \\ &= \int_{Y_i=0}^{\infty} \int_{x_i=Y_i}^{\infty} f_{Y_i}(Y_i) f_x(x_i) dx_i dY_i \\ &= \int_{Y_i=0}^{\infty} f_{Y_i}(Y_i) \int_{x_i=Y_i}^{\infty} \sum_{j=1}^{J} \alpha_{x,j} \\ &\quad \times \left[\frac{(\lambda_{x,j} x_i)^{m_{x,j}-1}}{(m_{x,j}-1)!}\right] \lambda_{x,j} e^{-\lambda_{x,j} x_i} dx_i dY_i \\ &= \int_{Y_i=0}^{\infty} f_{Y_i}(Y_i) \left\{ \sum_{j=1}^{J} \alpha_{x,j} \times \sum_{n_j=0}^{m_{x,j}-1} \left[\frac{(\lambda_{x,j} Y_i)^{n_j}}{n_j!}\right] \right. \\ &\qquad \left. e^{-\lambda_{x,j} Y_i} \right\} dY_i \\ &= \sum_{j=1}^{J} \alpha_{x,j} \left\{ \sum_{n_j=0}^{m_{x,j}-1} \left(\frac{\lambda_{x,j}^{n_j}}{n_j!}\right) \right. \\ &\qquad \left. \times \left[\int_{Y_i=0}^{\infty} Y_i^{n_j} f_{Y_i}(Y_i) e^{-\lambda_{x,j} Y_i}\right] dY_i \right\} \\ &= \sum_{j=1}^{J} \alpha_{x,j} \left\{ \sum_{n_j=0}^{m_{x,j}-1} \left(\frac{\lambda_{x,j}^{n_j}}{n_j!}\right) \right. \\ &\qquad \left. \times \left[\frac{(-1)^{n_j} d^{n_j} f^*_{Y_i}(s)}{ds^{n_j}}\bigg|_{s=\lambda_{x,j}}\right] \right\}. \end{aligned} \tag{10}$$

Therefore, the expected number $E[N_L]$ of lost packets can be computed by using (5), (9), and (10).

B. Expected Number of Buffered Packets

Consider the timing diagram in Fig. 8. Suppose that at time τ_0, the Update_PDP_Context_Response message is sent from the GGSN to SGSN2 [Step 5 in Fig. 5(b)]. SGSN2 receives this message at time t_1 and issues the Forward_Relocation_Response message to SGSN1 [Step 6 in Fig. 5(b)]. SGSN1 receives the message at t_2 and sends the Relocation_Command message to the source RNC [Step 7 in Fig. 5(b)]. The source RNC receives the message at time t_3, and transfers SRNS contexts to the target RNC by using the Relocation_Commit message [Step 8 in Fig. 5(c)]. The message arrives at the target RNC at time t_4. The transmission delay $t_4 - \tau_0$ can be represented by the random variable $D_2 + y + D_3 + z = x + y + z$. During this period, several packets may have been sent from the GGSN to the target RNC through SGSN2. We assume that z has a mixed-Erlang distribution with density function $f_z(t)$ and Laplace transform $f^*_z(s)$, where for $\sum_{p=1}^{P} \alpha_{z,p} = 1$, we have

$$\begin{aligned} f_z(t) &= \sum_{p=1}^{P} \alpha_{z,p} \left[\frac{(\lambda_{z,p} t)^{m_{z,p}-1}}{(m_{z,p}-1)!}\right] \lambda_{z,p} e^{-\lambda_{z,p} t} \\ f^*_z(s) &= \sum_{p=1}^{P} \alpha_{z,p} \left(\frac{\lambda_{z,p}}{s+\lambda_{z,p}}\right)^{m_{z,p}}. \end{aligned}$$

The ith packet was sent from the GGSN at time $\tau_i = \tau_0 + T_i$ through the new path $\langle \text{GGSN} \rightarrow \text{SGSN2} \rightarrow \text{target RNC} \rangle$, and its transmission delay can be represented by the random variable $x_i = D_2 + D_4$. Note that T_i has the same distribution as $\bar{T}_i$.

Suppose that N_B packets arrive at the target RNC during the period $[\tau_0, t_4]$ (i.e., during the transition of routing path switching). Then, these packets must be buffered in the target RNC. From the above discussion, the expected number $E[N_B]$ of buffered packets at the target RNC is

$$\begin{aligned} E[N_B] &= \sum_{i=1}^{\infty} \Pr[\text{the } i\text{th packet is queued in the buffer}] \\ &= \sum_{i=1}^{\infty} \Pr[T_i + x_i < x + y + z]. \end{aligned} \tag{11}$$

Let $w_i = T_i + x_i$ with density function $f_{w_i}(w)$. Then

$$\begin{aligned} f_{w_i}(w) &= \int_{t=0}^{w} f_{T_i}(w-t) f_x(t) dt \\ &= \int_{t=0}^{w} \left\{ \frac{[\lambda_a (w-t)]^{i-1}}{(i-1)!} \right\} \lambda_a e^{-\lambda_a (w-t)} \\ &\quad \times \left\{ \sum_{j=1}^{J} \alpha_{x,j} \left[\frac{(\lambda_{x,j} t)^{m_{x,j}-1}}{(m_{x,j}-1)!} \right] \lambda_{x,j} e^{-\lambda_{x,j} t} \right\} dt \\ &= \left[\frac{\lambda_a^i e^{-\lambda_a w}}{(i-1)!} \right] \\ &\quad \times \left\{ \sum_{j=1}^{J} \left[\frac{\alpha_{x,j} \lambda_{x,j}^{m_{x,j}}}{(m_{x,j}-1)!} \right] \right. \\ &\quad \times \int_{t=0}^{w} \left[\sum_{k=0}^{i-1} \binom{i-1}{k} w^{i-k-1} (-t)^k \right] \\ &\quad \left. \times e^{\Lambda_j t} t^{m_{x,j}-1} dt \right\} \\ &= \left[\frac{\lambda_a^i e^{-\lambda_a w}}{(i-1)!} \right] \\ &\quad \times \left\{ \sum_{j=1}^{J} \left[\frac{\alpha_{x,j} \lambda_{x,j}^{m_{x,j}}}{(m_{x,j}-1)!} \right] \right. \\ &\quad \times \sum_{k=0}^{i-1} \left[\binom{i-1}{k} \times (-1)^k w^{i-k-1} \right. \\ &\quad \left. \left. \times \int_{t=0}^{w} e^{\Lambda_j t} t^{k+m_{x,j}-1} dt \right] \right\} \end{aligned} \tag{12}$$

where $\Lambda_j = \lambda_a - \lambda_{x,j}$. Let $K_j(w)$ be the term $\int_{t=0}^{w} e^{(\lambda_a - \lambda_{x,j})t} t^{k+m_{x,j}-1} dt$ in (12), which can be expressed as shown in (13a) and (13b) at the bottom of the page. Then, (12) is rewritten as

$$\begin{aligned} f_{w_i}(w) = \left[\frac{\lambda_a^i e^{-\lambda_a w}}{(i-1)!} \right] &\left\{ \sum_{j=1}^{J} \left[\frac{\alpha_{x,j} \lambda_{x,j}^{m_{x,j}}}{(m_{x,j}-1)!} \right] \right. \\ &\left. \times \sum_{k=0}^{i-1} \left[\binom{i-1}{k} (-1)^k w^{i-k-1} K_j(w) \right] \right\}. \end{aligned} \tag{14}$$

Consider the case where $\lambda_a \neq \lambda_{x,j}$. From (13a), (14) is rewritten as

$$\begin{aligned} f_{w_i}(w) = &\sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{r=1}^{k+m_{x,j}} A(j,k,r) w^{i+m_{x,j}-r-1} e^{-\lambda_{x,j} w} \\ &+ \sum_{j=1}^{J} \sum_{k=0}^{i-1} B(j,k) w^{i-k-1} e^{-\lambda_a w} \end{aligned} \tag{15}$$

where

$$\begin{aligned} A(j,k,r) = (-1)^{k+1} &\binom{i-1}{k} \left[\frac{\alpha_{x,j} \lambda_a^i \lambda_{x,j}^{m_{x,j}}}{(\lambda_{x,j} - \lambda_a)^r (i-1)!} \right] \\ &\times \left[\frac{(k+m_{x,j}-1)!}{(m_{x,j}-1)!(k+m_{x,j}-r)!} \right] \end{aligned} \tag{16}$$

and

$$\begin{aligned} B(j,k) = (-1)^k &\binom{i-1}{k} \left[\frac{\alpha_{x,j} \lambda_a^i \lambda_{x,j}^{m_{x,j}}}{(\lambda_{x,j} - \lambda_a)^{k+m_{x,j}} (i-1)!} \right] \\ &\times \left[\frac{(k+m_{x,j}-1)!}{(m_{x,j}-1)!} \right]. \end{aligned} \tag{17}$$

From (15), we compute the complementary distribution function of f_{w_i} as

$$\begin{aligned} &\int_{w=\theta}^{\infty} f_{w_i}(w) dw \\ &\quad = \sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{r=1}^{k+m_{x,j}} A(j,k,r) \\ &\quad \quad \times \int_{w=\theta}^{\infty} w^{i+m_{x,j}-r-1} e^{-\lambda_{x,j} w} dw \\ &\quad \quad + \sum_{j=1}^{J} \sum_{k=0}^{i-1} B(j,k) \int_{w=\theta}^{\infty} w^{i-k-1} e^{-\lambda_a w} dw \\ &\quad = \sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{r=1}^{k+m_{x,j}} A(j,k,r) \\ &\quad \quad \times \left\{ e^{-\lambda_{x,j} w} \sum_{u=1}^{i+m_{x,j}-r} \left[\frac{-w^{i+m_{x,j}-r-u}}{\lambda_{x,j}^u} \right] \right. \\ &\quad \quad \left. \times \left[\frac{(i+m_{x,j}-r-1)!}{(i+m_{x,j}-r-u)!} \right] \Big|_{w=\theta}^{\infty} \right\} \\ &\quad \quad + \sum_{j=1}^{J} \sum_{k=0}^{i-1} B(j,k) \\ &\quad \quad \times \left[e^{-\lambda_a w} \sum_{v=1}^{i-k} \frac{-w^{i-k-v} (i-k-1)!}{\lambda_a^v (i-k-v)!} \Big|_{w=\theta}^{\infty} \right] \\ &\quad = \sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{r=1}^{k+m_{x,j}} \sum_{u=1}^{i+m_{x,j}-r} C(j,k,r,u) \\ &\quad \quad \times \theta^{i+m_{x,j}-r-u} e^{-\lambda_{x,j} \theta} \\ &\quad \quad + \sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{v=1}^{i-k} D(j,k,v) \theta^{i-k-v} e^{-\lambda_a \theta} \end{aligned} \tag{18}$$

$$K_j(w) = \begin{cases} e^{(\lambda_a - \lambda_{x,j}) w} \left\{ \sum_{r=1}^{k+m_{x,j}} w^{k+m_{x,j}-r} \times \left[\frac{-(k+m_{x,j}-1)!}{(\lambda_{x,j}-\lambda_a)^r (k+m_{x,j}-r)!} \right] \right\} + \left[\frac{(k+m_{x,j}-1)!}{(\lambda_{x,j}-\lambda_a)^{k+m_{x,j}}} \right], & \text{for } \lambda_a \neq \lambda_{x,j} \quad (13a) \\ \frac{w^{m_{x,j}+k}}{m_{x,j}+k}, & \text{for } \lambda_a = \lambda_{x,j} \quad (13b) \end{cases}$$

where

$$C(j,k,r,u) = A(j,k,r)\left[\frac{(i+m_{x,j}-r-1)!}{\lambda_{x,j}^{u}(i+m_{x,j}-r-u)!}\right] \quad (19)$$

and

$$D(j,k,v) = B(j,k)\left[\frac{(i-k-1)!}{\lambda_{a}^{v}(i-k-v)!}\right]. \quad (20)$$

Let $\theta = x+y+z$, which has the density function $f_\theta(\theta)$ and Laplace transform $f_\theta^*(s)$. From the convolution rule, it is clear that

$$f_\theta^*(s) = f_x^*(s) f_y^*(s) f_z^*(s) \quad (21)$$

We derive the probability that $w_i > \theta$ as follows:

$$\Pr[w_i > \theta] = \int_{\theta=0}^{\infty} f_\theta(\theta) \int_{w=\theta}^{\infty} f_{w_i}(w)\,dw\,d\theta. \quad (22)$$

From (18), (22) is rewritten as

$$\begin{aligned}
&\Pr[w_i > \theta] \\
&= \int_{\theta=0}^{\infty} f_\theta(\theta) \\
&\quad \times \left\{ \left[\sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{r=1}^{k+m_{x,j}} \right. \right. \\
&\qquad \left. \times \sum_{u=1}^{i+m_{x,j}-r} C(j,k,r,u)\theta^{i+m_{x,j}-r-u} e^{-\lambda_{x,j}\theta} \right] \\
&\qquad \left. + \left[\sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{v=1}^{i-k} D(j,k,v)\theta^{i-k-v} e^{-\lambda_a \theta} \right] \right\} d\theta \\
&= \left[\sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{r=1}^{k+m_{x,j}} \sum_{u=1}^{i+m_{x,j}-r} C(j,k,r,u) \right. \\
&\qquad \left. \times \int_{\theta=0}^{\infty} f_\theta(\theta) e^{-\lambda_{x,j}\theta} \theta^{i+m_{x,j}-r-u} d\theta \right] \\
&\quad + \left[\sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{v=1}^{i-k} D(j,k,v) \right. \\
&\qquad \left. \times \int_{\theta=0}^{\infty} f_\theta(\theta) e^{-\lambda_a \theta} \theta^{i-k-v} d\theta \right] \\
&= \left\{ \sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{r=1}^{k+m_{x,j}} \sum_{u=1}^{i+m_{x,j}-r} C(j,k,r,u) \right. \\
&\qquad \left. \times \left[\frac{(-1)^{q_c(i,j,r,u)} d^{q_c(i,j,r,u)} f_\theta^*(s)}{ds^{q_c(i,j,r,u)}} \Big|_{s=\lambda_{x,j}} \right] \right\} \\
&\quad + \left\{ \sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{v=1}^{i-k} D(j,k,v) \right. \\
&\qquad \left. \times \left[\frac{(-1)^{q_d(i,k,v)} d^{q_d(i,k,v)} f_\theta^*(s)}{ds^{q_d(i,k,v)}} \Big|_{s=\lambda_a} \right] \right\} \quad (23)
\end{aligned}$$

where $q_c(i,j,r,u) = i+m_{x,j}-r-u$ and $q_d(i,k,v) = i-k-v$.

Therefore, the expected number $E[N_B]$ of buffered packets can be computed by using (11), (23), and (21).

Similarly, for the case where $\lambda_a = \lambda_{x,j}$, we substitute (13b) in (14) to yield

$$\begin{aligned}
\Pr[w_i > \theta] &= \int_{\theta=0}^{\infty} f_\theta(\theta) \\
&\quad \times \left[\sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{r=0}^{i+m_{x,j}-1} C'(j,k,r)\theta^r e^{-\lambda_{x,j}\theta} \right] d\theta \\
&= \sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{r=0}^{i+m_{x,j}-1} C'(j,k,r) \\
&\quad \times \int_{\theta=0}^{\infty} f_\theta(\theta) e^{-\lambda_{x,j}\theta} \theta^r d\theta \\
&= \sum_{j=1}^{J} \sum_{k=0}^{i-1} \sum_{r=0}^{i+m_{x,j}-1} C'(j,k,r) \\
&\quad \times \left[\frac{(-1)^r d^r f_\theta^*(s)}{ds^r} \Big|_{s=\lambda_{x,j}} \right]
\end{aligned}$$

where

$$C'(j,k,r) = (-1)^k \binom{i-1}{k} \left[\frac{\alpha_{x,j} \lambda_a^i \lambda_{x,j}^{r-i} (i+m_{x,j}-1)!}{(i-1)!(m_{x,j}-1)!r!(m_{x,j}+k)} \right].$$

C. Validation of Simulation

By using C program, we have developed a discrete simulation model to validate against our analytic analysis. The simulation model follows the approach we developed in [10], and the details are omitted. Table II shows several numerical examples where $\bar{T}_2$ and T_2 have Erlang distribution with the density function given in (1), and their means (i.e., $E[\bar{T}_2]$ and $E[T_2]$) are $0.5/\lambda_{x,1}$ and $2/\lambda_{x,1}$, respectively. We consider the x (x_2), y and z intervals with Exponential, Erlang-2, and mixed-Erlang distributions, where $\lambda_{y,1} = \lambda_{z,1} = 2\lambda_{x,1}$, $\lambda_{y,2} = \lambda_{z,2} = 2\lambda_{x,2}$ and $\lambda_{x,2} = 2\lambda_{x,1}$. The output measures are $Pr[Y_2 < x_2]$ (i.e., the probability that the previous second packet arrives at source RNC later than the Relocation_Command message does) and $Pr[w_2 > \theta]$ (i.e., the probability that the second packet arrives at the target RNC later than Relocation_Commit message does). From Table II, $Pr[Y_2 < x_2]$ decreases and $Pr[w_2 > \theta]$ increases as $E[T_2]$ ($E[\bar{T}_2]$) increases. Table II shows that the analytic analysis and simulation experiments are consistent. Specifically, for all cases considered, the errors are within 0.5%. For other input parameter values, similar results are observed, which will not be presented in this paper.

V. Performance Evaluation

Based on the analysis in the previous section, we use numerical examples to investigate the performance of $E[N_L]$ (i.e., the expected number of lost packets) and $E[N_B]$ (i.e., the expected number of buffered packets) for the FSR approach. In our experiments, the mixed-Erlang distributions for x, y, and z have the parameters $\alpha_{x,i} = \alpha_{y,i} = \alpha_{z,i} = 0.5$ and $m_{x,i} = m_{y,i} =$

TABLE II
COMPARISON OF THE ANALYTIC AND SIMULATION RESULTS ($\lambda_{y,1} = \lambda_{z,1} = 2\lambda_{x,1}$, $\lambda_{y,2} = \lambda_{z,2} = 2\lambda_{x,2}$ AND $\lambda_{x,2} = 2\lambda_{x,1}$)

x, x_2, y and z are Exponentially distributed. ($\alpha_{x,1} = \alpha_{y,1} = \alpha_{z,1} = 1$) ($m_{x,1} = m_{y,1} = m_{z,1} = 1$)	$Pr[Y_2 < x_2]$ ($E[\bar{T}_2] = 0.5/\lambda_{x,1}$)	$Pr[Y_2 < x_2]$ ($E[\bar{T}_2] = 2/\lambda_{x,1}$)	$Pr[w_2 > \theta]$ ($E[T_2] = 0.5/\lambda_{x,1}$)	$Pr[w_2 > \theta]$ ($E[T_2] = 2.0/\lambda_{x,1}$)
Simulation	21.39%	8.35%	36.21%	68.53%
Analysis	21.33%	8.33%	36.20%	68.52%
Error	0.26%	0.15%	0.03%	0.02%
x, x_2, y and z are Erlang-2 distributed. ($\alpha_{x,1} = \alpha_{y,1} = \alpha_{z,1} = 1$) ($m_{x,1} = m_{y,1} = m_{z,1} = 2$)	$Pr[Y_2 < x_2]$ ($E[\bar{T}_2] = 0.5/\lambda_{x,1}$)	$Pr[Y_2 < x_2]$ ($E[\bar{T}_2] = 2/\lambda_{x,1}$)	$Pr[w_2 > \theta]$ ($E[T_2] = 0.5/\lambda_{x,1}$)	$Pr[w_2 > \theta]$ ($E[T_2] = 2.0/\lambda_{x,1}$)
Simulation	21.77%	10.18%	23.45%	48.66%
Analysis	21.81%	10.19%	23.45%	48.79%
Error	0.16%	0.02%	0.00%	0.39%
x, x_2, y and z are mixed-Erlang distributed. ($\alpha_{x,i} = \alpha_{y,i} = \alpha_{z,i} = 0.5$) ($m_{x,i} = m_{y,i} = m_{z,i} = 2$) ($i = 1$ and 2)	$Pr[Y_2 < x_2]$ ($E[\bar{T}_2] = 0.5/\lambda_{x,1}$)	$Pr[Y_2 < x_2]$ ($E[\bar{T}_2] = 2/\lambda_{x,1}$)	$Pr[w_2 > \theta]$ ($E[T_2] = 0.5/\lambda_{x,1}$)	$Pr[w_2 > \theta]$ ($E[T_2] = 2.0/\lambda_{x,1}$)
Simulation	21.20%	9.21%	27.84%	57.60%
Analysis	21.20%	9.22%	27.85%	57.63%
Error	0.00%	0.15%	0.02%	0.40%

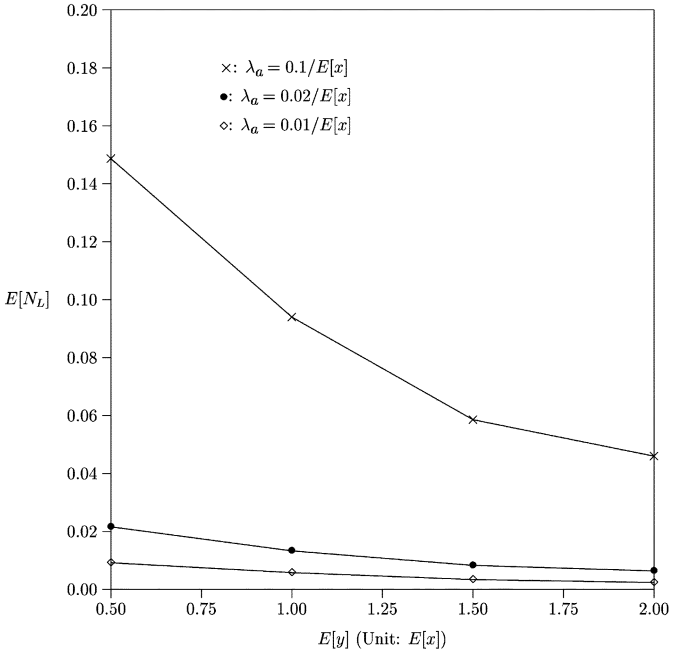

Fig. 9. $E[N_L]$ performance (x and y are mixed-Erlang distributed).

$m_{z,i} = 2$ for $i = 1$ and 2. Similar results are observed for other parameter values, which will not be presented here.

By using (5), (9), and (10), Fig. 9 plots $E[N_L]$ as a function of $E[y]$ (i.e., the expected number of y) ranging from $0.5E[x]$ to $2E[x]$. The $E[y]$ value is selected depending on whether the SGSNs are located in the same network or different networks. If the two SGSNs are in the same network, the transmission delay $E[y]$ is the same as that between the GGSN and the SGSN. Thus, $E[y] \simeq 0.5E[x]$. If these SGSNs are in the different networks, $E[y] \simeq 2E[x]$ may be appropriate. Depending on the applications being investigated, we consider $\lambda_a = 0.1/E[x]$, $0.02/E[x]$, and $0.01/E[x]$. Note that the 100 Mb/s Fast Ethernet and 155.52 Mbps STM-1/ATM have been commonly adopted for Gi (between the GGSN and the SGSN) and Iu (between the SGSN and the RNC). For real-time applications such as VoIP and video streaming services, the packet size typically ranges from 200 to 1500 bytes, and the interpacket arrival time $(1/\lambda_a)$ ranges from 10 to 40 ms. Therefore, our study selects the λ_a values in the range $[(0.01/E[x]), (0.1/E[x])]$. Fig. 9 shows intuitive results that $E[N_L]$ is an increasing function of λ_a, and is a decreasing function of $E[y]$. This figure indicates that the $E[N_L]$ performance is reasonably good. For example, when $E[y] = E[x]$, the expected number of lost packets $E[N_L]$ for VoIP application (i.e., $\lambda_a = 0.01/E[x]$) is 0.006. For video streaming services (i.e., $\lambda_a = 0.1/E[x]$), $E[N_L] = 0.09$. Also, when $E[y]$ increases from $E[x]$ to $2E[x]$, $E[N_L]$ is significantly reduced (i.e., 51%, 52% and 58% reductions for $\lambda_a = 0.1/E[x]$, $0.02/E[x]$, $0.01/E[x]$, respectively). In other words, the FSR performance can be improved by increasing the speed of the "x" link over the "y" link.

By using (11), (21), and (23), Fig. 10 shows the effects of $E[y]$ and λ_a on $E[N_B]$, where $\lambda_a = 0.1/E[x]$, $0.02/E[x]$ and $0.01/E[x]$. In this figure, we consider $E[y] = E[z]$ that ranges from $0.5E[x]$ to $2E[x]$. This figure intuitively indicates that $E[N_B]$ increases as $E[y]$ and λ_a increase. Since the expected number of buffered packets at the target RNC is below 3.5 for all cases considered in our study, it is clear that the packet buffering mechanism does not result in long packet delay (due to queuing).

VI. CONCLUSION

In 3GPP TR 25.936, SD and CNB were proposed to support real-time multimedia services in the UMTS all-IP network. Both approaches require packet duplication during SRNC

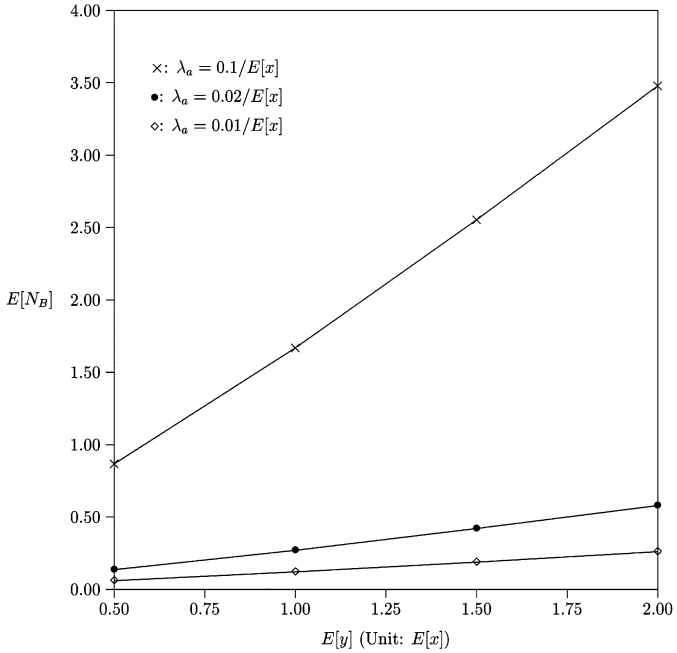

Fig. 10. $E[N_B]$ performance (x, y, and z are mixed-Erlang distributed).

relocation, which significantly consume the system resources. This paper proposed a FSR approach that provides real-time SRNC switching without packet duplication. In FSR, the packet buffering mechanism is implemented to avoid packet loss at the target RNC. We developed an analytic model to investigate the performance of FSR, which was validated against the simulation experiments. We note that packet loss cannot be avoided during SRNC relocation if we want to support real-time multimedia traffic in the UMTS all-IP network. 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. FSR can be implemented in the GGSN, SGSN, and RNC without introducing new message types to the existing 3GPP specifications. As a final remark, the FSR approach is a U.S. and an R.O.C. pending patents.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers. Their comments have significantly improved the quality of this paper.

REFERENCES

[1] Third-Generation Partnership Project (3GPP), "Architectural for an All IP Network," Tech. Spec. Group Serv. Syst. Aspects, Tech. Rep. 3G TR 23.922, version 1.0.0 (1999-10), 1999.

[2] Third-Generation Partnership Project (3GPP), "UTRAN Iu Interface RANAP Signaling for Release 1999," Tech. Spec. Group Radio Access Network, Tech. Spec. 3G TS 25.413, version 3.4.0 (2000-12), 2000.

[3] Third-Generation Partnership Project (3GPP), "Handovers for Real-Time Services From PS Domain," Technical Specification Group RAN 3, Tech. Rep. 3G TR 25.936, version 4.0.1 (2000-12), 2001.

[4] Third-Generation Partnership Project (3GPP), "General Packet Radio Service (GPRS); Service Description; Stage 2," Tech. Spec. Group Serv. Syst. Aspects, Tech. Spec. 3G TS 23.060, version 4.1.0 (2001-06), 2001.

[5] Third-Generation Partnership Project (3GPP), "Network Architecture," Tech. Spec. Group Serv. Syst. Aspects, Tech. Spec. T523.060 ,version 4.1.0 (2001-06), 2001.

[6] L. Bos and S. Leroy, "Toward an all-IP-based UMTS system architecture," *IEEE Network*, vol. 15, no. 1, pp. 36–45, 2001.

[7] Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling for PCS networks," *IEEE Trans. Commun.*, vol. 47, pp. 1062–1072, July 1999.

[8] H. Holma and A. Toskala, Eds., *WCDMA for UMTS*. New York: Wiley, 2000.

[9] F. P. Kelly, *Reversibility and Stochastic Networks*. New York: Wiley, 1979.

[10] Y.-B. Lin, H.-Y. Cheng, Y.-H. Cheng, and P. Agrawal, "Implementing automatic location update for follow-me database using VoIP and bluetooth technologies," *IEEE Trans. Comput.*, vol. 51, pp. 1154–1168, Oct. 2002.

[11] Y.-B. Lin, Y.-R. Huang, A.-C. Pang, and I. Chlamtac, "All-IP approach for third generation mobile networks," *IEEE Network*, vol. 16, no. 5, pp. 8–19, 2002.

[12] S. M. Ross, *Stochastic Processes*. New York: Wiley, 1996.

Ai-Chun Pang (S'00–M'02) was born in Hsinchu, Taiwan, R.O.C., in 1973. She received the B.S., M.S. and Ph.D. degrees in computer science and information engineering from National Chiao-Tung University (NCTU), Hsinchu, Taiwan, in 1996, 1998, and 2002, respectively.

She joined the Department of Computer Science and Information Engineering, National Taiwan University (NTU), Taipei, as an Assistant Professor in 2002. In August 2004, she will be an Assistant Professor of the Graduate Institute of Networking and Multimedia, NTU. Her research interests include design and analysis of personal communications services network, mobile computing, voice over IP, and performance modeling.

Yi-Bing Lin (M'96–SM'96–F'03) received the B.S.E.E. degree from the National Cheng Kung University, Tainan, Taiwan, in 1983 and the Ph.D. degree in computer science from the University of Washington, Seattle, in 1990.

From 1990 to 1995, he was with the Applied Research Area at Bell Communications Research (Bellcore), Morristown, NJ. In 1995, he was appointed as Professor of the Department of Computer Science and Information Engineering (CSIE), National Chiao-Tung University (NCTU), Hsinchu, Taiwan. In 1996, he was appointed as Deputy Director of Microelectronics and Information Systems Research Center, NCTU. From 1997 to 1999, he was elected as Chairman of CSIE, NCTU. He is an Adjunct Research Fellow of Academia Sinica, Chair Professor of Providence University. He is an Area Editor of the *ACM Mobile Computing and Communication Review*, a columnist of the *ACM Simulation Digest*, an Editor of the *International Journal of Communications Systems*, *ACM/Baltzer Wireless Networks*, *Computer Simulation Modeling and Analysis*, and the *Journal of Information Science and Engineering*, and Guest Editor of a special issue on "Personal Communications" for the *ACM/Baltzer MONET*. He is the author of the book *Wireless and Mobile Network Architecture*, coauthor with Imrich Chlamtac (New York: Wiley, 2001). His current research interests include design and analysis of personal communications services network, mobile computing, distributed simulation, and performance modeling.

Dr. Lin is an Associate Editor of the IEEE NETWORK, an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and the *IEEE Communications Survey and Tutorials*, an Editor of the *IEEE Personal Communications Magazine* and *Computer Networks*, and a Guest Editor of a special issue on "Mobile Computing" for the IEEE TRANSACTIONS ON COMPUTERS, special issue on "Wireless Internet" for the IEEE TRANSACTIONS ON COMPUTERS, and special issue on "Active, Programmable, and Mobile Code Networking" for the *IEEE Communications Magazine*. He has been Program Chair for the 8th Workshop on Distributed and Parallel Simulation, General Chair for the 9th Workshop on Distributed and Parallel Simulation, and Program Chair for the 2nd International Mobile Computing Conference.

Hsien-Ming Tsai (S'97–M'02) was born in Tainan, Taiwan, R.O.C., in 1973. He received the double B.S. degrees in computer science and information engineering, and communication engineering, and the M.S. and Ph.D. degrees in computer science and information engineering from the National Chiao-Tung University (NCTU), Hsinchu, Taiwan, in 1996, 1997, and 2002, respectively.

He is currently a Research Specialist with Quanta Research Institute, Quanta Computer, Inc., Taoyuan, Taiwan. His research interests are in the areas of cellular protocols (UMTS/GPRS/GSM/DECT), cellular Internet, cellular multimedia, and embedded systems.

Prathima Agrawal (S'74–M'77–SM'85–F'89) received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1977.

She is an Assistant Vice President of the Network Systems Research Laboratory and Executive Director of the Mobile Networking Research Department, Telcordia Technologies, Morristown, NJ, where she has worked since 1998. She was the Head of the Networked Computing Research Department, AT&T/Lucent Bell Laboratories, Murray Hill, NJ, where she worked from 1978 to 1998 in various capacities. Concurrently, for several years she was an Adjunct Faculty in the Electrical and Computer Engineering Department, Rutgers University, Piscataway, NJ. She has published over 150 papers and holds 30 patents. Her research interests are computer networks, mobile and wireless computing, and communication systems.

Dr. Agrawal is the recipient of the Distinguished Member of Technical Staff Award of AT&T Bell Laboratories in 1985, the Telcordia CEO Award in 2000, and the 2001 SAIC ESTC (Executive Science and Technology Council) Publication Award. She is the recipient of the IEEE Computer Society's Distinguished Service Award in 1990 and the IEEE Third Millennium Medal in 2000. She has chaired the IEEE Fellow Selection Committee during 1998–2000. She is a Member of the Association for Computing Machinery (ACM)