An All-IP Approach for UMTS Third-Generation Mobile Networks

Yi-Bing Lin and Ai-Chun Pang, National Chiao Tung University
Yieh-Ran Haung Industrial Technology Research Institute
Imrich Chlamtac, The University of Texas at Dallas

Abstract

This article describes the UMTS all-IP approach for third-generation mobile systems, with emphasis on the core network architecture. Following the introduction of the core network nodes, we elaborate on application-level registration, circuit-switched call origination, packet-switched call origination, and packet-switched call termination.

Next-generation telecommunications networks will provide global information access for users with mobility, which is anticipated to be achieved through integration of the Internet and third-generation (3G) wireless communication. The Internet provides ubiquitous connectivity for data communications, and it has become the most important vehicle for global information delivery. The flat-rate tariff structure and low entry cost of the Internet environment encourage global use. Furthermore, the recent introduction of 3G mobile systems has driven the Internet into new markets that support mobile users. As consumers become increasingly mobile, they will demand wireless access to services available from the Internet. The mobility, privacy, and immediacy offered by wireless access commonly create new opportunities for Internet business. Therefore, mobile networks are becoming a platform that provides leading-edge Internet services.

An All IP Architecture

The 3G Partnership Project (3GPP) [1–3] proposed the *Universal Mobile Telecommunication System* (UMTS) all-IP architecture to integrate IP and wireless technologies. This architecture evolved from the second-generation (2G) Global System for Mobile Communications (GSM), *General Packet Radio Service* (GPRS), *UMTS Release 1999* (UMTS R99), and *UMTS Release 2000* (UMTS R00). UMTS Release 2000 has been split up into Releases 4 and 5. Release 4 introduces a next-generation network architecture for the *circuit-switched* (CS) domain. Release 5 introduces the *IP multimedia* (IM) subsystem on top of the *packet-switched* (PS) domain. The evolution from UMTS R99 to all-IP network has the following benefits. First, mobile networks will benefit directly not only from existing Internet applications, but also from the huge momentum behind the Internet in terms of development and introduction of new services. Second, this evolution allows telecommunications operators to deploy a common backbone (e.g., IP) for all types of access, and thus greatly reduce capital and operating costs. Third, the new generation of applications will be developed in an all-IP environment, which guarantees optimal synergy between the ever-growing mobile world and Internet.

In this article we assume that the reader is familiar with GSM, GPRS, and UMTS R99 terminology. For details of these technologies the reader is referred to [4, 5, and references therein]. In the UMTS all-IP network, Switching System No. 7 (SS7) transport will be replaced by IP, and the common IP technology supports all services including multimedia and voice services controlled by *Session Initiation Protocol* (SIP) [6]. In UMTS R99, the PS domain supports packet-switched data services over the enhanced GPRS network, and the CS domain mainly supports voice-based services. On the other hand, the UMTS all-IP network supports voice applications through the PS domain using SIP. However, the CS domain call control mechanism in R99 may be reused to support CS domain services for the UMTS all-IP network. This article provides a tutorial on the UMTS all-IP approach. Two options exist for the UMTS all-IP network. *Option 1* architecture supports PS domain multimedia and data services. *Option 2* architecture extends the option 1 network by accommodating CS domain voice services over a packet-switched core network.

Option 1 for All IP Architecture

The UMTS all-IP network architecture option 1 consists of the following segments (Fig. 1).

The radio access network (RAN; Fig. 1a) can be a *UMTS Terrestrial Radio Access Network* (UTRAN) or *GSM Enhanced Data Rates for Global Evolution (EDGE) radio access network* (GERAN). The UTRAN is basically the same as the R99 version. Details on RANs are beyond the scope of this article; more information can be found in [5, 7, 8].

The home subscriber server (HSS; Fig. 1b) is the master database containing all 3G user-related subscription information. The HSS consists of:

- The IM functionality (i.e., IM user database)
- The subset of the *home location register* (HLR) functionality required by the PS domain (i.e., 3G GPRS HLR)
- The subset of the HLR functionality (i.e., 3G CS HLR) [5, 9] to support CS domain call handling entities

We briefly describe these functionalities later.

The GPRS network (Fig. 1c) consists of *serving GPRS support nodes* (SGSNs) and *gateway GPRS support nodes* (GGSNs) that provide mobility management and *Packet Data Protocol* (PDP) context activation services to mobile users [5]. An SGSN connects to the RAN, and a GGSN connects to the external *packet data network* (PDN). The GPRS network

0890-8044/02/$17.00 © 2002 IEEE

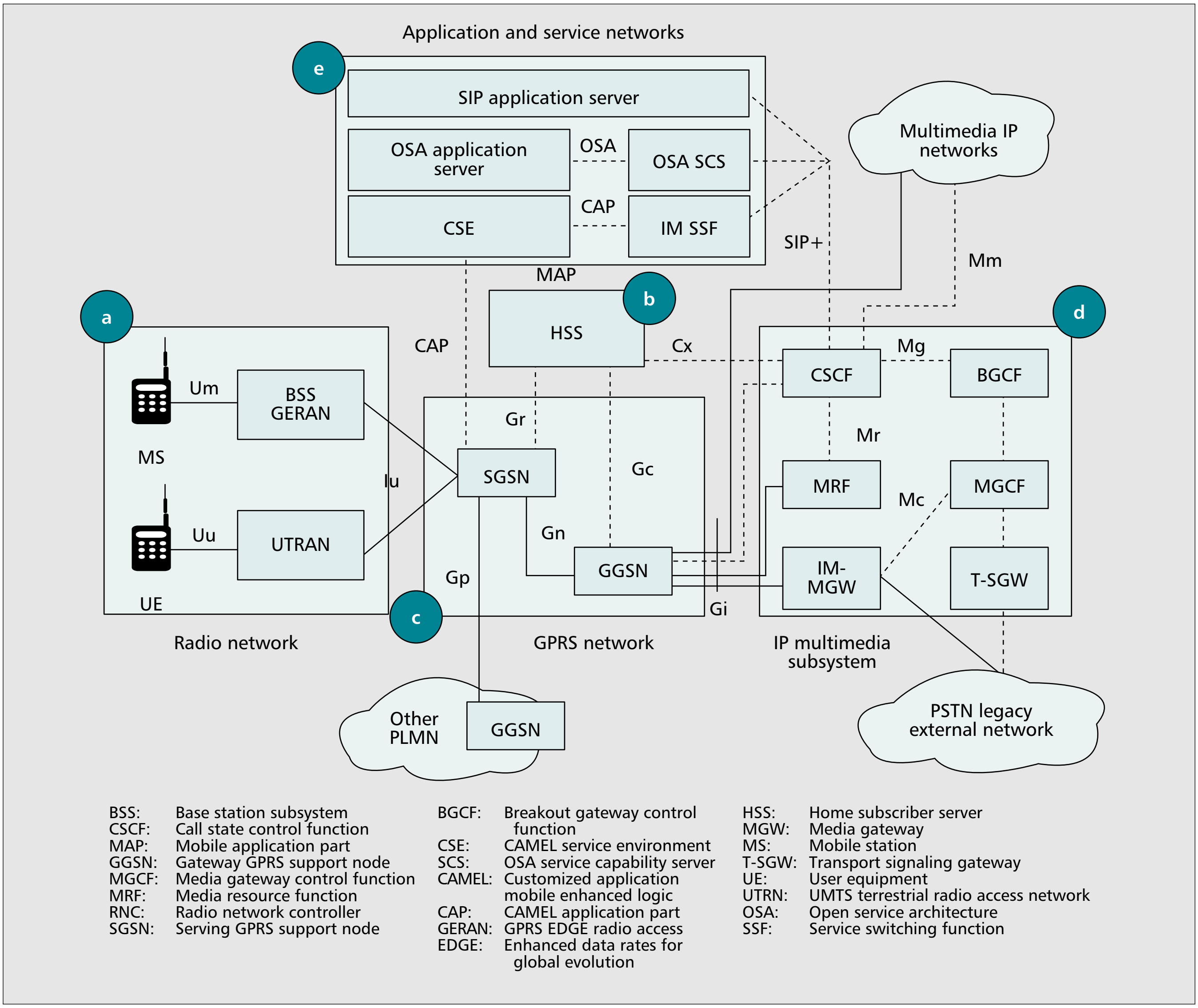

Application and service networks
e
SIP application server
OSA application server
OSA
OSA SCS
CSE
CAP
IM SSF
Multimedia IP networks
SIP+
Mm
MAP
b
HSS
a
d
CAP
Cx
Mg
CSCF
BGCF
Um
BSS GERAN
MS
Gr
SGSN
Gc
Mr
Iu
MRF
Mc
MGCF
Uu
UTRAN
Gn
GGSN
IM-MGW
T-SGW
Gp
UE
c
Gi
Radio network
GPRS network
IP multimedia subsystem
Other PLMN
GGSN
PSTN legacy external network
BSS: Base station subsystem
CSCF: Call state control function
MAP: Mobile application part
GGSN: Gateway GPRS support node
MGCF: Media gateway control function
MRF: Media resource function
RNC: Radio network controller
SGSN: Serving GPRS support node
BGCF: Breakout gateway control function
CSE: CAMEL service environment
SCS: OSA service capability server
CAMEL: Customized application mobile enhanced logic
CAP: CAMEL application part
GERAN: GPRS EDGE radio access
EDGE: Enhanced data rates for global evolution
HSS: Home subscriber server
MGW: Media gateway
MS: Mobile station
T-SGW: Transport signaling gateway
UE: User equipment
UTRN: UMTS terrestrial radio access network
OSA: Open service architecture
SSF: Service switching function

Figure 1. *UMTS all-IP network architecture (option 1).*

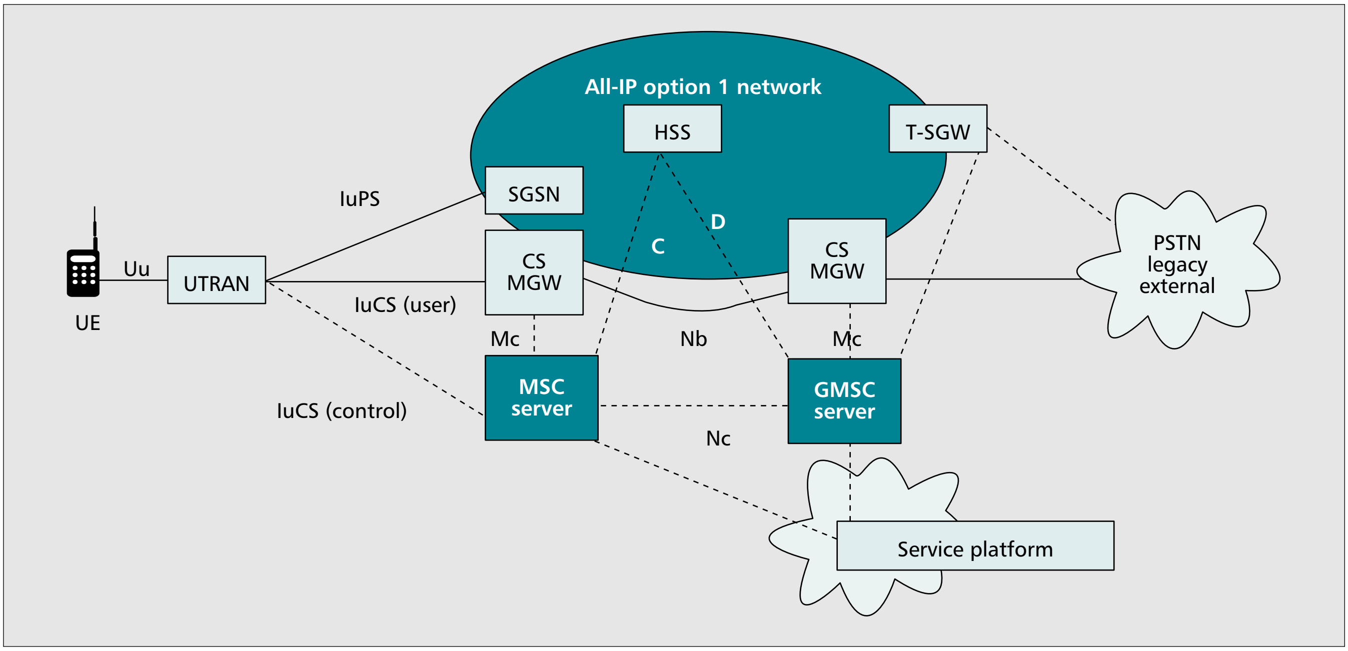
interfaces with a variety of RANs such as UTRAN and EDGE. The *Iu* interface between UTRAN and SGSN is IP based. Both the SGSN and GGSN communicate with 3G the GPRS HLR through *Gr* and *Gc* interfaces, respectively. These two interfaces are based on the *Mobile Application Part* (MAP) [5, 10]. SGSN communicates with GGSN through the *Gn* interface in the same network, and through the *Gp* interface in different networks. GGSN interacts with an external PDN through the *Gi* interface. *Gn*, *Gp*, and *Gi* are standard GPRS interfaces and are described in [5, 11].

The IM subsystem (Fig. 1d) is located behind the GGSN. In this subsystem, the *call state control function* (CSCF) is a SIP server, which is responsible for call control. Other nodes in the IM subsystem include the *breakout gateway control function* (BGCF), *media gateway control function* (MGCF), *IM-media gateway function* (IM-MGW), and *transport signaling gateway function* (T-SGW). These nodes are typically used in a voice over IP (VoIP) network [12, 13], and will be elaborated on in subsequent sections. Most interfaces among these nodes (i.e., *Mc*, *Mg*, *Mh*, *Mm*, *Mr*, and *Ms*) are IP-based gateway control protocols. We briefly describe these interfaces later.

The application and service network (Fig. 1e) supports flexible services through a service platform. The all-IP network architecture will provide a separation of service control from call/connection control, and the applications are implemented in dedicated application servers that host service-related databases or libraries. As shown in Fig. 1e, 3GPP defines three possible ways to provide flexible and global services.

- Direct SIP+[1] link between CSCF and SIP Application Server: This method will be used by mobile operators to provide new multimedia SIP applications. The SIP application services are either developed by the mobile operators or purchased from trusted third parties.
- SIP+ link between CSCF and *IM-service switching function* (IM-SSF) followed by *customized application mobile enhanced logic* (CAMEL) *application part* (CAP) link between IM-SSF and *CAMEL service environment* (CSE) [14]: This method will be used by mobile operators to provide popular CAMEL services (e.g., prepaid service) to the IM domain users. Note that similar services for the CS domain have already been provided via the CAMEL platform.
- SIP+ link between CSCF and *open service architecture* (OSA) *service capability server* (SCS) followed by OSA link between OSA SCS and OSA application server: This

[1] *SIP+ is SIP with extensions for service control to be defined later.*



■ Figure 2. *UMTS all-IP network architecture (option 2).*

method will be used to give third parties controlled access to the operator's network and let third parties run their own applications (in the third party application servers) using the IM capabilities of the operator's network.

The mobile terminals or *user equipment* (UE) are significantly influenced by the Internet and content availability. The contents for 3G services and applications need to scale to various display sizes of UEs, and also require interoperability to ensure wide end-user acceptance. Thus, a large variety of UEs targeted at different market segments will emerge. Multimode and multiband UEs will be the first step in the transition from 2G to 3G. This migration phase means that initial service quality may not be globally consistent [15]. In the UMTS all-IP network architecture, the GGSN is considered the border of the network toward the public IP network. The GGSN and MGW together are the network border toward the *public switched telephone network* (PSTN) and legacy mobile networks.

Option 2 for All-IP Architecture

All-IP network option 2, which supports R99 CS UEs, allows the R99 CS and PS domains to evolve independently. The UMTS all-IP network architecture option 2 is shown in Fig. 2. Two control elements, the MSC and GMSC servers, are introduced in option 2. The MSC servers and the HLR functionality in the HSS provide an evolution of R99 telephony services. MAP is the signaling interface between the HSS and the MSC server (GMSC server).

The R99 Iu interface separates transport of user data from control signals. Evolving from this interface, the option 2 UTRAN accesses the core network via a CS-MGW (user plane) separated from the MSC server (control plane). UTRAN communicates with MSC server using the *RAN application part* (RANAP) over the *Iu* interface. The *Iu* interface between UTRAN and CS-MGW is based on the *Iu user plane* (UP) protocol [16]. Notice that there are one or more CS-MGWs in the option 2 network. If more than one CS-MGWs exist, they communicate through the *Nb* interface. In our example, there are two CS-MGWs in the all-IP option 2 architecture (Fig. 2), one of which is connected to the PSTN and the other to the UTRAN via the *Iu*-CS interface. These two CS-MGWs are responsible for voice format conversion between PS and CS networks.

The Partitioning of All-IP Architecture in Horizontal Layers

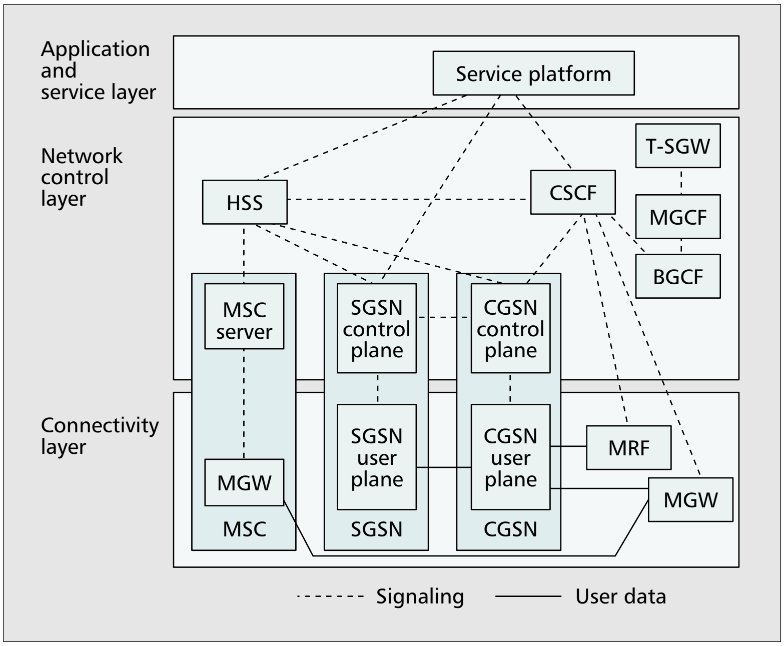
The all-IP network architecture can also be partitioned horizontally into three layers (Fig. 3):

- **The application and service layer** consists of service nodes (same as Fig. 1e).
- **The network control layer** is responsible for control signaling delivery, which consists of MSC server, HSS, CSCF, BGCF, MGCF, SGSN (control plane part), GGSN (control plane part), and T-SGW. In this layer, CSCF, MGCF, MSC server, and BGCF can serve as 3G call agents.
- **The connectivity layer** is a pure transport mechanism, capable of transporting any type of information via voice, data, and multimedia streams. The layer includes MGW, SGSN (user plane part), GGSN (user plane part), and MRF.

In UMTS all-IP option 1, the RAN (e.g., UTRAN) and GPRS network together are called the *bearer network*. Through the *Gm interface* (which includes radio, Iu, Gn, and Gi), the bearer network provides bearers for signaling (control plane) and data (user plane) exchange between the UE and the CSCF/gateways. Signaling between the bearer network and the PSTN is delivered through the CSCF, BGCF, MGCF, and T-SGW. The user plane bearer is connected to the PSTN via an MGW. The bearer network nodes (RAN, SGSN, and GGSN) are not aware of the multimedia signaling between the UE and the CSCF. However, the RAN may optimize the radio transmission by supporting specific *radio access bearers* (RABs) for individual flows of the multimedia user plane. These RABs are requested by the UE at PDP context activation.

All-IP Network Core Network Nodes

This section describes the nodes in the network control and connectivity layers. Among these network nodes, the GGSNs and SGSNs are basically the same as in R99 [4]. In all-IP network option 2, the MSC server controls CS services, and the SGSN controls PS bearer services. In the connectivity layer, the MGW uses open interfaces to connect different types of nodes. Just like the MSC structure in Fig. 3, the SGSN server handles control layer functions for PS domain communication. The media gateway provides user plane data transmission in the connectivity layer.



■ Figure 3. *The orizontal structure of a UMTS all-IP network.*

Call State Control Function

The CSCF communicates with the HSS for location information exchange and handles control layer functions related to application-level registration and SIP-based multimedia sessions. Through the *Mm interface*, the CSCF processes call requests from other VoIP call control servers or terminals in multimedia IP networks. The CSCF consists of the following logical components.

- **The incoming call gateway** (ICGW) communicates with the HSS to perform routing of incoming calls. The ICGW may trigger incoming call service (for call screening or call forwarding) and query address handling through the address handling component.
- **The call control function** (CCF) is responsible for call setup and call event report for billing and auditing. It receives and processes application level registration, provides service trigger mechanism (service capabilities features; toward application and service networks, and may invoke location-based services related to the serving network. It also checks whether the requested outgoing communication is allowed given the current subscription. The CCF interacts with the MRF through the *Mr* interface to support multiparty and other services (e.g., tones and announcements).
- **The serving profile database** (SPD) interacts with the HSS in the home network to receive profile information for the all-IP network and may store them depending on the *service level agreement* (SLA) with the home network. The SPD notifies the home network of initial user's access (includes, e.g., CSCF signaling transport address).
- **Address handling** (AH) analyzes and translates addresses. It supports address portability and alias address mapping (e.g., mapping between E.164 number and transport address). It may perform temporary address handling for internetwork routing.

There are three kinds of CSCFs: interrogating, proxy, and serving. The i*nterrogating CSCF* (I-CSCF) determines how to route mobile terminated calls to the destination UE. That is, the I-CSCF is the contact point for the home network of the destination UE, which may be used to hide the configuration, capacity, and topology of the home network from the outside world. When a UE attaches to the network and performs PDP context activation, a *proxy* CSCF (P-CSCF) is assigned to the UE. The P-CSCF contains limited CSCF function (i.e., address translation functions) to forward the request to the I-CSCF at the home network. Authorization for bearer resources in a network is performed by a P-CSCF within that network. By exercising the application-level registration described later, a *serving* CSCF (S-CSCF) is assigned to serve the UE. Through the Gm interface, this S-CSCF supports the signaling interactions with the UE for call setup and supplementary services control (e.g., service request and authentication). The S-CSCF provides SPD and AH functionality to the UE. Details of proxy, interrogating, and serving CSCFs are provided later.

Home Subscriber Server

The *home subscriber server* (HSS) keeps a master list of features and services (i.e., user profile information including user identities, subscribed services, numbering, and addressing information) associated with a user, and maintains the location of the user. The HSS provides the HLR functionality required by the PS and CS domains, and the IM functionality required by the IM subsystem to support the network entities (e.g., SGSN, GGSN, MSC server, and CSCF) actually handling calls. In other words, the HSS serves the following functionalities.

MAP termination. The HSS provides the HLR functionality to support an all-IP network. It stores mobility management and intersystem location information. Like the R99 HLR, the HSS communicates with the SGSN/GGSN in the PS domain and the GMSC/MSC server in the CS domain through interfaces Gr/Gc and C/D, respectively. Unlike the R99 HLR, these interfaces for the HSS are MAP transported over IP.

Addressing protocol termination. The HSS provides logical-name-to-transport-address translation for answering *Domain Name Server* (DNS) queries.

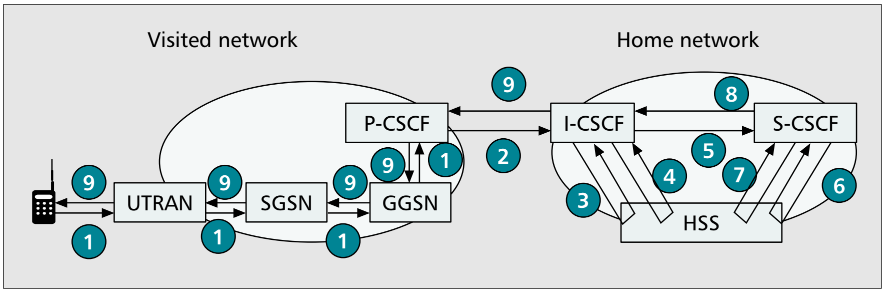
Authentication and authorization protocol termination. The HSS may also generate, store, and manage security data and policies used in the IM subsystem. CSCF-UE security parameters are sent from the HSS to the CSCF, which allow the CSCF and UE to communicate securely. The HSS stores the all-IP network service profiles and service mobility or S-CSCF-related information for the UE. The HSS also provides the S-CSCF with the service parameters (e.g., supplementary service parameters, application server address, triggers) of UE. The HSS communicates with a CSCF using Cx that can be implemented via MAP.

Other Network Nodes

This subsection describes the BGCF, MGCF, MSC server, and T-SGW in the network control layer, and MRF and MGW in the connectivity layer.

The breakout gateway control function (BGCF) is responsible for selecting an appropriate PSTN breakout point based on the received SIP request from the S-CSCF. If the BGCF determines that a breakout is to occur in the same network, the BGCF selects an MGCF that is responsible for interworking with the PSTN. If the breakout is in another network, depending on the configuration, the BGCF forwards this SIP request to another BGCF or an MGCF in the selected network.

The media gateway control function (MGCF) is the same as the MGC in a VoIP network [12], which controls the connection for media channels in an MGW. The MGCF communicates with CSCF through SIP over the *Mg interface*. It selects an appropriate CSCF, depending on the routing number for incoming calls from the legacy networks. The MGCF should support different call models.



■ Figure 4. *Application-level registration.*

The MSC (GMSC) server supports the MGC protocol to handle control layer functions related to CS domain services at the borders between the RAN and the UMTS all-IP core network, and between the PSTN and the UMTS all-IP core network. It comprises the call control and mobility control parts of UMTS R99 MSC (GMSC). An MSC (GMSC) server and associated MGW can be implemented as a single node that is equivalent to an R99 MSC (GMSC). For call setup, the GMSC server communicates with the MSC server through the *ISDN User Part* (ISUP) protocol [5] in the *Nc* interface.

The transport signaling gateway function (T-SGW) serves as the PSTN signaling termination point and provides the PSTN/legacy mobile network to IP transport level address mapping, which maps call-related signaling from/to the PSTN on an IP bearer and sends it to/from the MGCF or GMSC server. The T-SGW does not interpret the messages in the MAP or ISUP layer.

The media resource function (MRF) performs functions such as multiparty call, multimedia conferencing, and tone and announcement. Through the Mr interface, the MRF communicates with the S-CSCF for service validation of multiparty/multimedia sessions.

The media gateway function (MGW) provides user plane data transport in the UMTS core network. The MGW terminates bearer channels from the PSTN/legacy mobile networks and media streams from a packet network (e.g., *Real-Time Transport Protocol*, RTP [17], streams in an IP network). The implementation of a UMTS MGW should be consistent with existing/ongoing industry protocols/interfaces. Through the *Mc interface* (fully compliant with H.248), the MGW interacts with the MGCF, MSC server, and GMSC server (Figs. 1 and 2) for resource control. Two MGWs can communicate through the *Nb interface* (Fig. 2) where the transport for the user plane can be RTP/UDP/IP or ATM adaptation layer type 2 (AAL2).

In the option 2 network architecture, the MGW also interfaces the UTRAN with the all-IP core net-work over the Iu interface. The MGW provides flexible connection handling that supports different call models and media processing through different Iu options for CS services (AAL2/ATM-based as well as RTP/UDP/IP-based). Media processing includes media conversion, bearer control, and payload processing (e.g., codec, echo canceller, and conference bridge). The MGW bearer control and payload processing capabilities will also need to support mobile-specific functions such as serving radio network system relocation/handoff and anchoring [17].

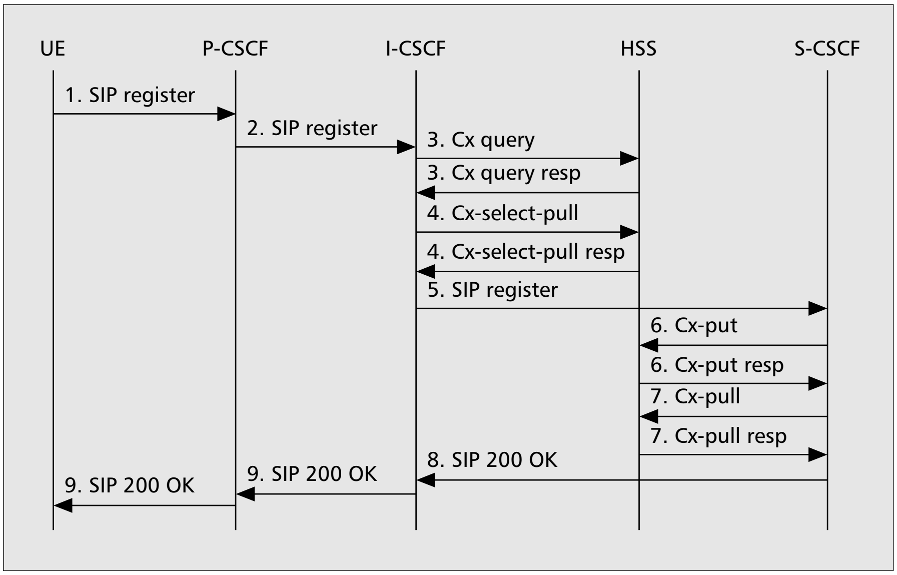
Registration and Call Control

In an all-IP network, the bearer network mobility management (for both CS and PS domains) follows UMTS R99 [17]. Specifically, mobility management procedures in the all-IP CS and PS domains are based on the *location area identifier* (LAI) and *routing area identifier* (RAI), respectively. A mechanism to convert the formats of identities among all-IP networks, R99, and 2G is needed. This section describes the application-level registration, CS and PS call originations, and PS call termination. For PS call origination and termination, we consider the scenarios where one party is the UE and the other is in the PSTN. The call agents involved in these scenarios include an S-CSCF (for the UE) and an MGCF (for the party in the PSTN). SIP is the signaling protocol for multimedia session setup, modification, and teardown. Although SIP can be used with any transport protocol, the media is normally delivered by RTP. RTP is a transport protocol on top of UDP that detects packet loss and ensures ordered delivery. An RTP packet also indicates the packet sampling time from the source media stream. The destination application can use this timestamp to calculate delay and jitter.

Application-Level Registration

In an all-IP network, UE conducts two types of registration. In *bearer-level registration*, the UE registers with the GPRS network following the standard UMTS routing area update or attach procedures [5]. After bearer-level registration, the UE can activate PDP contexts in the GPRS network. Bearer-level registration and authentication are required to support GPRS-based services. To offer IM services, *application-level registration* must be performed in the IM subsystem after bearer-level registration. In application-level registration, an S-CSCF is assigned to the UE. Specifically, the HSS interacts with the I-CSCF to determine the S-CSCF. This action is referred to as *CSCF selection* [18].

Before application-level registration can be initiated, the UE must have performed bearer-level registration to obtain an IP address and discover the P-CSCF.



■ Figure 5. *The message flow for application-level registration.*

P-CSCF discovery can be achieved, for example, through *Dynamic Host Configuration Protocol* (DHCP), which provides to the UE the domain name of the P-CSCF and the address of a DNS that can resolve the P-CSCF name. The UE stores the P-CSCF address to be used for mobile-originated signaling. The application-level registration procedure is described in the following steps (Figs. 4 and 5).

Step 1: The UE sends the SIP `REGISTER` message to the P-CSCF through the UTRAN, SGSN, and GGSN. The request includes the *home domain name* of the UE. In SIP, `REGISTER` is issued by a client (UE in our example) to the server (UMTS network) with an address at which the client can be reached for a SIP session.

Step 2: Based on the home domain name, the P-CSCF performs address translation (through a DNS-based mechanism) to find the I-CSCF address. Then it proxies the `REGISTER` message to the I-CSCF at the home network. Note that there may be multiple I-CSCFs within an operator's network.

Step 3: Based on the subscriber identity received from the P-CSCF and the home domain name, the I-CSCF determines the HSS address. Note that if there are more than one HSSs in the home network, the I-CSCF needs to query the *subscription location function* (SLF) to find the HSS address.

The I-CSCF sends the `Cx-Query` message to the HSS. The HSS checks if the subscriber has been registered. Then it returns the `Cx-Query Resp` message to the I-CSCF. By the end of this step, the user has been authenticated.

Step 4: The I-CSCF sends the `Cx-Select-Pull` message to the HSS to obtain the required S-CSCF capability information (supported service set and protocol version number). Based on the service network indication and subscriber identity provided by the I-CSCF, the location service of the HSS returns the required S-CSCF capabilities through the `Cx-Select-Pull-Resp` message.

Based on the information provided by the HSS, the I-CSCF selects the name of an appropriate S-CSCF. This S-CSCF must be in the home network.

Step 5: The I-CSCF sends the `REGISTER` request to the S-CSCF. The request includes the HSS name as a parameter.

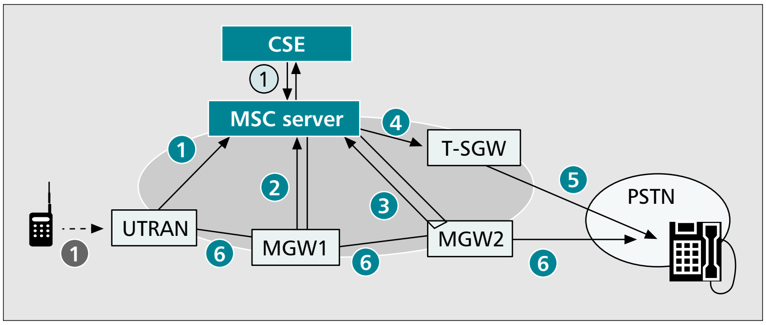
Step 6: Using `Cx-Put`, the S-CSCF sends its name and the subscriber identity to the HSS, which will be used by the HSS to route mobile terminated calls to the S-CSCF. The HSS replies the acknowledgment message `Cx-Put Resp`.

Step 7: The S-CSCF obtains the subscriber data from the HSS through the Cx-Pull and Cx-Pull-Resp exchange. The subscriber data are stored in the S-CSCF, including supplementary service parameters, application server address, triggers, and so on.

Step 8: The S-CSCF determines if the home contact name is the S-CSCF name or the I-CSCF name. If the contact name is for the S-CSCF, the P-CSCF can access the S-CSCF directly, and the internal configuration of the home network is known to the outside world. If the contact name is for the I-CSCF, the P-CSCF can only access the S-CSCF indirectly through the I-CSCF. In this case the home network configuration is hidden.

The S-CSCF sends its address and the home contact name to the I-CSCF through a `SIP OK` response message (the SIP status code is 200).

Step 9: With the `SIP OK` message, the I-CSCF returns the home contact name (either the I-CSCF or the S-CSCF address) to the P-CSCF. The P-CSCF stores the home contact name, and forwards the `SIP OK` message to the UE indicating that registration was successful.



■ Figure 6. *CS call origination.*

Notice that if the registration information expires or the registration status changes, the UE initiates *re-registration*. The re-registration procedure is basically the same as the registration procedure described above except that:

- In step 3, the HSS determines that the user is currently registered, and the S-CSCF name is sent back to the I-CSCF directly. Thus, step 4 can be omitted.
- Steps 6 and 7 may be skipped if the S-CSCF detects that this procedure is for re-registration.
- In step 8, the S-CSCF only sends the home contact name to the I-CSCF. The S-CSCF address need not be sent to the I-CSCF.

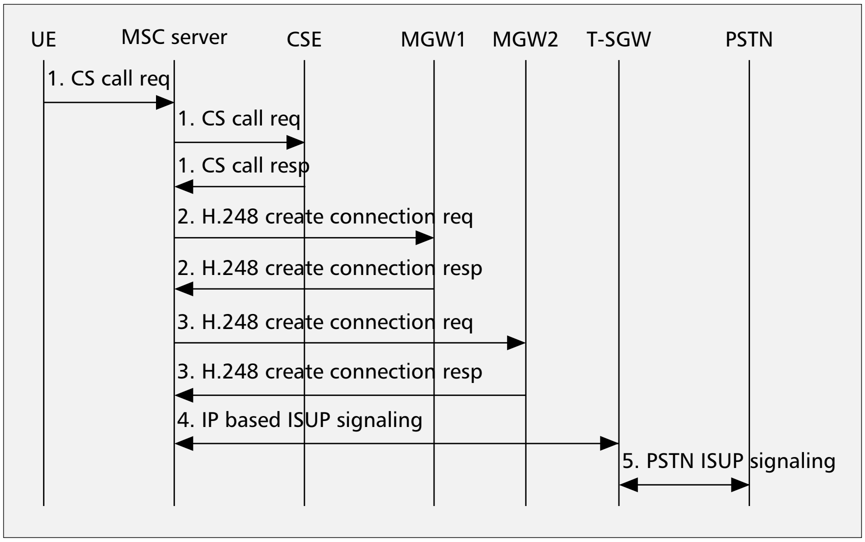
If the UE roams to a new network or is turned off, application-level deregistration is performed. Details of this procedure can be found in [18].

CS Mobile Call Origination

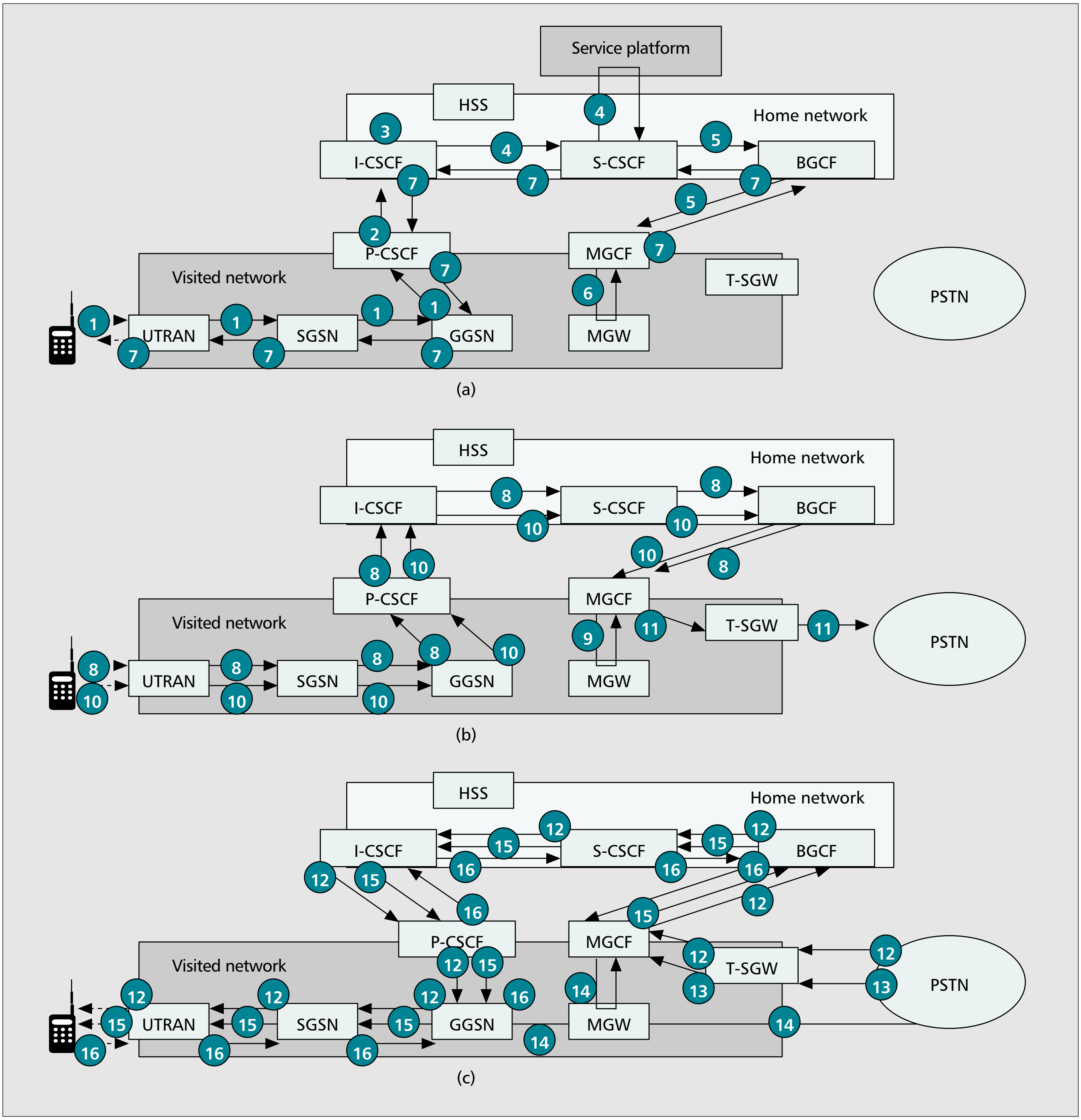
CS mobile call origination is similar to that in UMTS R99, which does not involve CSCFs and the HSS. Before CS call origination, the UE is already attached to the UMTS CS domain, and has registered to the VLR of an MSC server. The CS call origination message flow is described as follows (Figs. 6 and 7).

Step 1: The call request from the UE is forwarded to the MSC server through the UTRAN. The VLR function of the MSC server performs originating service control through the support of the CSE. Assume that the request is accepted.

Step 2: Based on the location of the UE, the MSC server selects the first media gateway, MGW1, that connects to the


UE
MSC server
CSE
MGW1
MGW2
T-SGW
PSTN
1. CS call req
1. CS call req
1. CS call resp
2. H.248 create connection req
2. H.248 create connection resp
3. H.248 create connection req
3. H.248 create connection resp
4. IP based ISUP signaling
5. PSTN ISUP signaling

■ Figure 7. *The message flow for CS call origination.*



■ Figure 8. *PS call origination: a) steps 1–7; b) steps 8–11; c) steps 12–16.*

UE through the UTRAN. This MGW is responsible for QoS provisioning and conversion between the ATM and IP protocols.

Step 3: Assume that the called party is in the PSTN. The MSC server selects the second media gateway MGW2 that serves as the termination to the PSTN. Note that this second media gateway can be MGW1 itself. Signaling based on protocols such as H.248 [19] is performed to reserve the MGW2 resources, that is, the ports to MGW1 and the ports to the PSTN.

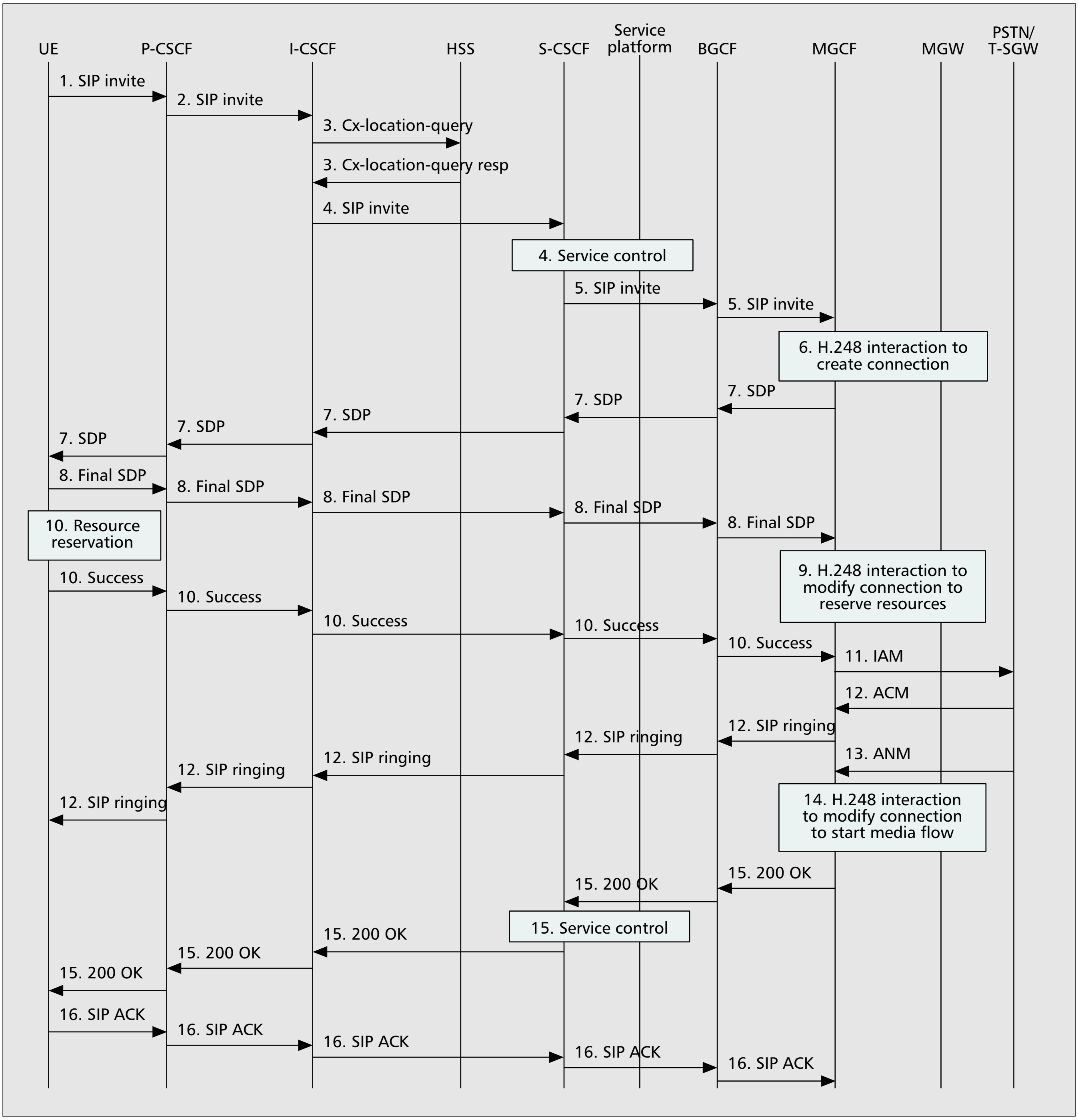
Steps 4 and 5: The MSC server and the PSTN exchange the ISUP call setup messages through the T-SGW.

When the call setup procedure is complete, the voice stream is UE ↔ UTRAN ↔ MGW1 ↔ MGW2 ↔ PSTN (see path (6) in Fig. 6).

PS Mobile Call Origination

Before PS (IM subsystem) call origination, the UE was already attached to the UMTS PS domain, and application-level registration was performed so that an S-CSCF was assigned to the UE (i.e., the user profile has been fetched from the HSS and stored in the S-CSCF). Assume that the UE is in a visited network. The PS mobile call origination to the PSTN is described as follows (Figs. 8 and 9).

Step 1: The UE sends a `SIP INVITE` request to the P-CSCF. The P-CSCF and UE must be located in the same network. The `INVITE` message is used to initiate a SIP media session with an initial *Session Description Protocol* (SDP). The SDP provides session information (e.g., RTP payload type, addresses, and ports) to potential session participants.



■ Figure 9. *The message flow for Ps call origination.*

Step 2: The P-CSCF resolves the UE's home network address (suppose that it is the I-CSCF name stored in step 9, "Application-Level Registration" section), and forwards the INVITE message to the I-CSCF.

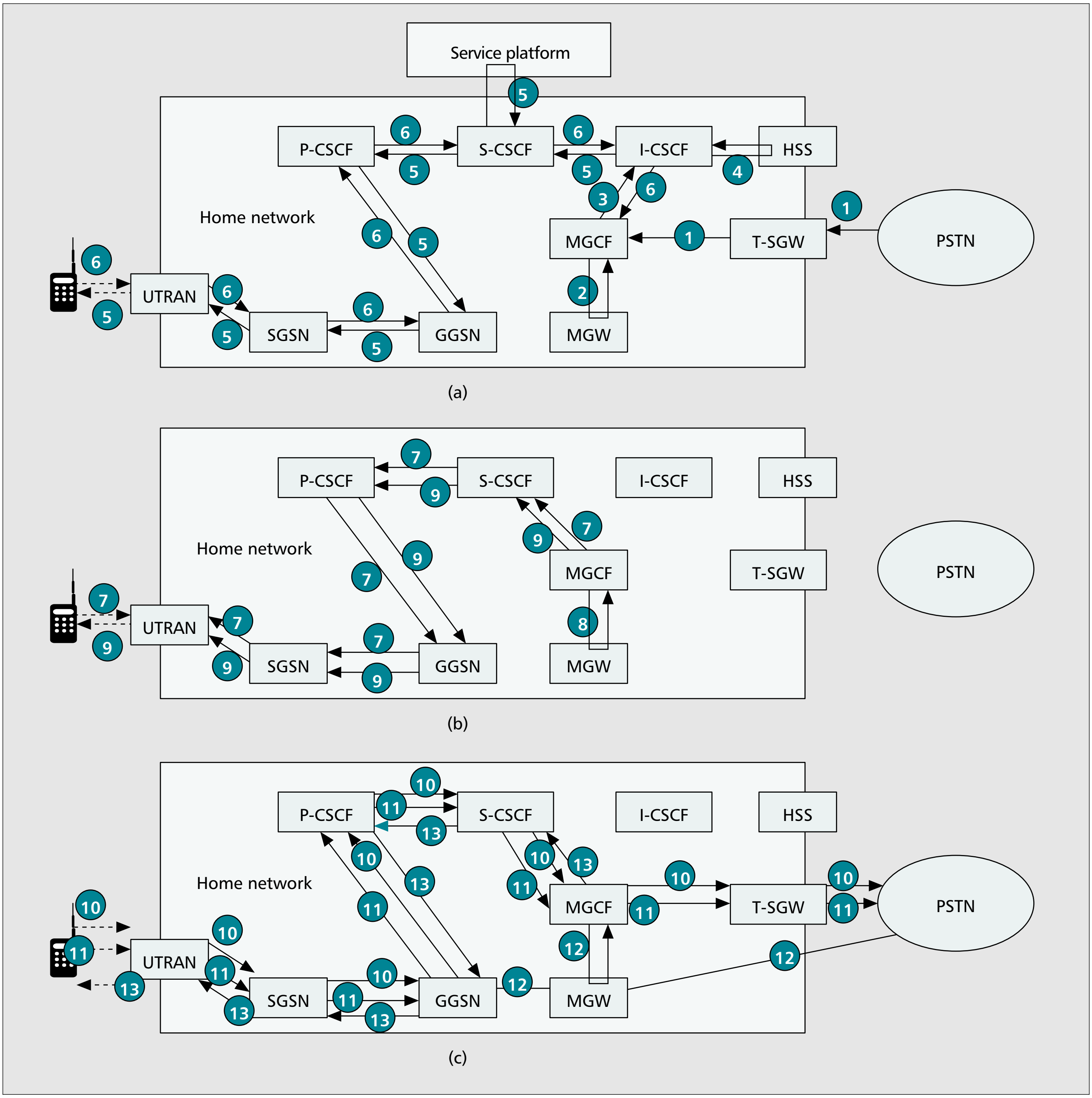
Step 3: The I-CSCF provides ICGW and AH functionality, and interrogates the location service of the HSS through the Cx-Location-Query and Resp message exchange to obtain the S-CSCF signaling transport parameters.

Step 4: The I-CSCF relays the INVITE message to the S-CSCF through the Mw interface. The S-CSCF will act as a host to the call control logic. It validates the service profile of the subscriber, and may contact the service platform to perform origination service control.

Step 5: The S-CSCF translates the destination address and determines that the call will break out to the PSTN. It therefore forwards the INVITE message to the BGCF in the home network. If the MGW is in the home network, the BGCF sends the INVITE message to the MGCF in the home network. In the case where the MGW is in the visited network, there are two possibilities. The BGCF may forward the INVITE message to the visited BGCF (which then selects an MGCF in the visited network for this call setup). Alternatively, the BGCF may directly forward the INVITE message to the MGCF in the visited network, as in this example.

Step 6: By using the H.248 protocol, the MGCF determines the MGW capabilities and allocates the MGW ports for the call connection.

Step 7: The MGCF returns the 183 SESSION IN



■ Figure 10. *PS call termination: a) steps 1–6; b) steps 7–9; c) steps 10–13.*

PROGRESS message to the P-CSCF. This message contains the SDP that indicates the media stream capabilities of the called party. The P-CSCF authorizes the required resources for this session and forwards the 183 SESSION IN PROGRESS (with SDP) message to the UE through the signaling path established by the INVITE message.

Step 8: The UE determines the final set of media streams, and sends the final SDP to the MGCF through the PRACK (provisional acknowledgment) message.

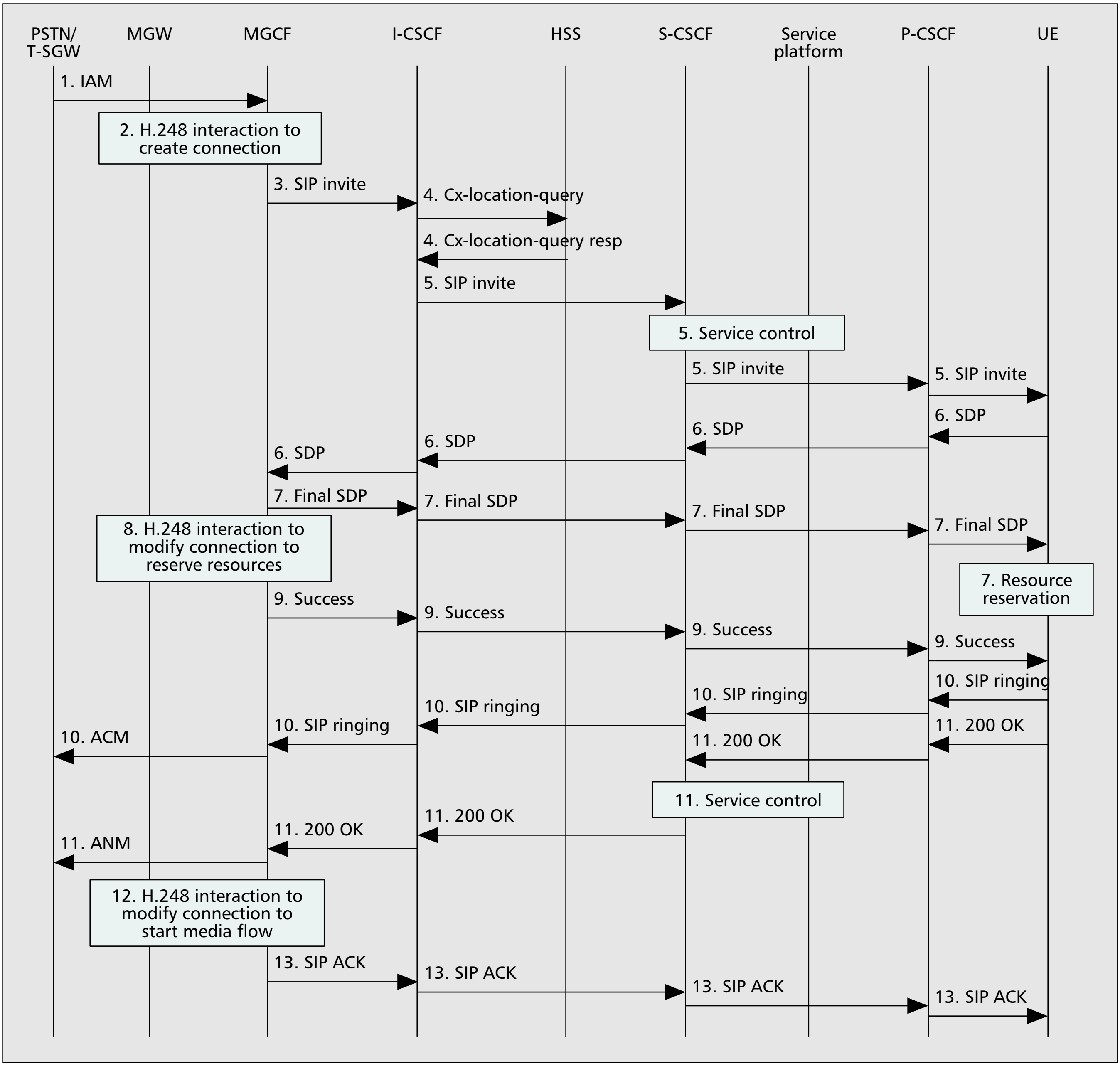
Step 9: The MGCF issues the H.248 command that instructs the MGW to reserve the required resources for the media streams.

Step 10: After Step 8, the UE reserves the resources for this session through the PDP context activation procedure. It then sends the Resource Reservation Successful message to the MGCF.

Step 11: The MGCF sends an IP IAM (Initial Address Message) to the T-SGW. The T-SGW translates the IP IAM message into the SS7 IAM message and forwards it to the PSTN. The IAM requests the PSTN to set up the PSTN call path toward the called party.

Step 12: The PSTN establishes the call path, alerts the called party, and returns the SS7 Address Complete Message (ACM) to the T-SGW. This message is translated into the IP ACM message and is forwarded to the MGCF. The MGCF sends the ring back message (SIP Ringing with status code 180) to the UE. The ACM indicates that the path to the destination has been established.

Step 13: When the called party answers, the PSTN sends the SS7 ANM (Answer Message) to the T-SGW. The SGW translates the message into the IP ANM message and forwards it to the MGCF.



■ Figure 11. *The message flow for PS call termination.*

Step 14: The MGCF instructs the MGW to make the bidirectional connection using the H.248 protocol.

Step 15: After Step 13, the MGCF sends the SIP OK response to the S-CSCF. The S-CSCF may perform service control for the call. It then forwards the SIP OK to the UE through the P-CSCF (which approves usage of the reserved resources). The UE starts the media flow for this session.

Step 16: The UE forwards the final SIP ACK message to the MGCF (through P-CSCF, I-CSCF, and S-CSCF). In SIP, if the client (i.e., the UE in our example) issuing the INVITE has received a final response, the client will reply an ACK.

Although several nodes are visited in call setup signaling, the shortest path is established for user plane media stream (i.e., UE $ UTRAN ↔ SGSN ↔ GGSN ↔ MGW ↔ PSTN).

PS Mobile Call Termination

Before PS mobile call termination, the UE was already attached to the UMTS PS domain, and has completed application-level registration. The PDP context has been established for SIP signaling message delivery. We assume that the UE is in the home network. The call termination procedure is given in the following steps (Figs. 10 and 11).

Step 1: The originating switch of the PSTN uses IAM to initiate a call. That is, the PSTN sends the SS7 IAM to the T-SGW. The T-SGW sends the IP IAM message to the MGCF in the home network.

Step 2: Using the H.248 protocol, the MGCF communicates with the MGW to reserve the resources (MGW ports) for user plane media streams. Note that the selection of the MGW and determination of the most optimal routing toward this MGW are open issues.

Step 3: The MGCF translates the destination address and determines that the called party is in the home network in our example. The MGCF sends the SIP INVITE request (including the initial SDP) to the I-CSCF.

Step 4: The I-CSCF exchanges the Cx-Location-Query and Cx-Location-Query-Resp message pair with the HSS to obtain the location information of the called party.

Step 5: The I-CSCF sends the INVITE message to the UE through S-CSCF, P-CSCF, GGSN, SGSN, and UTRAN. In this signaling flow, the S-CSCF validates the service profile, and may contact the service platform to perform termination service control.

Step 6: The INVITE message received by the UE indicates the requested media flows. Based on this information, the UE determines the resources to be allocated for the media streams. The UE then sends the 183 SESSION IN PROGRESS with SDP to the MGCF through the UTRAN, SGSN, GGSN, P-CSCF, and S-CSCF. The SDP information indicates the UE's media stream capabilities. Since different types of UE may support different media types, the UE capabilities have impact on the SDP description. In this signaling flow (which was established by the INVITE message in steps 3 and 5), the P-CSCF authorizes the necessary resources for this session.

Step 7: Based on theUE's capabilities, the MGCF determines the allocated media streams for this session, and sends the final SDP response to the UE. The UE initiates resource reservation for this session through the PDP context activation procedure [5, 11].

Step 8: After the MGCF sends the SDP in step 7, it instructs the MGW to reserve the resources for the session.

Step 9: The MGCF may perform service control as appropriate, and sends the Resource Reservation Successful message to the UE. If the UE has completed resource reservation in step 7, it alerts the subscriber (called person).

Step 10: The UE sends the SIP Ringing message (SIP response message with status code 180) to the MGCF. The MGCF sends the IP ACM message to the T-SGW, and the T-SGW sends the SS7 ACM to the PSTN. At this point, the PSTN knows that the call path toward the UE has been established.

Step 11: When the called person of the UE answers, the final SIP OK response is sent from the UE to the MGCF through the P-CSCF and S-CSCF. The UE starts the media flow for this session. In this signaling flow, the P-CSCF indicates that the reserved resources should be committed, and the S-CSCF may perform service control as appropriate. When the MGCF receives the SIP OK, it sends the IP ANM to the T-SGW, and the T-SGW sends the SS7 ANM to the PSTN.

Step 12: The MGCF sends H.248 command to the MGW, which instructs the MGW to make a bidirectional connection to the GGSN and the PSTN node.

Step 13: The MGCF then returns the SIP ACK message to the UE.

The user plane media streams are connected between the UE and the PSTN through path UE ↔ UTRAN ↔ SGSN ↔ GGSN ↔ MGW ↔ PSTN.

Efficiency of IP Packet Delivery

To support the UMTS all-IP network, IP packets must be delivered efficiently in the radio access bearer. Specifically, the UTRAN should meet the objectives of spectral efficiency and error robustness. Radio spectrum efficiency is significantly affected by IP packet overhead. The sizes of the combined packet (IP/UDP/RTP) headers are at least 40 bytes for IPv4 and at least 60 bytes for IPv6. Furthermore, the header part will require more error protection than the payload. If no error concealment or mitigation can be applied to a header, the header is lost, and the corresponding packet must be discarded.

For applications such as VoIP, the voice payload is typically shorter than the packet header. Packing more speech frames into one packet will reduce the relative overhead. However, the voice delay is increased, and thus voice quality is degraded. A more appropriate solution is to utilize *header adaptation techniques* that reduce the size of the header before radio transmission. Reduction of header size is achieved in two ways:

¶*Header compression* removes redundancy in the originally coded header information. In the UMTS all-IP network, header compression schemes must be developed based on the radio link reliability characteristics (i.e., high error rates and long round-trip times). Such schemes may be able to compress the packet headers down to 2 bytes [1]. To exercise header compression between the UE and the network, each maintains a consistent compressor and decompressor. Initially, uncompressed headers are transmitted. For subsequent packets, compressed headers (the "differences" from previous headers) are delivered over the air. The decompressor uses the received compressed information and knowledge of previous headers to reconstruct the next headers. Header compression should be efficient (the average header size is minimized), robust (no packet is lost due to header compression), and reliable (the decompressed header is identical to the header before compression) [20]. A major disadvantage of this technique is that compressed headers have variable sizes, which introduces extra overhead for end-to-end security (IPsec) and bandwidth management. Note that the header compression mechanism is typically implemented in the UTRAN.

¶*Header stripping* removes header field information. Packet headers are stripped before radio transmission and regenerated at the receiving end. Essentially only the payload is transmitted, but some additional header-related information needs to be transmitted to enable header regeneration. The degree of header transparency depends on the amount of transmitted header-related information. No header error protection is needed when header information is completely removed. When the payload is of constant size, the bandwidth management issue can be simplified since the payloads can be carried on a constant bit rate channel. This approach also mitigates QoS issues such as delay and jitter. Notice that the header stripping approach is typically adopted in the GERAN.

In selecting the header reduction solution, one should consider the impact on transparency and robustness to errors. Three approaches are considered for *user plane adaptation* (UPA):

- **Full opacity (no adaptation)**: The original packet header is sent over the air to achieve full transparency. The UPA supports IPsec on an end-to-end basis. High overhead of the header results in very poor spectrum efficiency.
- **Payload opacity (header adaptation only)**: The header is compressed or stripped. The UPA understands the header structure but not the payload.
- **No opacity (full adaptation)**: The UPA knows the structure of the headers and the payload. A header can be compressed or stripped. Payload transmission is optimized by techniques such as unequal bit protection. Channel and error coding are optimized for the payload structure.

To support VoIP, we may use header compression with equal bit protection in the payload.

Summary

This article describes the UMTS all-IP approach for 3G mobile systems. Our discussion focuses on the core network architecture. We first describe the core network nodes, then elaborate on application-level registration, circuit-switched call origination, packet-switched call origination, and packet-switched call termination.

UMTS should guarantee end-to-end QoS and radio spectrum efficiency. To provide the expected QoS across domains, operators must agree on the deployment of common IP protocols. The common IP protocols impact roaming (i.e., the interfaces between core networks) and the communications between terminals and networks in order to support end-to-end QoS and achieve maximum interoperability. QoS signaling and resource allocation schemes should be independent of the call control protocols, and the details can be found in [5, 18, 21]. To support radio spectrum efficiency, IP header compression or stripping is required, as discussed earlier.

As a final remark, we address some of the business issues for UMTS based on the materials provided in [15, 22, 23]. The wireless Internet services offered by UMTS (e.g., location-based services) will influence people's lifestyles, and can be expected to create new telecommunications paradigms, enable new technologies, and spur future demand. It is clear that we cannot directly apply the fixed Internet model to wireless Internet. The model of "almost-free" fixed Internet access should be re-investigated. Wireless Internet needs a new model so that end users will be willing to pay for secure convenient wireless data services. Accurate customer segmentation, customer value, and availability of services at the right price will be key factors to success. Specifically, prices should be pitched at affordable levels for the target segments. Profitability will be strongly dependent on packaging of applications and presentation of content based on the characteristics of people in different regions of the world with their cultural variety and particular needs.

Another drive for customers to use 3G services is transparency of charging that allows better cost control via easy-to-use interfaces. There must be a strong link between user-perceived QoS and charging. QoS support across multiple networks will require new forms of commercial agreements between operators, which will be radically different from the traditional peer-to-peer agreements from the Internet world or the roaming agreements known from the 2G mobile world.

Acknowledgment

We would like to thank the anonymous reviewers whose valuable comments have significantly improved the quality of this article. Lin's work was sponsored in part by MOE Program of Excellence Research under contract 89-E-FA04-4, TAHOE Network, Ericsson, InterVideo, FarEastone, National Science Council under contract NSC 90-2213-E-009-156, the Lee and MTI Center for Networking Research, NCTU.

References

[1] 3GPP, "Services and Systems Aspects, "Architectural for an All IP Network," Tech. rep. 3G TR 23.922 v. 1.0.0 (1999–10), 1999.

[2] 3GPP, "Services and Systems Aspects; Network Architecture," Tech. spec. 3G TS 23.002 version 5.3.0 (2001–06), 2001.

[3] L. Bos and S. Leroy, "Toward an All-IP-based UMTS System Architecture," *IEEE Network*, vol. 15, no. 1, 2001, pp. 36–45.

[4] Y.-B. Lin *et al.*, Mobility Management: From GPRS to UMTS, to appear, *WL Commun. and Mobile Comp.*, 2001.

[5] Y.-B. Lin and I. Chlamtac, *Wireless and Mobile Network Architectures*, Wiley, 2001.

[6] M. Handley *et al.*, "SIP: Session Initiation Protocol," IETF RFC 2543, Aug. 2000.

[7] H. Holma and A. Toskala, edited, *WCDMA for UMTS*, John Wiley & Sons, 2000.

[8] 3GPP, "Radio Access Network; UTRAN Overall Description," Tech. spec. 3G TS 25.401, v. 4.1.0 (2001–06), 2001.

[9] 3GPP, "Core Network; Subscriber Data Management; Stage 2," Tech. spec. 3G TS 23.016, v. 4.0.0 (2001-03), 2001.

[10] ETSI/TC, "Mobile Application Part (MAP) Specification," v. 7.3.0, Tech. rep. GSM 09.02, 2000.

[11] 3GPP, 3rd Generation Partnership Project, "Services and Systems Aspects; General Packet Radio Service (GPRS); Service Descripton; Stage 2," Tech. spec. 3G TS 23.060, v. 4.1.0 (2001-06), 2001.

[12] M. Hamdi *et al.*, "Voice Service Interworking for PSTN and IP Networks," *IEEE Commun. Mag.*, May 1999, pp. 104–11.

[13] H. Schulzrinne and J. Rosenberg, "The IETF Internet Telephony Architecture," *IEEE Commun. Mag.*, May 1999, pp. 18–23.

[14] 3GPP, "Series and System Aspects; Customized Applications for Mobile Network Enhanced Logic (CAMEL); Service Description (Stage 1)," Tech. spec. 3G TS 22.078, v. 5.1.0 (2001–01), 2001.

[15] UMTS Forum, "Enabling UMTS/Third Generation Services and Applications," Tech. rep. 11, 2000; http://www.umts-forum.org

[16] 3GPP, "Radio Access Network; UTRAN Iu Interface User Plane Protocols," Tech. spec. 3G TS 25.415, v. 4.1.0 (2001–06), 2001.

[17] H. Schulzrinne *et al.*, "RTP: A Transport Protocol for Real-Time Applications," IETF RFC 1889, Jan. 1996.

[18] 3GPP, "Services and Systems Aspects; IP Multimedia Subsystem Stage 2," Tech. spec. 3G TS 23.228, v. 5.1.0 (2001-06), 2001.

[19] ITU tech. rep. ITU-T H.248, "Gateway Control Protocol," 2000.

[20] K. Svanbro, "Voice over IP over Wireless: Performance and Principles," contact krister.svanbro@epl.ericsson.se

[21] 3GPP, "Services and Systems Aspects; End-to-End QoS Concept and Architecture," Tech. spec. 3G TS 23.207, v. 5.0.0 (2001-06), 2001.

[22] UMTS Forum, "The UMTS Third Generation Market – Structuring the Service Revenues Opportunities," Tech. rep. 9, 2000; http://www.umts-forum.org

[23] UMTS Forum, "Shaping the Mobile Multimedia Future – An Extended Vision from the UMTS Forum," Tech. rep. 10, 200; http://www.umts-forum.org

Biographies

YI-BING LIN (liny@csie.nctu.edu.tw)received his B.S.E.E. degree from National Cheng Kung University in 1983, and his Ph.D. degree in computer science from the University of Washington in 1990. From 1990 to 1995 he was with the Applied Research Area at (Bellcore, Morristown, New Jersey. In 1995 he was appointed a professor of the Department of Computer Science and Information Engineering (CSIE), National Chiao Tung University (NCTU). In 1996 he was appointed deputy director of the Microelectronics and Information Systems Research Center, NCTU. During 1997–1999 he was elected chair of CSIE, NCTU. His current research interests include design and analysis of personal communications services networks, mobile computing, distributed simulation, and performance modeling. He is an associate editor of *IEEE Network*, an editor of *IEEE J-SAC: Wireless Series*, *IEEE Wireless Communications*, and *Computer Networks*, an area editor of *ACM Mobile Computing and Communication Review*, a columnist in *ACM Simulation Digest*, as well as an editor of *International Journal of Communications Systems*, *ACM/Baltzer Wireless Networks*, *Computer Simulation Modeling and Analysis*, and *Journal of Information Science and Engineering*, Program Chair of the 8th Workshop on Distributed and Parallel Simulation, General Chair of the 9th Workshop on Distributed and Parallel Simulation, Program Chair for the 2nd International Mobile Computing Conference, Guest Editor for an *ACM/Baltzer MONET* Special Issue on Personal Communications, a Guest Editor for an *IEEE Transactions on Computers* Special Issue on Mobile Computing, and a Guest Editor for an *IEEE Communications Magazine* Special Issue on Active, Programmable, and Mobile Code Networking. He is the co-author with Imrich Chlamtac of the book *Wireless and Mobile Network Architecture* (Wiley). He received 1998 and 2000 Outstanding Research Awards from National Science Council, ROC, and 1998 Outstanding Youth Electrical Engineer Award from CIEE, ROC. He is an adjunct research fellow of Academia Sinica.

YIEH-RAN HAUNG received his B.S. degree in computer science from SooChow University, Taiwan, in 1991, and his M.S. and Ph.D. degrees in computer science and information engineering from National Chiao Tung University, Taiwan, in 1993 and 1997, respectively. From 1997 to 1999 he was a postdoctoral research fellow in the Institute of Information Science, Academia Sinica, Taipei, Taiwan. There, he was involved in R&D in QoS router and wireless Internet access. Currently he is with Computer and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan. His research interests include personal communication services networks, integrated services Internet, and wireless Internet access.

AI-CHUN PANG received her B.S.C.S.I.E. and M.S.C.S.I.E. degrees from National Chiao Tung University in 1996 and 1998, respectively. She is currently a Ph.D. candidate of the Department of Computer Science and Information Engineering, National Chiao Tung University. Her current research interests include design and analysis of personal communications services networks, mobile computing, and voice over IP.

IMRICH CHLAMTAC [F] (chlamtac@utdallas.edu) holds a Ph.D. degree in computer science from the University of Minnesota. Since 1997 he is Distinguished Chair in Telecommunications at the University of Texas at Dallas. He also holds the titles of the Sackler Professor at Tel Aviv University, Israel, Bruno Kessler Honorary Professor at the University of Trento, Italy, and university professor at the Technical University of Budapest, Hungary. He is a Fellow of the ACMr, a Fulbright Scholar, and an IEEE Distinguished Lecturer. He is the winner of the 2001 ACM award for contributions to wireless and mobile networks, and has received multiple best paper awards in wireless and optical networking. He published close to 300 papers in refereed journals and conferences, and is co-author of the first textbook on *Local Area Networks* (Lexington Books, 1981, 1982, 1984) and *Mobile and Wireless Networks Protocols and Services* (Wiley, 2000). He serves as founding Editor in Chief of *ACM/URSI/Kluwer Wireless Networks*, *ACM/Kluwer Mobile Networks and Applications*, and *SPIE/Kluwer Optical Networks Magazine*. He is also the founder and Steering Committee Chair of ACM/IEEE MobiCom and SPIE/IEEE/ACM OptiComm.