Interworking Internet Telephony and Wireless Telecommunications Networks

Jonathan Lennox
lennox@bell-labs.com

Kazutaka Murakami
kmurakami@bell-labs.com

Mehmet Karaul
karaul@bell-labs.com

Thomas F. La Porta
tlp@bell-labs.com

Bell Laboratories
Lucent Technologies
Holmdel, NJ 07733

ABSTRACT

Internet telephony and mobile telephony are both growing very rapidly. Directly interworking the two presents significant advantages over connecting them through an intermediate PSTN link. We propose three novel schemes for the most complex aspect of the interworking: call delivery from an Internet telephony (SIP) terminal to a mobile telephony (UMTS) terminal. We then evaluate the proposals both qualitatively and quantitatively. However, existing equipment may not support packet interfaces needed for such interworking. Therefore, we also consider techniques for backward compatibility, and analyze their performance as well.

1. INTRODUCTION

Two of the fastest growing areas of telecommunications are wireless mobile telephony and Internet telephony. Second and third-generation digital systems such as the Global System for Mobile communications (GSM) [3], the Universal Mobile Telecommunications System (UMTS) [4], and wideband CDMA [9] are bringing new levels of performance and capabilities to mobile communications. Meanwhile, both the Internet Engineering Task Force's Session Initiation Protocol (SIP) [6] and the International Telecommunications Union's H.323 [8] enable voice and multimedia telephone calls to be transported over an Internet Protocol (IP) network. Subscribers to each of these networks need to be able to contact subscribers on the other. There is, therefore, a need to interconnect the two networks, allowing calls to be placed between them.

Some research has been performed investigating various aspects of interworking mobile communication systems with IP-based systems. The iGSM system [14] allows an H.323 terminal to appear to the GSM network as a standard GSM terminal, so that a GSM subscriber can have his or her calls temporarily delivered to an H.323 terminal rather than a mobile device. Several papers [11, 12, 13] describe a system for interworking GSM's in-call handover procedures with H.323. However, neither of these approaches solves the general interworking question: what is the best way for calls to be delivered and routed between the two networks?

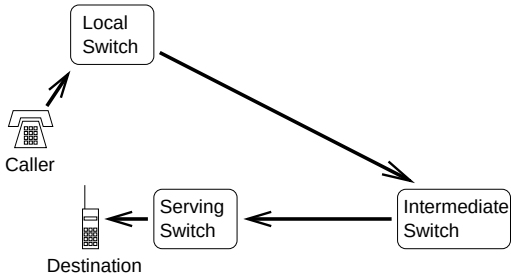

Figure 1: Illustration of triangular routing in mobile networks

As both mobile and Internet telephony are already designed to interconnect with the Public Switched Telephone Network (PSTN), the easiest way to interconnect them would be simply to use the PSTN as an intermediate link. This is, however, inefficient and suboptimal, as compared to connecting the networks by interworking the protocols directly, for a number of reasons.

First of all, routing calls via the PSTN can result in inefficient establishment of voice circuits. This is a common problem in circuit-switched wireless systems called "triangular routing," as illustrated in Figure 1. Because a caller's local switch does not have sufficient information to determine a mobile's correct current location, the signalling must travel to an intermediate switch which can locate the subscriber correctly.[1] This intermediate switch can be far away

[1]There is an architectural difference here between the American mobile system based on ANSI 41 [20] and the European systems based on GSM/UMTS MAP. In the American system, calls are always routed through a home mobile switching center, which is in a fixed location for each subscriber, so the voice traffic for all of the subscriber's calls travels through that switch. By contrast, GSM improves on this routing by sending calls through a gateway mobile switching center, which can be located close to the originating caller. However, as discussed in [1], there are some cases, such as international calls, where an originating PSTN switch does not have enough information to conclude that a call is destined for the GSM/UMTS network, and thus routes it to

from the caller and the destination even if the two are located in a geographically close area. Since voice circuits are established at the same time as the call signalling message is routed, the voice traffic could be transported over a long, inefficient route.

In Internet telephony, by contrast, the path of a call's media (its voice traffic, or other multimedia formats) is independent of the signalling path. Therefore, even if signalling takes a triangular route, the media travels directly between the devices which send and receive it. Since each device knows the other's Internet address, the packets making up this media stream are sent by the most efficient routes that the Internet routing protocols determine.

As we interwork Internet telephony with mobile telephony, we would like to maintain this advantage. We can accomplish this by supporting a direct IP connection between mobile base stations and IP terminals. With PSTN signalling, this is not possible, so IP telephony signalling must be used to establish this connection.

Another motivation for direct connection between mobile and Internet telephony is to eliminate unnecessary media transcoding. The Real-Time Transport Protocol (RTP) [18], the media transport protocol common to both H.323 and SIP, can transport almost any publicly-defined media encoding [17]. Most notably, the GSM 06.10 encoding [2] is implemented by many clients. If a GSM mobile device talks to an RTP-capable Internet telephone with an intermediate PSTN leg, the media channel would have to be converted from GSM 06.10 over the air, to uncompressed (μ-law or a-law) audio over a PSTN trunk, and then again (likely) to some compressed format over the RTP media channel. The degradation of sound quality from multiple codecs in tandem is well known, and multiple conversions induce unnecessary computation. A direct media channel between a base station and an IP endpoint allows, by contrast, communication directly using the GSM 06.10 encoding without any intermediate transcodings.

Finally, on a broader scale, an integrated architecture supporting Internet and mobile telephony will evolve naturally with the expected telecommunications architectures of the future. Third-generation wireless protocols will support wireless Internet access from mobile devices. New architectures such as RIMA [10] for Mobile Switching Centers (MSCs) are using IP-based networks for communications between MSCs and base stations. In the fixed network, meanwhile, IP telephony is increasingly becoming the long-haul transport of choice even for calls that originate in the PSTN. The direct connection between Internet telephony and mobile networks takes advantage of all these changes in architecture and allows us to build on them for the future.

In this paper, we will consider the issue of how to interwork Internet telephony and mobile telecommunications, such that all the issues discussed above are resolved. For concreteness, we will illustrate our architecture using SIP [6] for Internet telephony and UMTS [4] Release 1999 for mobile telephony. UMTS Release 1999 is an evolution of the older GSM [3] system, and as such is the most recent version of this widely deployed infrastructure. Newer UMTS releases will be directly IP-based, but systems based on GSM will likely persist for some time.

The rest of the paper is structured as follows. Section 2 gives an architectural background on the mobility and call delivery mechanisms of UMTS and SIP, to provide a basis for the following discussions. Section 3 proposes three different approaches to interworking UMTS and SIP, under the assumption that UMTS visited networks are IP-enabled. Section 4 provides mathematical and numerical analyses of the three proposals. In Section 5, we describe and analyze how efficiently the three proposals can interwork with existing non-IP-enabled infrastructure. We offer a higher-level discussion of the proposals' relative merits in Section 6, and we finish with some conclusions in Section 7.

2. BACKGROUND

In this section we review the mobility and call delivery mechanisms of UMTS and of SIP.

UMTS Mobility and Call Delivery

The key elements of a UMTS Release 1999 network are as follows. The MSC is a switching and control system in a wireless network. The MSC controlling the service area where a mobile is currently located is called its serving MSC. It routes calls to and from all the mobile devices within a certain serving area, and maintains call state for them. Associated with the serving MSC is a Visitor Location Register (VLR), a database which stores information about mobile devices in its serving area. (For the purposes of this paper we assume the predominant configuration in which the serving MSC and VLR are co-located.) Elsewhere in the fixed network we can find two other classes of entities. A Home Location Register (HLR) maintains profile information about a subscriber and keeps track of his or her current location. A gateway MSC directs calls from the PSTN into the mobile access network.

When a UMTS mobile device first powers up or enters the serving area of a new serving MSC, it transmits a unique identification code, its International Mobile Subscriber Identity (IMSI) to the MSC. From the IMSI, the serving MSC determines the mobile's HLR and informs this HLR of the mobile's current location using the UMTS Mobile Application Part (UMTS MAP) protocol. The HLR stores this information and responds with profile data for the subscriber.

When a call is placed to a mobile subscriber, the public telephone network determines from the telephone number called (the Mobile Station ISDN number, or MSISDN) that the call is destined for a mobile telephone. The call is then directed to an appropriate gateway MSC. Call delivery from the gateway MSC is performed in two phases. In the first phase, the gateway MSC obtains a temporary routing number called a Mobile Station Routing Number (MSRN) in order to route the call to the serving MSC. For this purpose, the gateway MSC first locates the subscriber's HLR based on the MSISDN and requests routing information from it using UMTS MAP. The HLR then contacts the VLR at the

the subscriber's home country. Because there is no way for circuit paths to be changed once they have been established, the call's voice traffic travels first to the user's home country and only then to his or her current location.

serving MSC. The VLR returns an MSRN that the HLR forwards to the gateway MSC. In the second phase, the gateway MSC routes the call to the serving MSC using the standard ISDN User Part (ISUP) protocol of the PSTN.

The MSRN is a temporarily assigned number which is allocated at the time the HLR contacts the VLR; it is valid only until the associated call is set up, and it is then recycled. This dynamic allocation of an MSRN is required because ISUP messages can only be directed to standard telephone numbers, and the quantity of these that can be allocated to a given serving MSC is limited. This has some costs, however, in the time needed to set up a call, as the serving MSC must be contacted twice during call setup.

When a subscriber moves from one location to another while a call is in progress, two possible scenarios result: intra-MSC or inter-MSC handovers. An intra-MSC handover occurs when a subscriber moves between the serving areas of two base stations controlled by the same serving MSC. In this case, the serving MSC simply redirects the destination of the media traffic. No signalling is necessary over the PSTN or UMTS MAP. An inter-MSC handover, on the other hand, occurs when the subscriber moves from one serving MSC's area to another. The old serving MSC contacts the new one in order to extend the call's media circuit over the PSTN. The old serving MSC then acts as an "anchor" for both signalling and voice traffic for the duration of the call.

All of the globally-significant numbers used by the UMTS system — in particular, for the purposes of this paper, the MSRN, and the identifying number of the MSCs, in addition to the MSISDN — have the form of standard E.164 [7] international telephone numbers. Therefore they can be used to route requests in Signalling System no. 7 (SS7), the telephone system's signalling transport network.

SIP Mobility and Call Delivery

Architecturally, a pure SIP network is rather simpler than a UMTS network, as it is significantly more homogeneous and much of the work takes place at the network layer, not the application layer. All devices communicate using IP, and all signalling occurs with SIP.

When a SIP subscriber becomes reachable at a new network address (either because she is using a new network device or because her device has obtained a new IP address through a mobility mechanism), the SIP device sends a SIP REGISTER to the user's registrar to inform it of the new contact location. This registration is then valid for only a limited period of time. Because end systems are assumed not to be totally reliable, registration information must be refreshed periodically (typically, once per hour) to ensure that a device has not disappeared before it could successfully de-register itself.

Unlike systems that use traditional telephone-network numbering plans, addresses in SIP are based on a "user@domain" format, similar to that of e-mail addresses. Any domain can, therefore, freely create an essentially unlimited number of addresses for itself. For the purposes of this discussion, it is useful to consider two types of addresses — "user addresses," analogous to an MSISDN number, to which external calls are placed, and "device addresses," roughly comparable to a non-transient MSRN. A device can create a temporary address for itself and have it persist for any period it wishes.

Table 1: Analogous entities in SIP and UMTS

UMTS	SIP
HLR	Registrar
Gateway MSC	Home proxy server
Serving MSC	End system (for REGISTER)
MSISDN	User address (in INVITE)
IMSI	User address (in REGISTER)
MSRN	Device address

When a SIP call is placed to a subscriber's user address, a SIP INVITE message is directed to a proxy server in the domain serving this address. The proxy server consults the recipient's registrar and obtains his or her current device address. The proxy server then forwards the INVITE message directly to the device. Because the device address is not transient, the two-stage process used by UMTS is not necessary. Once the call is established, media flows directly between the endpoints of the call, independently of the path the signalling has taken.

Though not explicitly defined as part of the basic SIP specification, in-call handover mobility is also possible within SIP. A mechanism for an environment based entirely on SIP, with mobile devices which have an Internet presence, is described in [21]. This mechanism does not use Mobile IP, as it suffers from a similar triangular routing issue as does circuit switching, and its handovers can be slow. Instead, it exploits SIP's in-call media renegotiation capabilities to alter the Internet address to which media is sent, once a device obtains a new visiting address through the standard mobile IP means. Therefore, Internet telephony calls can send their media streams to mobile devices' visiting addresses directly, rather than forcing them to be sent to the home addresses and then relayed by a home agent as in mobile IP.

There are two significant architectural differences between mobility in SIP and UMTS. First of all, a SIP network does not have an intermediate device analogous to the serving MSC. Instead, end systems contact their registrars directly, and proxy servers directly contact end systems. Second, in SIP a two-phase process is not needed to contact the device during call establishment.

Table 1 lists some analogous entities in UMTS and SIP networks.

3. ARCHITECTURE

In this section we describe our proposals for interworking SIP and UMTS networks. In our design UMTS mobile devices and their air interfaces and protocols are assumed to be unmodified. They use standard UMTS access signalling protocols and media encodings atop the standard underlying framing and radio protocols. Some UMTS entities within the fixed part of the network, however, are upgraded to have Internet presences in addition to their standard UMTS MAP and ISUP interfaces. Serving MSCs send and receive RTP

packets and SIP signalling. In some of the proposals other UMTS fixed entities, such as HLRs, have Internet presences as well. These entities still communicate with each other using UMTS MAP and other SS7 signalling protocols, however.[2]

Section 5 will discuss compatibility with existing infrastructure, in the case where serving MSCs are not IP-enabled.

There are two primary issues to consider when addressing this interworking: how calls may be placed from SIP to UMTS, and how they may be placed from UMTS to SIP. The latter point is relatively straightforward, and we will address it first. The former is more challenging and represents the main focus of this paper.

SIP/UMTS Interworking: Calls from UMTS to SIP

Calls originating from a UMTS device and directed at a SIP subscriber are not, in principle, different from calls from the PSTN to a SIP subscriber. The primary issue when placing calls from a traditional telephone network to SIP is that traditional telephones can typically only dial telephone numbers, whereas SIP addresses are of a more general form, based roughly on e-mail addresses, which cannot be dialed on a keypad. Work is ongoing to resolve this problem, but one currently envisioned solution is to use a distributed database based atop the domain name system, known as "Enum," [5] which can take an E.164 international telephone address and return a SIP universal resource locator. For example, the E.164 number +1 732 332 6063 could be resolved to the SIP URI 'sip:lennox@bell-labs.com'. A SIP subscriber wishing to be reachable from the PSTN would obtain a telephone number in a special telephone exchange controlled by a switch which understands SIP. This switch would perform this Enum lookup to obtain a SIP address, and then place the call over SIP.

Since globally significant UMTS numbers take the form of E.164 numbers, several of the proposals below use Enum-style globally distributed databases in order to locate Internet servers corresponding to these addresses. However, for such databases it would not be desirable to use the actual global Enum domain, as the semantics of the URIs returned is different.

SIP/UMTS Interworking: Mobile-Terminated Calls

The most complex point of SIP/UMTS interworking is the means by which a SIP call can be placed to a UMTS device. As discussed in the introduction, it is desirable to set up media streams directly between the calling party and the serving MSC. In order to accomplish this, SIP signalling must travel all the way to the serving MSC, as only the serving MSC will know the necessary IP address, port assignment conventions, and media characteristics.

In our model, the signalling between the serving MSC and the mobile device is unchanged from standard UMTS. This is actually a rather complicated procedure, involving communication between the serving MSC, base-station controllers, and base-stations. Devices may be in standby mode, requiring initiation of paging to locate them, or they may be turned off or in a region where no service is available, causing them to be unreachable. All these points, however, are elided in our descriptions of our architecture, as this complexity does not affect the nature of our arguments. Thus, for the purposes of discussion, this communication can be simplified into a simple pair of alert-answer messages between the serving MSC and the mobile device.

We propose three methods as to how SIP devices can determine the current MSC at which a UMTS device is registered. These have various trade-offs in terms of complexity, amount of signalling traffic, and call setup delay.

Proposal 1: modified registration

Our first proposal is to enhance a serving MSC's registration behavior. The basic idea is that a serving MSC registers not only with the subscriber's HLR, but also with a "Home SIP Registrar." This registrar maintains mobile location information for SIP calls.

The principal complexity with this technique lies in how the serving MSC locates the SIP registrar. Our proposal, illustrated in Figure 2, is to use a variant of the Enum database described above. Once the serving MSC has performed a UMTS registration for a mobile device, it knows the mobile's MSISDN number. From this information, an Enum database is consulted to determine the address of the device's home SIP registrar, and the serving MSC performs a standard SIP registration on behalf of the device.[3] A SIP call placed to the device then uses standard SIP procedures.

Because of authentication needs, this proposal uses either eight or ten UMTS MAP messages (depending on whether authentication keys are still valid at the VLR) and six DNS messages[4] per initial registration, and four SIP messages per

[2]It is possible that this SS7 signalling itself takes place over an IP network, using mechanisms such as the Stream Control Transmission Protocol (SCTP) [19].

[3]Because they travel over the public internet, SIP registrations must be authenticated. In this model, the serving MSC and the SIP proxy must have some sort of pre-existing trust relationship established. The exact mechanism for this is for future study; however, most likely some sort of public key system, with a root certificate authenticating that a MSC is a legitimate UMTS provider, would be the best approach.

[4]Only two of these six DNS messages are shown in Figure 2. In addition, four DNS messages (two request/response pairs) are necessary to resolve the destination of a SIP request. The originator of the request must first perform an SRV query on the destination, which will return an A record giving an actual hostname. The returned hostname, or the original name if no SRV record was present for the host, must then be resolved with another query, to return the actual IP network address. (Some DNS servers may optimize these queries so that a response to an SRV query also contains response information to the corresponding A query, pre-empting it, but this is not always possible.) Thus, all the message counts in this section, and in Section 5, include four DNS messages for every SIP request sent, in addition to any DNS messages used for Enum queries.
However, these DNS queries can often be cached, so the computations of signalling load in Sections 4 and 5 adjust the weight due to DNS queries by a probabilistic factor of how likely it is that the query was cached. In cases where we can be *certain* the query will be cached — as for refreshed

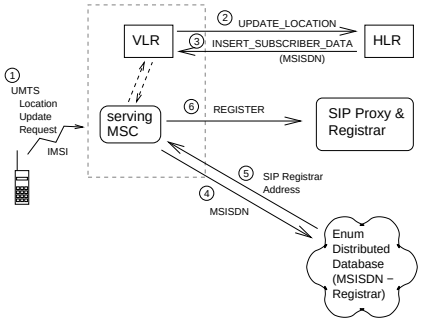

Figure 2: Registration procedure for proposal 1

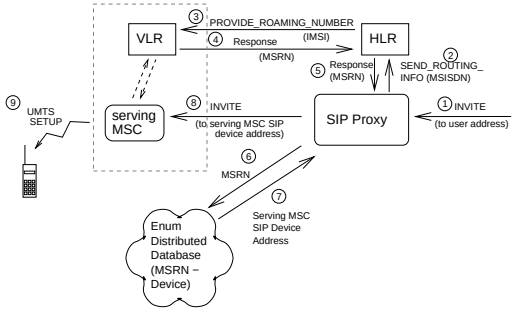

Figure 3: Call setup procedure for proposal 2

initial or refreshed registration. Call setup requires a single SIP message and four DNS messages, though some DNS queries may be cached.

Compared to our other proposals, this proposal has two primary advantages. First, the only changes to the existing infrastructure are the modifications in the serving MSC and the addition of a variant Enum database to find registrars. Neither the SIP registrar and proxy server, nor the UMTS HLR and gateway MSC, need to be altered. Second, because the complexity of the proposal occurs only in registration, call setup shares the single-lookup efficiency of SIP and is therefore relatively fast.

The disadvantages of this proposal, however, also arise due to the separation of the two registration databases. First, once a system requires the maintenance of two separate databases with rather incomparable data, the possibility arises that the information in the databases becomes inconsistent due to errors or partial system failure. This is especially true because of the differing semantics of SIP and UMTS registrations — UMTS registrations persist until explicitly removed, whereas SIP registrations have a timeout period and must be refreshed by the registering entity. Furthermore, when mobility rates are low, the dual registration procedure imposes significantly more signalling overhead than UMTS registration alone, since SIP registrations must be refreshed frequently.

Proposal 2: modified call setup

By contrast, our second proposal does not modify the UMTS registration procedure. Instead, it adds complexity to the call setup procedure. Essentially it adapts the UMTS call setup to SIP. This is illustrated in Figure 3. When a SIP call is placed to a UMTS user, the user's home SIP proxy server determines the MSISDN corresponding to the SIP user address, and queries the UMTS HLR for an MSRN. The HLR obtains this through the normal UMTS procedure of requesting it from the serving MSC's VLR. The SIP proxy server then performs an Enum lookup on this MSRN, and obtains a SIP address at the serving MSC to which the SIP INVITE message is then sent.

This approach uses either eight or ten MAP messages, as with standard UMTS, for registration, and four MAP messages, six DNS messages, and one SIP message for a call setup.

Because this proposal does not modify the UMTS registration database, it has several advantages over the previous proposal. Specifically, there is no possibility for data to become inconsistent, and the overhead of registration is as low as it is for standard UMTS. However, both the signalling load and the call setup delay are high, as call setup now involves a *triple*-phase query: a UMTS MAP query for the MSRN, an Enum lookup for the SIP device address, and finally the actual call initiation. Additionally, we have a new requirement that the SIP proxy server and the HLR need to be able to communicate with each other. This imposes additional complexity in both these devices, as it requires new protocols or interfaces.

Proposal 3: modified HLR

Our final proposal is to modify the UMTS HLR. In this proposal, the serving MSC registers the mobile at the HLR through standard UMTS means. The HLR then has the responsibility to determine the mobile's SIP device address at the serving MSC.

The overall registration procedure for this proposal is illustrated in Figure 4. When a serving MSC communicates with an HLR, the HLR is informed of the serving MSC's address, which, as mentioned earlier, is an E.164 number. The HLR performs a query to a specialized Enum database to obtain the name of the serving MSC's SIP domain, based on the serving MSC's address. While the previous two proposals treat the SIP device address as an opaque unit of information whose structure is known only to the serving MSC, this proposal takes advantage of its structure.

Figure 5 shows how a SIP call is placed. The SIP proxy server queries the HLR for a SIP address and the HLR returns an address of the form "MSISDN@hostname.of.serving.MSC" to which the SIP proxy then sends the call. This proposal uses either eight or ten MAP messages, and two DNS messages, for registration, and four DNS messages and one SIP message for call setup. Because in this proposal the HLR and the SIP proxy are assumed to be co-located, the communication between them is local and therefore can be considered as "free."

This approach has the advantage that its overhead is relatively low for registration and quite low for call setup. The

registrations — no DNS queries are listed, or included in the computations.

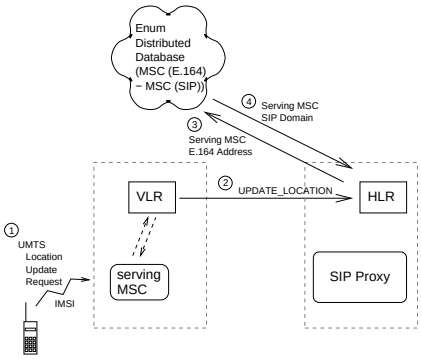

Figure 4: Registration procedure for proposal 3

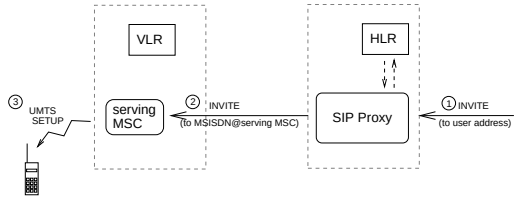

Figure 5: Call setup procedure for proposal 3

time requirements for call setup are similarly low. It does, however, require invasive modifications of HLRs. Additionally, the SIP proxy server and the HLR must be co-located, or else they must also have a protocol defined to interface them.

4. ANALYSIS

Two important criteria for evaluating the signalling performance of these three proposals for interworking SIP and UMTS are signalling load and call setup delay. A detailed study of call setup delay remains for future investigation. In this paper we focus on performance in terms of signalling load.

Each of the proposals involves the use of several different protocols, in varying ratios. In order to compare total signalling load imposed by each protocol, we assigned signalling messages of each protocol a weight. The default values of these weights are listed in Table 2. The weights represent the impact each protocol has on the total signalling load of the system. The weights were chosen to reflect the complexity of each protocol, as well as the number of nodes and geographical distance each message must cross. We discuss the effect of these weights on the total signalling load in our sensitivity analysis later in this section.

Tables 3 and 4 list the parameters for our model. We assume equal rates of call delivery r_{in} and r_{out}, as is commonly observed in European settings. We assign an exponential distribution to the probability $P_t(t)$ that a mobile remains in a particular MSC's serving area for longer than time t. DNS caching was accounted for by assigning the probabilities P_{nr}, P_{ur}, and P_{ns} to the likelihood that particular DNS queries have been performed recently, within the DNS time-to-live period.

Table 2: Message weights

Symbol	Parameter	Value
w_{sip}	Weight of a SIP message	1.0
w_{isup}	Weight of an ISUP message	1.0
w_{dns}	Weight of a DNS message	0.5
w_{map}	Weight of a MAP message	1.5

Table 3: Mobility parameters

Symbol	Parameter	Value
r_{in}, r_{out}	Rate of call delivery / origination	variable
r_{bc}	Average boundary crossing rate	variable
$P_t(t)$	Boundary crossing rate prob. distribution ($P(t_0 \geq t)$)	$e^{-r_{\text{bc}}t}$
s	Call / mobility ratio	$\frac{r_{\text{out}}+r_{\text{in}}}{r_{\text{bc}}}$
P_{nr}	Prob. that a device is new to a serving MSC	50%
P_{ur}	Prob. that a device has a unique registrar at its serving MSC	20%
P_{us}	Prob. that a device has a unique serving MSC at its HLR/registrar	20%

Table 5 shows the equations for the weighted signalling loads for registration and call establishment in each proposal. These equations are based on the packet counts for each proposal in Section 3.

Figure 6 graphs the total weighted signalling load (registration plus call setup costs) for each of the three proposals, as both the incoming call rate and the call / mobility ratio vary. The intersection line at which modified registration and modified call setup are equal is shown in bold.

From this graph, we can observe some general characteristics of the proposals' signalling load. First, the modified HLR proposal consistently has the lowest signalling load of the three, typically 20 – 30% less than the others. This corresponds to intuition, as it combines the "best" aspects of each of the other two proposals, unifying both an efficient registration and an efficient call setup procedure.

Second, the relative signalling loads for the other two proposals depend on the values of the traffic parameters. Modified call setup is more efficient for a low incoming call rate or a low call / mobility ratio (i.e., fast mobility), while modified registration is more efficient when both parameters are high. A closer look at the equations in Table 5 reveals the reasons. Consider the relative efficiency of the two approaches for varying incoming call rates: modified call setup performs

Table 4: Protocol parameters

Symbol	Parameter	Value
t_{sip}	SIP registration refresh interval	3 hr
t_{dns}	DNS cache time-to-live	24 hr
c_{auth}	Number of pieces of authentication data cached at VLR	5

Table 5: Weighted packet counts for each proposal

Case	Formula
Modified Registration	
Registration	$r_{\mathrm{bc}}\big((8+2/c_{\mathrm{auth}})\,w_{\mathrm{map}}+$ $(2P_{\mathrm{nr}}+4P_{\mathrm{ur}})\,w_{\mathrm{dns}}+$ $4\left(1+\sum_{i=1}^{\infty}P_t(it_{\mathrm{sip}})\right)w_{\mathrm{sip}}\big)$
Call setup	$r_{\mathrm{in}}\,(4P_{\mathrm{us}}w_{\mathrm{dns}}+1w_{\mathrm{sip}})$
Modified Call Setup	
Registration	$r_{\mathrm{bc}}\,(8+2/c_{\mathrm{auth}})\,w_{\mathrm{map}}$
Call setup	$r_{\mathrm{in}}\,(4w_{\mathrm{map}}+6P_{\mathrm{us}}w_{\mathrm{dns}}+1w_{\mathrm{sip}})$
Modified HLR	
Registration	$r_{\mathrm{bc}}\big((8+2/c_{\mathrm{auth}})\,w_{\mathrm{map}}$ $+2P_{\mathrm{us}}w_{\mathrm{dns}}\big)$
Call setup	$r_{\mathrm{in}}\,(4P_{\mathrm{us}}w_{\mathrm{dns}}+1w_{\mathrm{sip}})$

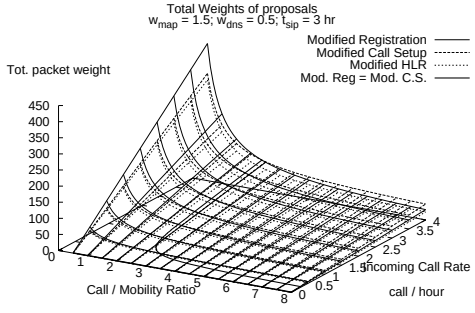

Figure 6: Weighted signalling load of the three proposals

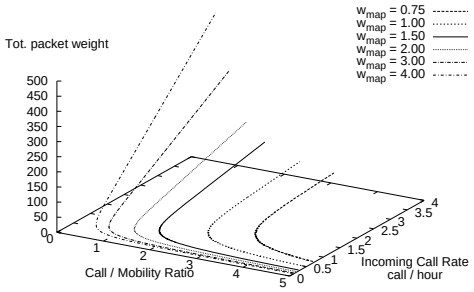

Figure 7: Line of Intersection: Mod. C.S. = Mod. Reg. (w_{map} varying)

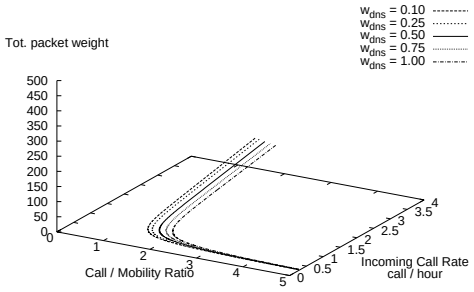

Figure 8: Line of Intersection: Mod. C.S. = Mod. Reg. (w_{dns} varying)

less well for high incoming call rates because its call setup procedure requires four additional UMTS MAP messages and possibly two additional DNS messages compared to that of modified registration. Similarly, modified call setup outperforms modified registration for low call / mobility ratios because the latter has higher registration message overhead due to dual registration and SIP registration soft-state.

In order to increase the confidence in the above results, we performed sensitivity analyses to validate our choice of various parameters.

Sensitivity analyses for the weights assigned to MAP and DNS messages are shown in Figures 7 and 8, respectively. These graphs illustrate how, as the protocol weighting changes, the position of the intersection line in Figure 6 changes.

Figure 7 shows that as the weight assigned to the MAP protocol increases, the area in which modified registration is more efficient — the right-hand side of the graph, where call rate and call/mobility ratio are both high — increases as well. This fits with the intuitive understanding of the approaches, as modified registration uses fewer MAP messages than modified call setup. Similarly, Figure 8 shows that as the weight assigned to the DNS protocol increases, the area in which modified registration is more efficient shrinks slightly. This also fits with intuition, as modified registra-

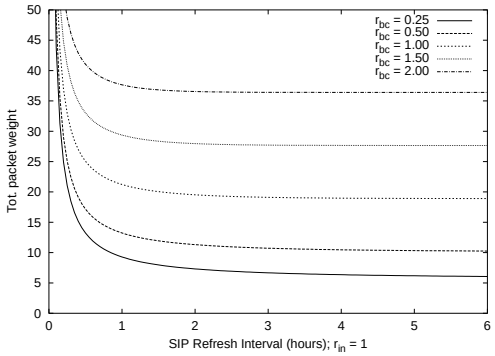

Figure 9: Total weight of modified registration

tion uses more DNS packets. However, the total packet load is generally less sensitive to the weight assigned to DNS messages, which explains why the lines in Figure 8 are relatively close to each other.

The signalling load of the modified HLR proposal is always less than the other two. Thus, it is not shown in our sensitivity graphs. In regards to the other two protocols, though the crossover point moves as the weights assigned to the protocols vary, these sensitivity analyses show that the general shape of the graph, and therefore the conclusions we draw from it, do not change.

Figure 9 shows the effect of various choices of values for the SIP registration timeout period. (This value only affects the modified registration proposal, as the other proposals do not use SIP registration.) The value for this parameter should be chosen so that the additional cost of SIP registration is relatively minor, that is, so that the graph has roughly flattened out. This optimal value therefore depends on the boundary crossing rate, but generally, a timeout of three hours is a good choice for most reasonable boundary crossing rates. This value can be larger than the standard value of one hour used by SIP, as serving MSCs can be assumed to be more reliable and available than regular SIP end systems.

5. COMPATIBILITY WITH NON-IP-ENABLED VISITED NETWORKS

As we have demonstrated, using IP for wide-area communication to a serving MSC can be much more efficient than using the circuit-switched network. However, the existing deployed circuit-switched networks cannot be ignored, and any system for connecting voice over IP networks to mobile telephony networks will have to be able to connect to networks which have not been upgraded to the new protocols.

As discussed in Section 1, both SIP and UMTS are designed to be able to interwork with the public switched telephone network. The entity which connects SIP to a circuit-switched network is called a *SIP gateway*. This gateway can terminates SIP and RTP connections from IP, and translate them into equivalent ISUP and circuit trunks on its circuit-switched side.

This same device can be used to interwork SIP and UMTS networks.[5] Conceptually, this can be viewed as decomposing the SIP-enabled serving MSC into two devices: a traditional circuit-switched serving MSC, and a SIP-enabled gateway that communicates with it. Indeed, each of the schemes described above could be implemented in this manner. However, in the general case, we must assume that the user's visited network has no support for voice over IP networks at all. In this case, we must assume that the SIP system does not have the cooperation of the VLR and SMSC for registration, and no Enum database has records for the serving network's E.164 number space.

The Telephony Routing for IP (TRIP) protocol [16, 15] is used to locate an appropriate gateway from SIP to the PSTN, based on a telephone number and on a provider's routing policy. Gateways can advertise routes to telephone numbers, with parameters indicating the "quality" of the route based on various criteria such as cost or geographic proximity. For SIP to UMTS routing, this means that we can locate a gateway close to a telephone number, minimizing the amount of triangular routing needed to reach that number. This route advertisement takes place off-line — the advertised data is stored in a local database in or near a device which needs to consume the data, and therefore these lookups are "free" in terms of the call setup message flows.

Interoperation approaches for the three proposals

Each of the three proposals for SIP-to-UMTS calls in Section 3 can support interoperation with non-IP-enabled systems in a different way. In this section we review techniques for interoperation for each of the three proposals, and review their relative signalling performance.

Non-IP-enabled visited networks with modified registration

The first proposal, modified registration, requires the serving MSC in the visited network to alter its registration procedure. The HLR and the SIP proxy server, in this case, are each unmodified.

In the interoperation case, however, we must assume the serving MSC is a standard UMTS device. In this case, therefore, the "modified registration" scenario does *not* actually involve a modified registration. Registration will simply be the standard UMTS registration procedure described in Section 2. We are left with no devices at all that have special knowledge of SIP and UMTS interworking, and so we must fall back to SIP–PSTN and PSTN–UMTS interworking.

In this scenario, when a SIP call is initiated, the SIP proxy discovers that the user is not at any SIP-enabled location. It does not know whether the user is at a non-SIP-enabled location, or is simply unreachable. To attempt to reach the user, it routes the call toward the user's MSISDN in the PSTN through an appropriate SIP gateway, and the PSTN

[5]In standard UMTS, a pure SIP/RTP—ISUP/Circuit gateway can be used. If UMTS with Route Optimization, or ANSI 41, is used instead, the gateway will also need to be able to understand some UMTS MAP or ANSI MAP messages for some supplementary services.

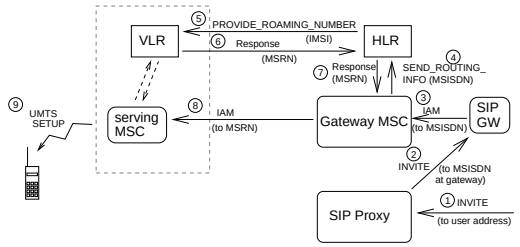

Figure 10: Call setup procedure for proposal 1 — non-IP-enabled visited network

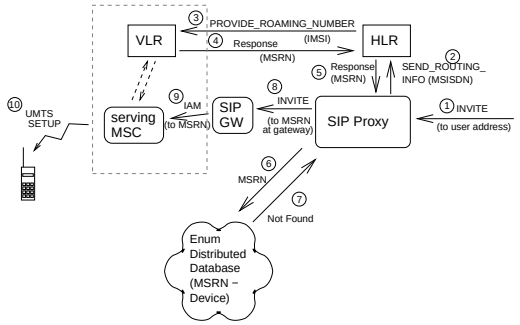

Figure 11: Call setup procedure for proposal 2 — non-IP-enabled visited network

then routes the call to a gateway MSC. The SIP gateway can either be discovered through TRIP, or pre-configured.

Thus, as shown in Figure 10, the call setup procedure for this procedure consists of a SIP INVITE message for the MSISDN at a SIP gateway, followed by the standard UMTS call setup procedure. Because the call must be directed to the MSISDN via the PSTN, connections to non-IP-enabled visited networks, under this proposal, do not avoid triangular routing.

In the non-IP-enabled visited network case, this proposal uses the standard eight or ten UMTS MAP messages for registration. Call setup requires one SIP message, two ISUP messages, and four MAP messages. We can assume that the SIP proxy has only a small number of SIP gateways which it wants to use to reach gateway MSCs, and therefore the DNS lookup for the SIP gateway can be amortized widely over all the users and therefore be ignored.

Non-IP-enabled visited networks with modified call setup

In the modified call setup proposal, the SIP Proxy discovers that a serving MSC does not support SIP. As shown in figure 11, this occurs at call setup time, when the Enum MSRN mapping database does not return a mapping from the MSRN to a SIP address.

In this case, the SIP proxy knows the MSRN to use to reach the user. Using TRIP, the proxy can thus locate a SIP gateway close to the serving MSC. Assuming that such a gateway is available, therefore, this proposal therefore largely eliminates triangular routing even when visited networks do not support IP.

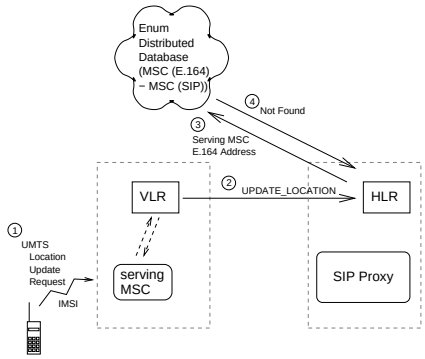

Figure 12: Registration procedure for proposal 3 — non-IP-enabled visited network

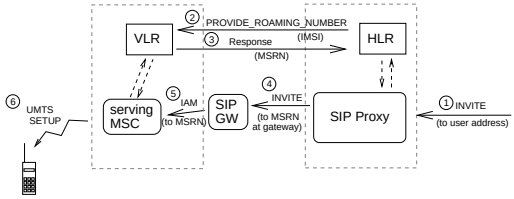

Figure 13: Call setup procedure for proposal 3 — non-IP-enabled visited network

However, interoperation with non-IP-enabled visited networks makes this scenario's primary disadvantage, slow call setup, even worse. In this case, the lookup may potentially require *four* round trips between the originating and serving systems — the MSRN lookup; the failing Enum lookup; potentially, the DNS lookup of the SIP gateway; and finally the SIP INVITE message to the SIP gateway. If we assume the SIP gateway is close to the serving MSC, however, the ISUP message sent from the SIP gateway to the serving MSC does not require another round trip.

This proposal uses the standard eight or ten UMTS MAP messages for registration. Call setup involves four MAP messages, six DNS messages, one SIP message, and one ISUP message.

Non-IP-enabled visited networks with modified HLR

Finally, the proposal to modify the UMTS HLR is different from the other two proposals in that it can detect non-IP-enabled visited networks at registration time. As shown in Figure 12, when the modified HLR attempts to determine the serving MSC's SIP domain based on its E.164 address, it discovers that there is no such domain available. It therefore knows that calls for this user must be handled in a circuit-compatible manner.

Figure 13 shows the resulting call setup procedure. Because the call must reach the serving MSC through UMTS means, the HLR must initiate the standard MSRN lookup procedure. Once a MSRN has been assigned, a SIP gateway can be located for it, using TRIP. (This TRIP lookup can be done either by the HLR or by the SIP Proxy.) The call is then placed through the SIP proxy to the serving MSC.

Registration in this proposal requires eight or ten MAP

Table 6: Weighted packet counts for each proposal: non-IP-enabled visited network

Case	Formula
Modified Registration	
Registration	$r_{\mathrm{bc}}\left(8 + 2/c_{\mathrm{auth}}\right) w_{\mathrm{map}}$
Call setup	$r_{\mathrm{in}}\left(4w_{\mathrm{map}} + 1w_{\mathrm{sip}} + 2w_{\mathrm{isup}}\right)$
Modified Call Setup	
Registration	$r_{\mathrm{bc}}\left(8 + 2/c_{\mathrm{auth}}\right) w_{\mathrm{map}}$
Call setup	$r_{\mathrm{in}}\left(4w_{\mathrm{map}} + 6P_{\mathrm{us}}w_{\mathrm{dns}} + 1w_{\mathrm{sip}} + 1w_{\mathrm{isup}}\right)$
Modified HLR	
Registration	$r_{\mathrm{bc}}\big(\left(8 + 2/c_{\mathrm{auth}}\right) w_{\mathrm{map}} + 2P_{\mathrm{us}}w_{\mathrm{dns}}\big)$
Call setup	$r_{\mathrm{in}}\left(2w_{\mathrm{map}} + 4P_{\mathrm{us}}w_{\mathrm{dns}} + 1w_{\mathrm{sip}} + 1w_{\mathrm{isup}}\right)$

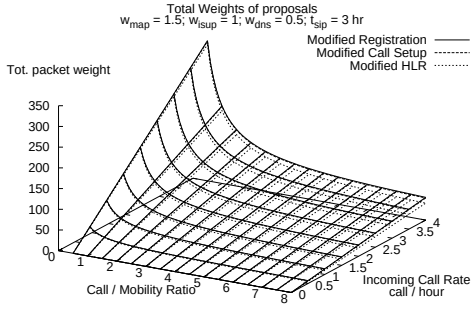

Figure 14: Weighted signalling load of the three proposals: non-IP-enabled visited network

messages and two DNS messages. Call setup requires two MAP messages, four DNS messages, one SIP message, and one ISUP message. As in the case when serving MSCs are IP-enabled, communication between the SIP proxy and the HLR can be considered to be "free."

Because this proposal discovers early on, at registration time, that visited networks do not support IP, in this environment this proposal is better than the other two both for the call setup delay and for the total message load. Additionally, as with the second scenario but in contrast to the first, triangular routing is still largely avoided. Because of the need for MSRN lookup, however, call setup for non-IP-enabled visited networks is still significantly heavier-weight than it is with IP-enabled networks.

Analysis of non-IP-enabled scenarios

In Section 4, we analyzed the performance of the three proposals in the ordinary cases, by assigning weights to every message (Table 2) and considering the total signalling load each protocol imposes on the network under a range of possible user behaviors (Table 4).

The behavior of the non-IP-enabled scenarios for the three protocols can be analyzed similarly. Table 6 shows the equations for the weighted signalling load for the three proposals in this case.

Figure 14 graphs Table 6 given the same assumptions as

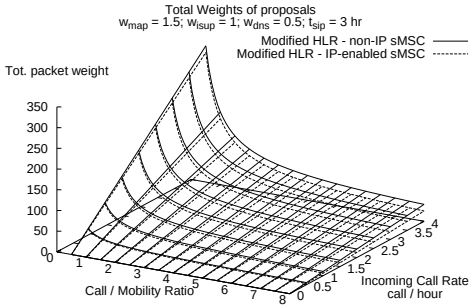

Figure 15: Comparison of modified HLR signalling load with and without IP-enabled visited network.

used in Figure 6. The graph shows that when the visited network is not IP-enabled, the signalling load of the modified registration and modified call setup procedures are nearly equal. Indeed, analysis of the equations quickly shows that in this scenario the load of modified registration exceeds that of modified call setup by only $r_{\mathrm{bc}}\left(w_{\mathrm{isup}} - 6P_{\mathrm{us}}w_{\mathrm{dns}}\right)$, or $0.4r_{\mathrm{bc}}$ given the parameter values used for the graph. (Because this is a constant factor, the weights of modified registration and modified call setup never cross in this graph, so no line of intersection is shown in Figure 14.)

The modified HLR procedure is consistently better than the other two proposals in this environment as well. The amount by which modified HLR outperforms the other proposals depends strongly on the degree to which call setup dominates the weight, since the three proposals have very similar registration procedures in these scenarios. The signalling load of modified HLR is lower by a factor of only 2% when the call-mobility ratio is very low (0.5), but is 20% lower with a moderate call-mobility ratio (4.0) and 30% lower with a high call-mobility ration (8.0).

Figure 15 compares the weights of the modified HLR proposal with and without an IP-enabled visited network. We can see that the IP-enabled case is significantly more efficient than the non-IP-enabled case. As would be expected, since the registration procedure uses the same number of messages in both cases, the relative benefit of the IP-enabled case depends on how much the message flow is dominated by call setup. The load advantage of the IP-enabled case varies, from approximately 5% when the call-mobility ratio is very low (0.5), through 36% for a moderate ratio (4.0), to approximately 65% when the ratio is high (8.0). The relative loads of the other two proposals are not shown, but are generally similar.

The comparative merits of the three proposals in the case of a non-IP-enabled visited network are therefore relatively similar to what they are in the case of the IP-enabled visited network described in Sections 3 and 4. Modified registration and modified call setup are roughly similar, and their relative merits depend on the exact assumptions made about packet weights and network characteristics. The modified HLR case is significantly better, though again it requires fairly invasive modifications of HLRs.

6. DISCUSSION

The three proposed schemes to interconnect UMTS mobile and SIP Internet telephony impose different signalling burdens on the network. The modified HLR scheme always imposes the least signalling burden, typically $20 - 30\%$ less than the other schemes. The efficiency of the other two proposals, modified registration and modified call setup, depends on the traffic parameters. When the incoming call rate and call / mobility ratio are both high, modified registration is more efficient. Modified call setup performs better otherwise.

In the case when we must interoperate with visited networks that do not support IP, the total signalling burden is higher, by about 36% in an average case. The modified HLR scheme is still the most efficient in this scenario, with typically 20% less load than the other two proposals. The modified call setup and modified registration schemes result in nearly identical signalling load.

The modified HLR case therefore appears to be the most efficient of the three proposed scenarios that we have studied. However, it requires significantly greater modification to UMTS equipment. The other two proposals are roughly similar in efficiency. There relative merits depend on the environment in which they would be deployed.

Further work

Our work addresses the issue of how calls can be set up to a SIP-enabled serving MSC. Full support of SIP-UMTS interconnection will require another issue to be resolved: interworking in-call handovers, in which a terminal moves during a call.

As explained in Section 2, there are two categories of in-call handover: intra-MSC and inter-MSC. Intra-MSC handover does not need to be treated specially for SIP-UMTS interworking. Because this happens between the serving MSC and the base stations, the network beyond the serving MSC is not affected. As an optimization, however, a serving MSC could use different IP addresses corresponding to different base stations under its control. In this case, a mechanism for SIP mobility as described before could be used to change the media endpoint address in mid-call.

Inter-MSC handover does affect SIP-UMTS interworking, and this issue remains for future study. We anticipate that a mechanism similar to that of [13], as described in the introduction, could be adapted to SIP for this purpose.

7. CONCLUSION

We proposed three novel schemes to directly interconnect UMTS mobile and SIP Internet telephony systems. Compared with the conventional approach of routing a call through the PSTN, direct interconnection prevents triangular routing and eliminates unnecessary transcodings along its path. We analyzed the signalling message load of three proposals under a wide range of call and mobility conditions. The modified HLR scheme always imposes less signalling burden, although it requires significantly greater modification to UMTS equipment. The efficiency of the other two proposals, modified registration and modified call setup, depends on the traffic parameters. In the case when we must interoperate with visited networks that do not support IP, the total signalling burden is higher. The modified HLR scheme is still the most efficient in this scenario. We therefore conclude that the modified HLR scheme is the best of the three proposals.

8. REFERENCES

[1] Y.-J. Cho, Y.-B. Lin, and H. C.-H. Rao. Reducing the network cost of call delivery to GSM roamers. *IEEE Network*, 11(5):19–25, Sept. 1997.

[2] European Telecommunications Standards Institute. Digital cellular telecommunications system, full rate speech. GSM 06.10 version 5.0.1, European Telecommunications Standards Institute, Sophia Antipolis, France, May 1997.

[3] European Telecommunications Standards Institute. Digital cellular telecommunications system, network architecture. GSM 03.02 version 7.1.0 release 1998, European Telecommunications Standards Institute, Sophia Antipolis, France, Feb. 2000.

[4] European Telecommunications Standards Institute. Universal mobile telecommunications system (UMTS), general UMTS architecture. 3G TS 23.101 version 3.0.1 release 1999, European Telecommunications Standards Institute, Sophia Antipolis, France, Jan. 2000.

[5] P. Faltstrom. E.164 number and DNS. Request for Comments 2916, Internet Engineering Task Force, Sept. 2000.

[6] M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg. SIP: session initiation protocol. Request for Comments 2543, Internet Engineering Task Force, Mar. 1999.

[7] International Telecommunication Union. The international public telecommunication numbering plan. Recommendation E.164, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, May 1997.

[8] International Telecommunication Union. Packet based multimedia communication systems. Recommendation H.323, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, Feb. 1998.

[9] B. Jabbari, Ed. Special issue on wideband CDMA. *IEEE Communications Magazine*, 36(9), Sept. 1998.

[10] T. F. La Porta, K. Murakami, and R. Ramjee. RIMA: router for integrated mobile access. In *Proceedings of the 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communication (PIMRC)*, London, United Kingdom, Sept. 2000.

[11] W. Liao. Mobile internet telephony: Mobile extensions to H.323. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, New York, Mar. 1999.

[12] W. Liao. Mobile internet telephony protocol: An application layer protocol for mobile internet telephony services. In *Conference Record of the International Conference on Communications (ICC)*, Vancouver, British Columbia, June 1999.

[13] W. Liao and J.-C. Liu. VoIP mobility in IP/cellular network internetworking. *IEEE Communications Magazine*, 38(4):70–75, Apr. 2000.

[14] H. C. H. Rao, Y.-B. Lin, and S.-L. Cho. iGSM: VoIP service for mobile networks. *IEEE Communications Magazine*, 38(4):62–69, Apr. 2000.

[15] J. Rosenberg, H. Salama, and M. Squire. Telephony routing over IP (TRIP). Internet Draft, Internet Engineering Task Force, Aug. 2001. Work in progress.

[16] J. Rosenberg and H. Schulzrinne. A framework for telephony routing over IP. Request for Comments 2871, Internet Engineering Task Force, June 2000.

[17] H. Schulzrinne. RTP profile for audio and video conferences with minimal control. Request for Comments 1890, Internet Engineering Task Force, Jan. 1996.

[18] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: a transport protocol for real-time applications. Request for Comments 1889, Internet Engineering Task Force, Jan. 1996.

[19] R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, and V. Paxson. Stream control transmission protocol. Request for Comments 2960, Internet Engineering Task Force, Oct. 2000.

[20] Telecommunications Industry Association and Electronics Industry Association. Cellular radiotelecommunications intersystem operations. TIA/EIA ANSI-41-D, Telecommunications Industry Association, Arlington, Virginia, Dec. 1997.

[21] E. Wedlund and H. Schulzrinne. Mobility support using SIP. In *Second ACM/IEEE International Conference on Wireless and Mobile Multimedia (WoWMoM'99)*, Seattle, Washington, Aug. 1999.