

Overview of VoIP Systems

In their simplest form, Voice over IP protocols simply enable two (or more) devices to transmit and receive real-time audio traffic that allows their respective users to communicate. In general, VoIP architectures are partitioned in two main components: signaling and media transfer. Signaling covers both abstract notions, such as endpoint naming and addressing, and concrete protocol functions such as parameter negotiation, access control, billing, proxying, and NAT traversal. Depending on the architecture, quality of service (QoS) and device configuration/management may also be part of the signaling protocol (or protocol family). The media transfer aspect of VoIP systems generally includes a comparatively simpler protocol for encapsulating data, with support for multiple codecs and (often, but not always) content security. A commonly used media transfer protocol is RTP [219]. There exists an RTP profile (named Secure RTP, or SRTP [131]) that supports encryption and integrity protection, but it is not yet widely used. The RTP protocol family also includes RTCP, which is used to control certain RTP parameters between communicating endpoints.

However, a variety of other features are generally also desired by users and offered by providers as a means for differentiation by competing technologies and services, such as video, integration with calendaring and file sharing, and bridging to other networks (*e.g.*, to the “regular” telephony network). Furthermore, a number of different decisions may be made when designing a VoIP system, reflecting different requirements and approaches to addressing, billing, mobility, security and access control, usability, and other issues. Consequently, there exist a variety of different VoIP protocols and architectures. For concreteness, we will focus our attention on a popular and widely deployed technology: the Session Initiation Protocol (SIP) [212]. We will also discuss the Unlicensed Mobile Access (UMA) architecture [1], as a different approach to VoIP that is gaining traction among wireless telephony operators. In the rest of this chapter, we give a high-level overview of SIP and UMA, followed by a brief description of the salient points of a few other popular VoIP systems, such as H.323 and Skype. We will refer back to this overview when discussing the threat space and specific vulnerabilities in Sec. 3.

2.1 Session Initiation Protocol

SIP is a protocol standardized by the Internet Engineering Task Force (IETF), and is designed to support the setup of bidirectional communication sessions including, but not limited to, VoIP calls. It is similar in some ways to HTTP, in that it is text-based, has a request-response structure, and even uses a mechanism based on the HTTP Digest Authentication [88] for user authentication. However, it is an inherently stateful protocol that supports interaction with multiple network components (e.g., middleboxes such as PSTN bridges), and asynchronous notifications. While its finite state machine is seemingly simple, in practice it has become quite large and complicated — an observation supported by the fact that the main SIP RFC [212] is one of the longest ever defined (after the encyclopedic “Internet Security Glossary” RFC 4949), with additional RFCs further extending the specification. [Figure 1](#) shows the number of SIP-related RFCs (and the number of total bytes in these) per year (until May 2009), and a size comparison of the main SIP RFC with respect to the TCP RFC, the 5 main MIME RFCs, the 2 Secure MIME (S/MIME) RFCs, and the 4 main IPsec RFCs. These graphs should provide a quantitative, if indirect, indication of the complexity of SIP.

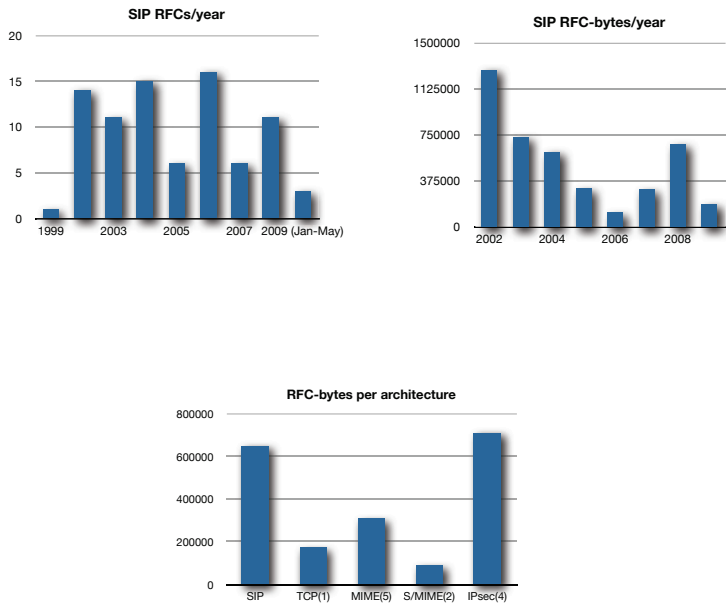


Fig. 1 Breakdown of SIP-related RFCs and their sizes

SIP can operate over a number of transport protocols, including TCP [190], UDP [189] and SCTP [179]. UDP is generally the preferred method due to simplicity

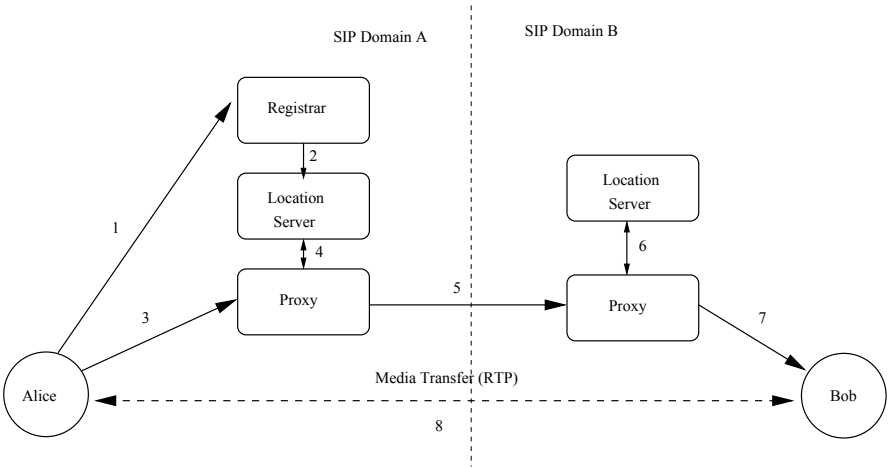


Fig. 2 Session Initiation Protocol (SIP) entity interactions. User Alice registers with her domain’s Registrar (1), which stores the information in the Location Server (2). When placing a call, Alice contacts her local Proxy Server (3), which may consult the Location Server (4). A call may be forwarded to another Proxy Server (5), which will consult its domain Location Server (6) before forwarding the call to the final recipient. After the SIP negotiation terminates, RTP is used directly between Alice and Bob to transfer media content. For simplicity, this diagram does not show the possible interaction between Alice and a Redirection Server (which would, in turn, interact with the Location Server).

and performance, although TCP has the advantage of supporting TLS protection of call setup. However, recent work on Datagram TLS (DTLS) [205] may render this irrelevant. SCTP, on the other hand, offers several advantages over both TCP and UDP, including DoS resistance [114], multi-homing and mobility support, and logical connection multiplexing over a single channel.

In the SIP architecture, the main entities are end points (whether softphones or physical devices), a proxy server, a registrar, a redirect server, and a location server. [Figure 2](#) shows a high-level view of the SIP entity interactions. The registrar, proxy and redirect servers may be combined, or they may be separate entities operated independently. Endpoints communicate with a registrar to indicate their presence. This information is stored in the location server. A user may be registered via multiple endpoints simultaneously.

During call setup, the endpoint communicates with the proxy which uses the location server to determine where the call should be routed to. This may be another endpoint in the same network (e.g., within the same enterprise), or another proxy server in another network. Alternatively, endpoints may use a redirect server to directly determine where a call should be directed to; redirect servers consult with the location server in the same way that proxy servers operate during call setup. Once an end-to-end channel has been established (through one or more proxies) between the two endpoints, SIP negotiates the actual session parameters (such as the codecs, RTP ports, etc.) using the Session Description Protocol (SDP) [113].

Figure 3 shows the message exchanges during a two-party call setup. Alice sends an INVITE message to the proxy server, optionally containing session parameter information encoded within SDP. The proxy forwards this message directly to Bob, if Alice and Bob are users of the same domain. If Bob is registered in a different domain, the message will be relayed to Bob’s proxy, and from there to Bob. Note that the message may be forwarded to multiple endpoints, if bob is registered from multiple locations. While these are ringing (or otherwise indicating that a call setup is being requested), RINGING messages are sent back to Alice. Once the call has been accepted, an OK message is sent to Alice, containing his preferred parameters encoded within SDP. Alice responds with an ACK message. Alice’s session parameter preferences may be encoded in the INVITE or the ACK message.

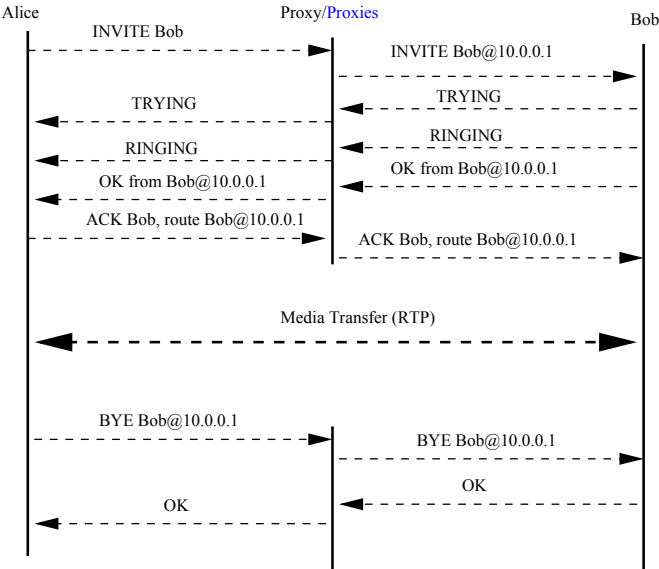


Fig. 3 Message exchanges during a SIP-based two-party call setup.

Following this exchange, the two endpoints can begin transmitting voice, video or other content (as negotiated) using the agreed-upon media transport protocol, typically RTP. While the signaling traffic may be relayed through a number of SIP proxies, the media traffic is exchanged directly between the two endpoints. When bridging different networks, e.g., PSTN and SIP, media gateways may disrupt the end-to-end nature of the media transfer. These entities translate content (e.g., audio) between the formats that are supported by the different networks.

Because signaling and media transfer operate independent of each other, the endpoints are responsible for indicating to the proxies that the call has been terminated, using a BYE message which is relayed through the proxies along the same path as the call setup messages.

There are many other protocol interactions supported by SIP, that cover many common (and uncommon) scenarios including call forwarding (manual or automatic), conference calling, voicemail, etc. Typically, this is done by semantically overloading SIP messages such that they can play various roles in different parts of the call. We shall see in Sec. 3 examples of how this flexibility and protocol modularity can be used to attack the system. It is worth pointing out that many of the vulnerabilities we will discuss in Sec. 3 are at least partially caused by this complexity. Some efforts to formally define and analyze parts of the protocol have pointed out subtle problems [285], but such efforts have not (yet) been extended to cover significant portions of the specifications due to their size and complexity.

All SIP traffic is typically transmitted over port 5060 (UDP or TCP), although that is configurable. The ports used for the media traffic, however, are dynamic and negotiated via SDP during call setup. This poses some problems when Network Address Translation (NAT) or firewalls are traversed. Typically, these have to be stateful and understand the SIP exchanges so that they can open the appropriate RTP ports for the media transfer. In the case of NAT traversal, endpoints may use protocols like STUN to enable communication. Alternatively, the Universal Plug-and-Play (uPnP) protocol ¹ may be used in some environments, such as residential broadband networks consisting of a single subnet behind a NAT gateway.

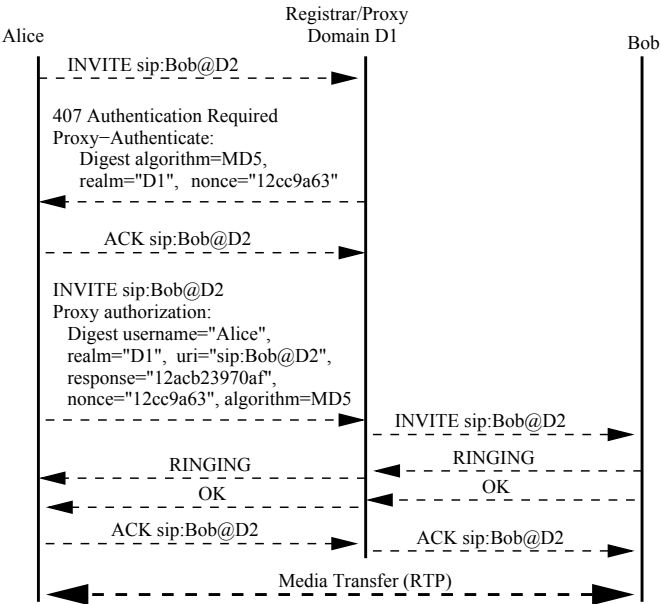


Fig. 4 SIP Digest Authentication

¹ <http://www.upnp.org/>

For authenticating endpoints, the registrar and the proxy typically use HTTP Digest Authentication, as shown in Fig. 4. This is a simple challenge-response protocol that uses a shared secret key along with a username, domain name, a nonce, and specific fields from the SIP message to compute a cryptographic hash. Using this mechanism, passwords are not transmitted in plaintext form over the network. It is worth noting that authentication may be requested at almost any point during a call setup. We shall later see an example where this can be abused by a malicious party to conduct toll fraud in some environments.

For more complex authentication scenarios, SIP can use S/MIME encapsulation [196] to carry complex payloads, including public keys and certificates. When TCP is used as the transport protocol for SIP, TLS can be used to protect the SIP messages. TLS is required for communication among proxies, registrars and redirect servers, but only recommended between endpoints and proxies or registrars. Alternatively, IPsec [135] may be used to protect all communications, regardless of the transport protocol. However, because few implementations integrate SIP, RTP and IPsec, it is left to system administrators to figure out how to setup and manage such configurations.

2.2 Unlicensed Mobile Access

UMA is a 3GPP standard for enabling transparent access to mobile circuit-switched voice networks, packet-switch data networks and IMS services using any IP-based substrate. Handsets supporting UMA can roam between the operator's wireless network (usually referred to as a Radio Access Network, or RAN) and the Internet without losing access. For example, a call that is initiated over the RAN can then be routed, without being dropped and with no user intervention, over the public Internet if conditions are more favorable (e.g., stronger WiFi signal in the user's premises, or in a hotel wireless hotspot while traveling abroad). For consumers, UMA offers better connectivity and the possibility of lower cost by enabling new business models and reducing roaming charges (under some scenarios). For operators, UMA reduces the need for additional spectrum, cellphone towers and related equipment. A variety of cellphones supporting UMA over WiFi currently exist, along with home gateways and USB-stick softphones. More recently, some operators have introduced femto-cells (ultra-low power RAN cells intended for consumer-directed deployment) that can act as UMA gateways, allowing any mobile handset to take advantage of UMA where such devices are deployed.

The basic approach behind UMA is to encapsulate complete GSM and 3G radio frames (except for the over-the-air crypto) inside IP packets. These can then be transmitted over any IP network, including the Internet. This means that the mobile operator can continue to use the existing back-end equipment; all that is needed is a gateway that decapsulates the GSM/3G frames and injects them to the existing circuit-switched network (for voice calls), as can be seen in Fig. 5.

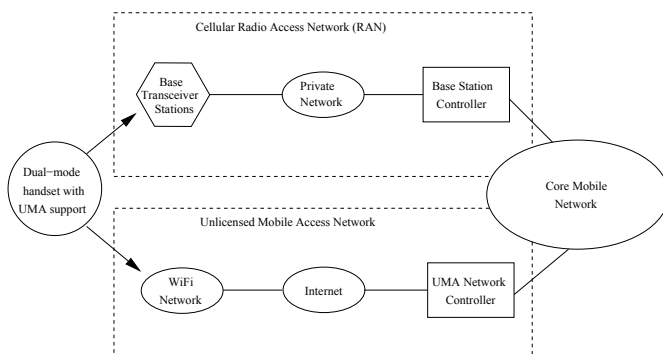


Fig. 5 Unlicensed Mobile Access (UMA) conceptual architecture

To protect both signaling and media traffic confidentiality and integrity while traversing untrusted (and untrustworthy) networks, UMA uses IPsec. All traffic between the handset (or, more generally, UMA endpoint) and the provider's UMA Network Controller (or a firewall/VPN concentrator screening traffic) is encrypted and integrity-protected using ESP [134]. The use of IPsec provides a high level of security for the traffic, once keys and other parameters have been negotiated. For that purpose, the IKEv2 key management protocol [133] is used. Authentication uses the EAP-SIM [120] (for GSM handsets) and EAP-AKA [18] (for UMTS handsets) profiles. Authentication is asymmetric: the provider authenticates to the handset using digital signatures and public key certificates, while the handset authenticates using a SIM-embedded secret key. It is worth pointing out that UMA provides stronger authentication guarantees than the baseline cellphone network, in that the provider does not authenticate to the handset in a RAN. Furthermore, the cryptographic algorithms used in IPsec (AES and 3DES) are considered significantly stronger than the on-the-air algorithms used in GSM.

Despite the use of strong cryptography and sound protocols, UMA introduces some new risks in the operator networks, since these now have to be connected to the public Internet in a much more intimate fashion. In particular, the security gateway must process IPsec traffic, including the relatively complex IKEv2 protocol, and a number of UMA-related discovery and configuration protocols. These increase the attack surface and overall security exposure of the operators significantly.

2.3 Other VoIP Systems

H.323 is an ITU-defined protocol family for VoIP (audio and video) over packet-switched data networks. The various subprotocols are encoded in ASN.1 format. In the H.323 world, the main entities are terminals (software or physical phones), a gateway, a gatekeeper and a back-end service. The gatekeeper is responsible for ad-

dress resolution, controlling bandwidth use and other management functions, while the gateway connects the H.323 network with other networks (*e.g.*, PSTN, or a SIP network). The back-end service maintains data about the terminals, including configuration, access and billing rights, etc. An optional multipoint control unit may also exist to enable multipoint communications, such as a teleconference. To setup a H.323 call, terminals first interact with the gatekeeper using the H.225 protocol over either TCP or UDP to receive authorization and perform address resolution. Using the same protocol, they then establish the end-to-end connection to the remote terminal (possibly through one or more gateways). At that point, H.245 over TCP is used to negotiate the parameters for the actual media transfer, including ports, which uses RTP (as in the case of SIP). A number of other protocols within the H.323 framework covering security, interoperability with PSTN, teleconferencing, and others. Authentication may be requested at several steps during call setup, and typically depends on symmetric keys but may also use digital signatures. Voice encryption is also supported through SRTP and MIKEY [19]. Unlike SIP, H.323 does not use a well-known port, making firewall traversal even more complicated.

Skype² is a peer-to-peer VoIP system that was originally available as a softphone for desktop computers but has since been integrated into cellphones and other handheld devices, either as an add-on or as the exclusive communication mechanism. It offers voice, video, and text messaging to all other Skype users free of charge, and provides bridging (typically for a fee) to the PSTN both for outgoing and incoming calls and text messages (SMS). The underlying protocol is proprietary, and the software itself incorporates several anti-reverse engineering techniques. Nonetheless, some analysis [26, 32] and reverse engineering [38] have taken place, indicating both the ubiquitous use of strong cryptography and the presence of some software bugs (at the time of the work). The system uses a centralized login server but is otherwise fully distributed with respect to intra-Skype communications.

A number of chat (IM) networks, such as the AOL Instant Messenger, Microsoft's Live Messenger, Yahoo! Messenger, and Google Talk offer voice and video capabilities as well. Although each network uses its own (often proprietary) protocol, there exist bridges between most of them, allowing inter-IM communication at the text level. In most of these networks, users can place outgoing voice calls to the PSTN. Some popular IM clients also integrate SIP support.

² <http://www.skype.com/>



<http://www.springer.com/978-1-4419-9865-1>

Voice over IP Security

A Comprehensive Survey of Vulnerabilities and
Academic Research

Keromytis, A.D.

2011, XIII, 83 p., Softcover

ISBN: 978-1-4419-9865-1