TECHNICAL REPORT

ISO/IEC TR 16166

First edition
2010-08-01

Information technology — Telecommunications and information exchange between systems — Next Generation Corporate Networks (NGCN) — Security of session-based communications

Technologies de l'information — Téléinformatique — Réseaux d'entreprise de prochaine génération (NGCN) — Sécurité des communications sur la base de sessions

Reference number
ISO/IEC TR 16166:2010(E)

ISO copyright office
Case postale 56 CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC TR 16166 was prepared by Ecma International (as ECMA TR/100) and was adopted, under a special “fast-track procedure”, by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, in parallel with its approval by national bodies of ISO and IEC.

Introduction

This Technical Report is one of a series of Ecma publications that explore IP-based enterprise communication involving Corporate telecommunication Networks (CNs) (also known as enterprise networks) and in particular Next Generation Corporate Networks (NGCN). The series particularly focuses on inter-domain communication, including communication between parts of the same enterprise, between enterprises and between enterprises and carriers. This particular Technical Report discusses issues related to the security of session-based communications and builds upon concepts introduced in ISO/IEC TR 12860.

This Technical Report is based upon the practical experience of Ecma member companies and the results of their active and continuous participation in the work of ISO/IEC JTC1, ITU-T, ETSI, IETF and other international and national standardization bodies. It represents a pragmatic and widely based consensus. In particular, Ecma acknowledges valuable input from experts in ETSI TISPAN.

Information technology — Telecommunications and information exchange between systems — Next Generation Corporate Networks (NGCN) — Security of session-based communications

1 Scope

This Technical Report is one of a series of publications that provides an overview of IP-based enterprise communication involving Corporate telecommunication Networks (CNs) (also known as enterprise networks) and in particular Next Generation Corporate Networks (NGCN). The series particularly focuses on session level communication based on the Session Initiation Protocol (SIP) [4], with an emphasis on inter-domain communication. This includes communication between parts of the same enterprise (on dedicated infrastructures and/or hosted), between enterprises and between enterprises and public networks. Particular consideration is given to Next Generation Networks (NGN) as public networks and as providers of hosted enterprise capabilities. Key technical issues are investigated, current standardisation work and gaps in this area are identified, and a number of requirements and recommendations are stated. Among other uses, this series of publications can act as a reference for other standardisation bodies working in this field, including ETSI TISPAN, 3GPP, IETF and ITU-T.

This particular Technical Report discusses security of session-based communications. It uses terminology and concepts developed in ISO/IEC TR 12860 [1]. It identifies a number of requirements impacting NGN standardisation and makes a number of recommendations concerning deployment of enterprise networks. Also a number of standardisation gaps are identified. Both signalling security and media security are considered.

The scope of this Technical Report is limited to communications with a real-time element, including but not limited to voice, video, real-time text, instant messaging and combinations of these (multi-media). The non-real-time streaming of media is not considered. For media, only security of transport (e.g., securing the Real-time Transport Protocol, RTP [6]) is considered, and higher level security measures (e.g., digital rights management) are not considered. Peer-to-peer signalling between SIP user agents (without involving SIP intermediaries) is not considered.

Detailed considerations for lawful interception are outside the scope of this Technical Report, although general considerations for call recording and audit are discussed.

2 References

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

[1] ISO/IEC TR 12860, Information technology — Telecommunications and information exchange between systems — Next Generation Corporate Networks (NGCN) — General

[2] ISO/IEC TR 12861, Information technology — Telecommunications and information exchange between systems — Next Generation Corporate Networks (NGCN) — Identification and routing

[3] ISO/IEC TR 16167, Information technology — Telecommunications and information exchange between systems — Next Generation Corporate Networks (NGCN) — Emergency calls

[4] IETF RFC 3261, SIP: Session Initiation Protocol

[5] IETF RFC 3325, Private Extensions to the Session Initiation Protocol (SIP) for Asserted Identity within Trusted Networks

[6] IETF RFC 3550, RTP: A Transport Protocol for Real-Time Applications

[7] IETF RFC 3711, The Secure Real-time Transport Protocol (SRTP)

[8] IETF RFC 3830, MIKEY: Multimedia Internet KEYing

[9] IETF RFC 3893, Session Initiation Protocol (SIP) Authenticated Identity Body (AIB) Format

[10] IETF RFC 4119, A Presence-based GEOPRIV Location Object Format

[11] IETF RFC 4301, Security Architecture for the Internet Protocol

[12] IETF RFC 4346, The Transport Layer Security (TLS) Protocol Version 1.1

[13] IETF RFC 4347, Datagram Transport Layer Security

[14] IETF RFC 4474, Enhancements for Authenticated Identity Management in the Session Initiation Protocol (SIP)

[15] IETF RFC 4567, Key Management Extensions for Session Description Protocol (SDP) and Real Time Streaming Protocol (RTSP)

[16] IETF RFC 4568, Session Description Protocol (SDP) Security Descriptions for Media Streams

[17] IETF RFC 4650, HMAC-Authenticated Diffie-Hellman for Multimedia Internet KEYing (MIKEY)

[18] IETF RFC 4738, MIKEY-RSA-R: An Additional Mode of Key Distribution in Multimedia Internet KEYing (MIKEY)

[19] IETF RFC 4916, Connected Identity in the Session Initiation Protocol (SIP)

[20] IETF RFC 4961, Symmetric RTP / RTP Control Protocol (RTCP)

[21] IETF RFC 5626, Managing Client-Initiated Connections in the Session Initiation Protocol (SIP)

[22] IETF RFC 5630, The Use of the SIPS URI Scheme in the Session Initiation Protocol (SIP)

[23] IETF RFC 5761, Multiplexing RTP Data and Control Packets on a Single Port

[24] IETF RFC 5763, Framework for Establishing a Secure Real-time Transport Protocol (SRTP) Security Context Using Datagram Transport Layer Security (DTLS)

[25] IETF RFC 5764, Datagram Transport Layer Security (DTLS) Extension to Establish Keys for the Secure Real-time Transport Protocol (SRTP)

[26] IETF draft-ietf-sip-connect-reuse-14, Connection Reuse in the Session Initiation Protocol (SIP)

NOTE At the time of publication of this Technical Report, the IETF had approved this draft as a standards track RFC but had not published the RFC and had not allocated an RFC number. If the draft is no longer available, readers should look for the RFC with the same title.

[27] IETF draft-ietf-sipcore-location-conveyance-02, Location Conveyance for the Session Initiation Protocol

NOTE At the time of publication of this Technical Report, the IETF had not completed the approval process for this draft and had not allocated an RFC number. If the draft (or a later version) is no longer available, readers should look for the RFC with the same title.

[28] IETF draft-zimmermann-avt-zrtp-16, ZRTP: Media Path Key Agreement for Secure RTP

NOTE At the time of publication of this Technical Report, the IETF had not published this as an informational RFC. If the draft (or a later version) is no longer available, readers should look for the RFC with the same title.

[29] ITU-T Recommendation E.164, The international public telecommunication numbering plan

[30] ISO/IEC 9594-8 | ITU-T Rec. X.509, Information technology - Open Systems Interconnection - The Directory: Public-key and attribute certificate frameworks

[31] 3GPP TS 33.203, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3G security; Access security for IP-based services (Release 8)

[32] 3GPP TS 33.210, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3G security; Network domain security; IP network layer security (Release 8)

[33] 3GPP TS 33.310, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Network domain security; Authentication Framework (AF) (Release 8)

[34] ETSI TS 187 003, Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); NGN Security; Security Architecture

[35] IEEE 802.1x, IEEE Standard for Local and metropolitan area networks - Port-Based Network Access Control (2004)

[36] IEEE 802.11, IEEE Standard for Information Technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific Requirements - Part 11: Wireless LAN Media Access Control (MAC) and Physical Layer (PHY) Specifications (2007)

[37] OASIS, Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0 (March 2005)

[38] ISO/IEC 27001, Information technology - Security techniques - Information security management systems - Requirements

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

3.1 External definitions

This Technical Report uses the following terms defined in ISO/IEC TR 12860 [1]:

- Domain
- Enterprise network
- Next Generation Corporate Network (NGCN)
- Next Generation Network (NGN)
- Private network traffic
- Public network traffic
- Session Service Provider (SSP)

- SIP intermediary

3.2 Other definitions

None.

4 Abbreviations

AIB	Authenticated Identity Body
AKA	Authentication and Key Agreement
CA	Certification Authority
B2BUA	Back-to-Back UA
DECT	Digital Enhanced Cordless Telecommunications
DoS	Denial of Service
DTLS	Datagram Transport Layer Security
DNS	Domain Name System
GAN	Generic Access Network
IMS	IP Multimedia Subsystem
IP	Internet Protocol
IPsec	Internet Protocol Security
LAN	Local Area Network
MIKEY	Multimedia Internet KEYing
NAT	Network Address Translation
NGCN	Next Generation Corporate Network
NGN	Next Generation Network
PAI	P-Asserted-Identity
PIN	Personal Identification Number
PKI	Public Key Infrastructure
PLMN	Public Land Mobile Network
PSTN	Public Switched Telephone Network
RTCP	Real-time Transport Control Protocol
RTP	Real-time Transport Protocol
S/MIME	Secure Multi-media Internet Mail Extensions
SBC	Session Border Controller
SDP	Session Description Protocol
SIP	Session Initiation Protocol
SRTCP	Secure Real-time Transport Control Protocol
SRTP	Secure RTP
SSP	Session Service Provider
TCP	Transaction Control Protocol
TLS	Transport Layer Security
UA	User Agent

UAC	User Agent Client
UAS	User Agent Server
UDP	User Datagram Protocol
URI	Universal Resource Identifier
VPN	Virtual Private Network
WLAN	Wireless LAN

5 Background

General concepts of NGCNs are discussed in ISO/IEC TR 12860 [1]. In particular, that document describes use of the Session Initiation Protocol (SIP) [4] for session level communications within enterprise networks and with other domains. It focuses on enterprise networks based on enterprise infrastructure (NGCN), but also covers hosting on other networks, in particular NGNs, using the same infrastructure that supports public networks.

ISO/IEC TR 12860 describes the basic communications architecture of an NGCN as comprising three levels (transport, session and application), together with security and management capabilities spanning all three levels. This reflects the fact that security vulnerabilities can arise at all three levels, and therefore appropriate security measures need to be put in place at and across all three levels. Security measures aim to ensure data integrity and confidentiality, authentication of parties, authorisation, and prevention of denial of service (DoS) attacks. For example, at the transport level, measures may be taken to authenticate equipment when connecting to a Local Area Network (LAN), e.g., using IEEE 802.1x [35] or to protect communications on a Wireless LAN (WLAN), e.g., using Robust Security Network Association (RSNA) [36] (known commercially by the WiFi Alliance as Wireless Protected Access 2, WPA2). Alternatively Virtual Private Network (VPN) technologies can be used (e.g., based on TLS [12] or IPsec [11]), particularly when accessing an NGCN from an untrusted LAN or WLAN. Session level and application level communications can to some extent rely on underlying security at the transport level, but in general this is insufficient for achieving end-to-end security.

At the session level, the signalling protocol (SIP) is used to negotiate the parameters needed to allow session-related media streams to flow between endpoints. In the case of real-time media (audio, video), transport is achieved using the Real-time Transport Protocol (RTP) [6]. For non-real-time media conventional transports such as the Transport Control Protocol (TCP) can be used directly. Both the security of SIP signalling and the security of media need to be considered.

This Technical Report analyses the needs and available mechanisms for securing SIP signalling and for securing real-time media transported over RTP. Signalling security is discussed in clause 7. Media security, including the security of any signalling in support of media security, is discussed in clause 8. The securing of non-real-time media is not considered.

6 General principles

6.1 Threats and counter-measures

Information technology security involves the careful balancing of threats and counter-measures. Threats can be assessed in terms of:

- how easily can a vulnerability be exploited; and
- the amount of damage that can be inflicted by a successful attack.

Counter-measures generally incur some costs, such as:

- the cost of purchasing or licensing additional hardware or software (e.g., card readers, biometric devices);

- ongoing operational costs (in terms of dealing with forgotten passwords, handing out and replacing certificates, etc.);
- inconvenience to the user (e.g., the need to enter passwords and Personal Identification Numbers (PINs));
- reduced performance arising from increased computation times.

Thus costly measures are not normally put in place to counter an attack that is difficult to stage and can inflict relatively little damage. On the other hand, an attack that is easily staged and can inflict substantial damage will almost certainly need to be countered.

In enterprises, damage is often assessed in terms of the amount of financial damage that can be inflicted on the enterprise, either directly (e.g., by stealing money or goods) or indirectly (e.g., by stealing trade secrets or marketing plans, or by causing disruption to an enterprise's normal operations). Also enterprises have certain legal and moral obligations (e.g., data protection, retention of data), and damage can be done to a company's reputation if these obligations fail to be met because of inadequate security. The term enterprise network is often applied to similar communication networks operated by non-commercial organisations (e.g., military, educational, medical), and similar security principles apply to these networks, although weightings might be different. For example, in a military network the consequences of certain attacks might be considered greater, and therefore more costly counter-measures might be justified.

For a given enterprise, security policy [38] will determine the particular measures taken. Security policy can change as an enterprise becomes aware of or the victim of new threats. Tools exist to assist in analysing risks and recommending counter-measures.

6.2 Threats to session level security

A secure network aims to support communications with the following properties (among others):

- authenticity, whereby the claimed identity of an entity (e.g., a user) is correct;
- integrity, whereby information has not been altered or destroyed in an unauthorised manner;
- confidentiality, whereby information has not been disclosed to an unauthorised individual;
- privacy, which is closely related to confidentiality and concerns the right of an individual to control how information related to the individual is held or disclosed to others;
- availability, whereby network services to legitimate users are not denied through unauthorised intervention;
- non-repudiation, whereby proof of involvement of a party in certain actions (e.g., sending or receiving a message, stating something during a call) is secured and stored for later use.

For session level communications, successful attacks can compromise these properties in various ways, e.g.:

- eavesdropping on media (e.g., voice, video, messaging);
- discovery of call details (who called whom and when);
- discovery of private information relating to a user (e.g., his current geographic location);
- discovering a user's password, PIN or other credentials;
- masquerading, such as calling somebody and presenting a false caller identifier;
- injecting unwanted media into a call between two legitimate users;

- injecting unreasonable amounts of unwanted calls or messages towards users;
- causing network equipment to malfunction in a way that denies service completely, impairs performance or causes incorrect behaviour (DoS);
- refuting involvement in an earlier action (e.g., refuting earlier participation in a session or refuting the sending or receiving of a message).

Threats come from inside or outside the enterprise. Even internal communications can use a variety resources, not all of which are owned by the enterprise or located on enterprise premises, for example:

- the enterprise's own IP infrastructure and session level resource;
- those of a hosting organisation; or
- any IP infrastructure used for interconnection of remote sites or users, including the public Internet.

All these form part of the enterprise network as a whole. For external communications, with parties outside the enterprise network, threats can also arise in the public networks or other enterprise networks involved.

Threats can arise through various means, such as:

- wire tapping, through access to the physical interconnection media;
- unauthorised access to or inappropriate use of wireless infrastructure;
- unauthorised access to or inappropriate use of network resources, such as SIP intermediaries, routers, switches, call recorders, etc.;
- unauthorised access to or inappropriate use of authorised devices attached to the network, e.g., PCs, fixed or mobile phones, PDAs;
- the connection of unauthorised devices to the network;
- unauthorised access to the network from other networks such as the public Internet.

Many of these threats can be countered by measures below the session layer, such as physical security (locked rooms or buildings), LAN access security (e.g., IEEE 802.1x), WiFi security and firewalls. However, this still leaves considerable scope for attacking session level communications, and measures need to be put in place at the session level to counter such threats.

It should also be noted that threats can arise outside the IP environment. For example when Digital Enhanced Cordless Telecommunications (DECT) terminals are used (connecting via a gateway to the IP infrastructure), some of the threats described above are viable if appropriate DECT security measures are not taken. Threats arising outside the IP environment are not addressed in this Technical Report.

6.3 Authorisation

At the session level, certain things are subject to authorisation, e.g.:

- whether a call can be established;
- whether and to what degree a resource can be used;
- whether an incoming call can be accepted; and
- whether a particular feature can be used.

Authorisation is generally based on the authenticated identity of the requesting user. The higher the 'value' of the resource requested, the stronger the authentication needs to be. This Technical Report discusses authentication but does not discuss authorisation, which is a policy matter for the enterprise concerned.

6.4 Security and mobile users

Additional security considerations can apply to mobile users. Security measures at the transport level (e.g., WLAN security, Generic Access Network (GAN) security, VPN technologies) are relevant. For example, where a mobile user has VPN connectivity back to the NGCN, mechanisms as discussed in this Technical Report should be sufficient for providing the necessary session level security. On the other hand, where session level capabilities of a visited network (e.g., an NGN or Public Land Mobile Network, PLMN) are involved, additional security considerations arise. These are outside the scope of this Technical Report and were the subject of separate ongoing studies in Ecma at the time of publication of this Technical Report.

6.5 Security and NGN

Much work has been done on security of the IP Multimedia Subsystem (IMS) in 3GPP and on NGN security in ETSI TISPAN and ITU-T Study Group 13 (in collaboration with Study Group 17). NGN security (including IMS security) has an impact on enterprise networks when NGCNs interwork with NGN and when enterprise functionality is hosted on NGN infrastructure. Enterprise security policy needs to consider NGN involvement, particularly for private network traffic (between NGCN sites or for NGN-hosted enterprise users). Other traffic will be subject to NGN policy for public network traffic, but enterprise security policy should at least consider the security of public network traffic at the NGCN-NGN interface. Enterprise security policy will influence contractual negotiations with NGN operators.

RECOMMENDATION 1: As part of their security policy, enterprises should consider the level of security needed for private network traffic within an NGN, as well as for public network traffic at the NGCN-NGN interface. This policy should be taken into account during contractual negotiations with NGN operators.

6.6 Security and software status

Security of signalling and media can also be impacted by the status of relevant software in participating devices, including devices that contain SIP UAs and/or SIP intermediaries. Within an enterprise, this is generally governed by policy. In particular, an incorrect version or patch level of software such as the operating system, the TCP/UDP/IP stack, the SIP stack, the RTP stack, relevant security modules or anti-malware software can introduce security vulnerabilities. This aspect is not considered further in this Technical Report.

6.7 Call recording and audit

Enterprises have various needs to keep certain data for future reference, e.g., for non-repudiation, training and quality control, compliance with regulations and other standards, or cooperation with law enforcement agencies. Data retained might just comprise session details (e.g., session times and participants) or might extend to the actual media (see also 8.4). This represents a challenge, in that information needs to be transmitted securely between participants and network intermediaries, yet when recorded needs to be available in unencrypted form to authorised personnel.

7 Signalling security

Signalling security is important for several reasons. First, it is important to authenticate the source of a message, in order to authorise any action to be taken on the message. Otherwise an attacker could gain unauthorised access to services or information or carry out a denial of service attack. Secondly, even if the source of a message is authenticated, it is important to be sure that data integrity has not been compromised en route. Thirdly, signalling can contain sensitive information (e.g., the identities of parties in communication, IP addresses and ports to be used for media reception, etc.). This information needs to be protected against

eavesdropping. Fourthly, non-repudiation may be important. For example, signalling reveals which entities are participating in a session, and if those entities have been strongly authenticated and evidence of this is stored securely for future use, it becomes difficult to refute involvement of those entities. If those entities are strongly tied to human users, then those users would find it difficult to refute involvement.

Although there may be protection provided by the underlying network infrastructure, this is very much dependent on the particular network segment, its type and security measures. A locally secure network infrastructure (e.g., security over a WLAN segment) does not guarantee security over the entire signalling path to the next SIP entity, and certainly does not guarantee end-to-end security when SIP intermediaries are involved.

7.1 Security of access to session level services

Session level services are supported by SIP intermediaries, which provide functions such as registrar, proxy, etc.. Except where peer-to-peer SIP signalling is used, which is outside the scope of this Technical Report, all SIP signalling is between a UA and a SIP intermediary, or between two SIP intermediaries. Often a UA will not accept a SIP request that has not arrived via a known SIP intermediary.

When a UA issues a SIP request, it sends it to a SIP intermediary. Before the SIP intermediary can carry out the requested function (e.g., register the UA or establish or clear down a call), it needs to seek authorisation, and that generally requires obtaining an authenticated identity of the entity making the request. SIP provides a means of doing this through SIP digest authentication, which is specified in [4] and is adapted from Hyper-Text Transfer Protocol (HTTP) digest authentication. In this way, a SIP intermediary can challenge a SIP UA to provide evidence that it is in possession of credentials in the form of a shared secret. These credentials relate to a SIP or SIPS URI identifying the UA's user, which means the SIP intermediary is able to authenticate the UA as being an agent for a given user, and it is on this basis that it authorises access to SIP services. The risks involved in allowing an unauthenticated entity to use session level services are fairly high (e.g., misuse of resources, fraudulent use of services that attract a charge, misleading other users), and therefore SIP digest authentication can always be expected to be used in enterprise networks.

In the case of a human user, this assumes that the user concerned is indeed the person using the UA to access SIP services. UAs can provide their own means to help ensure this is the case, e.g., by requiring the user to submit one or more forms of identification (e.g., a password, a smart card or biometric evidence) in order to "log on" to the UA and activate SIP services. Alternatively a single sign-on mechanism can be used, whereby a user who has already logged on to another service can gain access to session level services based on the credentials already submitted. The Security Assertion Mark-up Language (SAML) [37] could be a basis for this.

Of course, even a strong sign-on mechanism doesn't protect against bad practices by the user, e.g., leaving the device logged on but unattended and accessible to others, logging on but then inviting a colleague to use the device, or disclosing the password to a colleague. Therefore except in very strictly controlled environments, authentication of a UA only provides evidence that the UA possesses credentials associated with a user, and does not provide evidence that the human user is actually present.

NOTE In effect, SIP digest authentication is a form of device authentication rather than user authentication. It takes place at the session level, and is independent of any device authentication that might take place at the transport level, e.g., using IEEE 802.1x.

Call detail recording and other accounting measures can provide a record of what session level services a user has used and can be used for billing, traceability, etc.. It can help to track down excessive or inappropriate use, and perhaps help to uncover a person masquerading as an authorised user. Furthermore, if recorded details are based on authenticated and integrity protected signalling messages, and if records are stored securely, this can provide satisfactory proof of involvement of an entity In a session, thereby supporting non-repudiation.

7.2 Securing a SIP signalling hop

Because SIP uses SIP intermediaries, signalling between two UAs in general comprises one or more hops (e.g., from UA to SIP intermediary, SIP intermediary to next SIP intermediary, etc.). Signalling can be secured

separately on each hop using general purpose security protocols at the transport layer (TLS [12]) or the network layer (IPsec [11]). These can provide authentication, integrity protection and secrecy. SIP digest authentication plays only a limited role.

Security should normally be considered for any SIP signalling hop within and at the boundaries of an enterprise, unless the infrastructure is protected by other means (e.g., physical protection by lock and key, Media Access Control (MAC) layer security). Risks involved include eavesdropping on potentially sensitive information (e.g., who is communicating with whom, identifiers that are subject to privacy) and the ability to inject or modify signalling information for malicious purposes.

7.2.1 TLS for securing SIP signalling

With TLS, authentication is achieved using public key cryptography, whereby an entity has a private key for signing messages and a certificate that it can publish to allow other entities to verify its signature. Between two SIP intermediaries or between a SIP intermediary and a UA such as a gateway or media server, mutual certificate-based authentication is the norm. However, between a SIP UA serving an individual user (e.g., a phone) and the SIP intermediary to which it connects, TLS server authentication alone is often used. This is because there is often no Public Key Infrastructure (PKI) available capable of deploying signed certificates to large numbers of these UAs, and therefore TLS mutual authentication is not feasible. Although using TLS server authentication these UAs can still authenticate the SIP intermediary, SIP digest authentication (see 7.1) is normally used to authenticate the UA.

NOTE In this case the connection must always be established in the direction UA to SIP intermediary, so that the SIP intermediary can act as the server. TLS does not provide client-only authentication.

An established secure connection should normally be retained on a semi-permanent basis for use by multiple SIP transactions in either direction, because of the overhead involved in establishing a new secure connection for each transaction. In particular, if connections cannot be established in the direction SIP intermediary to UA, because of the absence of a certificate at the UA, connections established by the UA should be retained in order for inbound requests to be delivered to the UA. A mechanism for achieving this is specified in [21]. For other cases (where TLS mutual authentication is used), a mechanism for connection reuse is specified in [26].

7.2.2 IPsec for security SIP signalling

With IPsec, there are various means for achieving authentication. For example, the method specified by 3GPP for use at the accesses of 3rd generation mobile networks, and also adopted by ETSI for the accesses of NGNs, is known as IP Multimedia Subsystem (IMS) Authentication and Key Agreement (AKA) [31].

Because TLS operates directly below the SIP layer, SIP software is able to force the use of TLS, where required. With IPsec, the SIP layer uses an unsecured transport protocol (e.g., TCP, UDP) and SIP software has no visibility of any underlying security at the network layer and must rely on management to ensure that security is in place. IPsec also requires special provisions for Network Address Translation (NAT) traversal.

TLS runs only over TCP transport, whereas IPsec can also run over UDP. The use of Datagram TLS (DTLS) to secure SIP signalling over UDP has not been standardised. UDP for SIP suffers from the problem that many SIP messages can exceed the maximum payload size of UDP, leading to packet fragmentation and reassembly, which is not reliable. For this reason, TCP is to be preferred, in which case TLS can be used.

7.2.3 The role of SIP digest authentication

Although SIP digest authentication can be used to authenticate a UAC that does not have a certificate by which it can be authenticated using TLS, SIP digest authentication is not a general solution for authenticating a SIP entity. The following points are relevant:

1. SIP digest authentication operates between two SIP entities that possess a shared secret, typically between a UA and the SIP intermediary that provides the domain proxy/registrar function. There may be other SIP intermediaries between the two, e.g., an edge proxy. In this case, TLS authentication operates separately across each hop, i.e., within the context of a single TCP connection. SIP digest authentication operates "end-to-end" across these two (or more) hops.

2. SIP digest authentication applies only to requests. TLS authentication applies to any SIP messages transported over the TLS connection, and therefore applies to responses as well as requests.

3. SIP digest authentication allows a UA to authenticate another UA, but this depends on possession of a shared secret, which is unlikely to the be the case between two UAs representing end users. A more likely use is where one of the UAs represents a central resource, such as a presence server , a PSTN gateway or a media server.

4. SIP digest authentication does not allow a SIP proxy to be authenticated by a UA or another SIP proxy. In theory a SIP intermediary in the form of a B2BUA, as opposed to a proxy, can be authenticated in this way, and in practice this is sometimes done. For example, it is sometimes used by between a Session Service Provider (SSP, e.g., an NGN) and an NGCN.

Point 1 above leads to the following consideration. TLS authentication takes place once, on establishment of the TLS session, and applies to all data transported during the session. On the other hand, SIP digest authentication applies only to a single SIP request and needs to be used on potentially every request. Just because two requests arrive over the same TLS session and the sender of the first request has been SIP digest authenticated, does not mean that the sender of the second request does not need to be SIP digest authenticated too. The requests could come from two different UAs, via a SIP intermediary that terminates the TLS session. Unless there is some means by which the UAC is bound to the termination of the TLS session (TLS client or TLS server), the two UACs could be different, with the risk that one might be masquerading.

Similarly, if a SIP response arrives over a TLS session, it does not necessarily mean that it comes from a UA that has been digest-authenticated on an earlier request over that TLS session.

7.3 Ensuring that all SIP signalling hops are secured

For any given signalling transaction, security will be expected on every hop linking the peer UAs (hop-by-hop security). The risks involved in any one link being unsecured are as mentioned in 7.2. A UA has visibility only of its local hop (between the UA and the next SIP intermediary), and there is no reliable way for a UA to be aware of the security status of further hops. The SIPS URI scheme, as specified in RFC 3261, is aimed at assuring hop-by-hop security, although it suffers from a number of weaknesses:

- Originally, in RFC 3261, TLS was mandated on all hops except the last (from the target domain proxy to the UAS). Although the IETF has now deprecated this last hop exception [22], current implementations that permit this last hop exception in accordance with RFC 3261 are likely to be around for some time.

- The SIPS URI scheme mandates TLS and does not allow for other means of security that might give equivalent protection (e.g., IPsec).

- The SIPS URI scheme does not mandate particular cryptographic algorithms for use with TLS, and therefore the precise level of protection cannot be determined - it will depend on algorithms used by the weakest link.

- UAs have to trust all SIP intermediaries to honour the SIPS URI in accordance with RFC 3261 by ensuring each hop is secured.

- There is no "best effort" version of the SIPS URI scheme mechanism, i.e., a request will be rejected if TLS is not available, rather than falling back to unsecured (effectively the corresponding SIP URI) and informing the UAC.

- The UAS cannot tell reliably whether an incoming request is secured on all hops.

- There were lots of instances of lack of clarity and ambiguity in RFC 3261 concerning the SIPS URI scheme, although these have now been addressed by the IETF [22].

Although specification of the SIPS URI scheme is much improved in [22], compared to [4], until there is widespread deployment of implementations compliant with [22], the security status of a call established using SIPS will be questionable.

Within a single domain, or within a multi-domain enterprise network, it might be known that all signalling hops are provisioned as secure, and hence there might be less need for deploying the SIPS URI scheme. However, for more general inter-domain deployments, the SIPS URI scheme might be useful.

7.4 End-to-end signalling security

Notwithstanding the limitations of the SIPS URI scheme for ensuring all SIP signalling hops are secured, hop-by-hop signalling security, even when present on all hops, may be considered insufficient for some purposes, since SIP intermediaries still have visibility of and can modify all signalling information. This means that the risks mentioned in 7.2 can arise at a compromised SIP intermediary. Of course, for some information visibility and modification are necessary for normal routing purposes. Some SIP intermediaries examine or modify information for other legitimate purposes, such as call admission control (including QoS provision) or NAT/firewall traversal. This leaves information of a pure end-to-end nature such as security keys for media protection, geographic location information and instant messages that ideally should be concealed from SIP intermediaries. For this last type of information, end-to-end signalling security would appear to be required.

If only hop-by-hop security is available, the UA has to trust the first SIP intermediary to behave correctly by not unnecessarily disclosing or modifying message content of an end-to-end nature and by ensuring that the next signalling hop is appropriately secured. Similarly, the first SIP intermediary has to trust the second SIP intermediary, and so on. This transitive trust model is generally viewed as leading to poor security. Each entity might be happy to trust the next entity (e.g., a SIP entity in an NGCN might trust a SIP entity in an NGN with which the NGCN has a business relationship), but probably has no idea what entities are involved beyond the next entity and no idea whether they can be trusted.

Within a domain, or perhaps within a multi-domain enterprise network, end-to-end security may not be so important if all signalling hops are provisioned as secure or the SIPS URI scheme is used, provided SIP intermediaries can be trusted (e.g., they are known to comply with the enterprise security policy). For inter-domain working outside the enterprise network and when roaming, end-to-end security is likely to be of greater importance. As more domains are involved, the threats increase, and the need to be confident of signalling security on each hop is much greater. It can be regarded as essential when signalling over the public Internet without any underlying security such as a VPN tunnel.

Two mechanisms address end-to-end signalling security: S/MIME and SIP Identity.

7.4.1 End-to-end security using S/MIME

SIP allows bodies of SIP messages to be encrypted, authenticated and integrity protected using S/MIME (Secure Multi-media Internet Mail Extensions). To use this capability, UAs need private keys and certificates, implying the need for a PKI to provide this information. Partly for this reason, and partly because difficulties obtaining a peer's certificate prior to sending encrypted information, S/MIME for SIP bodies has not been adopted by the market, even though it has been specified since the publication of RFC 3261. There has even been talk in the IETF of deprecating S/MIME.

One potential use of S/MIME is for geographic location conveyance in SIP in accordance with [27]. This specifies two mechanisms for location conveyance in SIP: location by value, whereby the location object [10] is conveyed within the body of a SIP request; and location by reference, whereby a URI pointing to a resource from which the location can be obtained is conveyed in the header of the SIP request. Location by reference does not require end-to-end security, since security can be achieved during dereferencing (by authenticating and authorising the dereferencing entity). Location by value frequently does require that the location be encrypted to prevent eavesdropping. However, one potential use of location by value is during emergency call establishment (see [3]), where end-to-end encryption cannot be used because of the need for SIP intermediaries to access location to determine where to route the call (although hop-by-hop encryption should of course be used if available). Location by reference might incur unnecessary delays at SIP intermediaries during emergency call establishment, as well as being an additional point of potential failure. In general, it should be sufficient to use location by reference where end-to-end security is required and location by value with hop-by-hop security (e.g., TLS) where SIP intermediaries need fast access to location information (such as for emergency call establishment). Therefore location conveyance does not seem to be a compelling reason to use S/MIME in enterprise networks.

RECOMMENDATION 2: Enterprise networks should not use S/MIME as a means of achieving end-to-end signalling security in SIP.

RECOMMENDATION 3: For conveyance of geographic location in SIP, enterprise networks should use location by reference where end-to-end security is required and rely on hop-by-hop security with location by value where SIP intermediaries need fast access to location information (e.g., for emergency call establishment).

7.4.2 Near end-to-end security using SIP Identity

SIP Identity (RFC 4474 [14]) provides near end-to-end authentication of requests and partial integrity protection of request contents. However, it does not provide encryption and does not protect responses in any way. The basis for SIP Identity is the Identity header field (and associated Identity-Info header field), whereby the SIP intermediary at the originating domain can sign an assertion that the URI in the From header field is correct, having authenticated the UA by other means (typically SIP digest authentication). This cryptographically signed assertion can be passed to the destination, where it can be verified, subject to the signer's certificate chain being acceptable. Besides the From header field URI, the signature covers several other parts of the request, including the entire body and information from the To, Date, Contact, Call-Id and CSeq header fields, thereby providing integrity protection across these parts of the message and some replay protection. RFC 4474 as it stands is for requests outside the context of an existing dialog, but RFC 4916 [19] extends its use to mid-dialog requests.

NOTE SIP Identity is applicable only to SIP requests, not responses. This is because of difficulties authenticating a UAS.

The near end-to-end authentication and integrity protection provided by SIP Identity will be broken if any domain modifies SIP bodies or certain SIP header fields when forwarding a SIP request. This is particularly true of Session Border Controllers (SBCs) that modify information in Session Description Protocol (SDP) bodies. One common reasons for this is for media steering, where an SBC modifies IP addresses and ports in SDP in order to steer media over particular routes, e.g., to ensure quality of service. This is a problem for which the IETF does not yet have a solution, even though the IETF is working on other security solutions that rely on SIP Identity. Notwithstanding these limitations with the current version of SIP Identity, it seems fairly certain that the technique will become an important component of SIP security, for example for underpinning key management for media security (see **8.3**).

STANDARDISATION GAP 1. Problems with deploying SIP Identity need to be addressed, so that a satisfactory solution is available for near end-to-end authentication and integrity protection.

RECOMMENDATION 4: Enterprise networks should use SIP Identity [14] [19] for near end-to-end authentication and integrity protection of inter-domain signalling traffic, subject to the finding solutions to or work-arounds for its known problems. See also 7.5 and 8.3.

7.5 Authenticated identity delivery

An important aspect of SIP is its ability to deliver to a UA information identifying the other participant(s) in a call, as discussed in [2]. This introduces an important security issue, namely the authenticity of a delivered identifier. A delivered identifier can be used for many purposes, such as:

- helping a user to decide whether to answer a call or accept a message;
- helping a user to decide whether to disclose sensitive information or to trust information received;
- helping an automatic filtering application to filter out unwanted calls or messages (often known as "spit" and "spam");
- helping an application to pull down caller-related information from a database;
- recording answered or missed messages or calls and facilitating subsequent return messages or calls.

For some of these purposes, it is important that the delivered identifier be authenticated, and risks involved in not doing so can include subjecting a user to unwanted traffic, tricking a user into disclosing sensitive information, etc.. An authenticated identifier is also needed for non-repudiation purposes.

Considering a SIP request, the From header field URI (From URI) provides the identity of the user on behalf of whom the request is issued. Unfortunately that can be forged by the UAC and, if not rectified by a SIP intermediary, can result in incorrect information being delivered to the UAS. Furthermore, even a correct From URI can be altered by a well-meaning or malicious SIP intermediary, again resulting in incorrect information being delivered to the UAS. In the case of a SIP INVITE request, for example, it could result in the request passing a white-list check that it would not otherwise pass, resulting in an unwanted call being presented to the called user. It could also result in the called user placing undue trust in the identity of the caller, and perhaps disclosing sensitive information through one or more media. In the case of a SIP MESSAGE request it might lead a user to trust the message contents and perhaps even reply with sensitive information. Therefore a mechanism is required that provides evidence of authenticity of a URI representing the source of a SIP request and provides integrity protection to prevent tampering by SIP intermediaries.

SIP digest authentication is not in general considered a suitable method for achieving authenticated identity between UAs, because, as discussed in 7.2.3, it depends on shared secrets.

7.5.1 P-Asserted-Identity (PAI)

A preliminary and partial solution to the problem was the P-Asserted-Identity (PAI) header field defined in RFC 3325 [5]. A SIP intermediary that has authenticated the UAC (e.g., using SIP digest authentication) can insert a PAI header field in the forwarded request, asserting the correct identity of the user on behalf of whom the request is issued. Any downstream SIP intermediary, if it trusts the entity from which it has received the request, will simply forward the PAI header field further downstream. This has several weaknesses: there is no integrity protection to prevent tampering by downstream SIP intermediaries; and there is no indication of which entity is making the assertion. As the PAI header field is passed on from one SIP intermediary to the next and eventually to the UAS, perhaps through multiple domains, each entity has to trust the previous one. This is the transitive trust model described in 7.4. Although the use of PAI will often be sufficient within a single domain enterprise network or perhaps even within a multi-domain enterprise network, for more general inter-domain communication it does not represent a secure solution.

7.5.2 Authenticated Identity Body (AIB)

A cryptographic solution is the Authenticated Identity Body (AIB) defined in RFC 3893 [9]. Using S/MIME, the UAC inserts a signed fragment of the SIP message header, including the From header field, as an S/MIME body or body part. This requires the UAC to have a certificate whose subject matches the identity it is asserting. This solution, whilst having good security properties, has not seen deployment, probably for the same reasons S/MIME in SIP in general has not seen deployment. Furthermore, if S/MIME were to be deprecated, so too would AIB.

7.5.3 SIP Identity

SIP Identity (see 7.4.2) provides near end-to-end (or end-domain-to-end-domain) authenticated identity. Within the IETF, this is the generally accepted mechanism for authenticated identity, rather than AIB. Although much newer than AIB, the SIP Identity too has seen little deployment. This is partly due to the problem described in 7.4.2, whereby the signature can be broken by SIP intermediaries such as SBCs.

Another concern is with SIP or SIPS URIs based on E.164 [29] numbers. As described in ISO/IEC TR 12861 [2], a global E.164 number does not require the domain part of the SIP URI to make it globally unique. Hence it has become common practice to place in the domain part any domain through which the user of the E.164 number can be reached, and to change the domain part as a SIP URI within a SIP message traverses domains. For example, an SSP might use its own domain part, even for an E.164 number that is assigned to an enterprise. Only an entity with a certificate for the domain part in the From URI can sign a request in accordance with SIP Identity. However, if a domain substitutes its own domain name for that in the received From URI when forwarding a request, that domain can sign the request again (replacing any previous signature). Moreover, the fact that actions of SBCs and other B2BUAs can break a SIP Identity signature is motivation for a domain to change the domain part of the From URI so that it can sign the request again. With

other forms of SIP URIs where there is a dependency on the domain part to make the URI globally unique, changing the domain part and signing again is not feasible, but with a globally unique user part (e.g., a global E.164 number) this is possible and likely to occur.

NOTE In fact there are ways of circumventing the problem of URIs that rely on the domain part for global uniqueness. For example, a domain could modify the user part by appending the received domain part, and then substituting its own domain name in the domain part. For example,

sip:john@example1.com

could become

sip:john_example1.com@example2.com

Technically, the practice of a domain other than the originating domain signing a request is likely to be in violation of RFC 4474, because there is often no way for an SSP to authenticate a particular user in an enterprise, even it is sure (e.g., based on TLS authentication) that the request came from the enterprise.

Such practices, whereby the From URI is changed as a SIP request passes through a domain, can cause problems for the destination user. The user might receive a URI in which the domain part is misleading, e.g., representing the user's own SSP or (in the case of an enterprise user) the enterprise's SSP, rather than that of the true origin of the call or instant message. Therefore the user might not know whether the call or message originates at a domain the user wishes to communicate with. This can impact the user's willingness to accept the communication and, in the case of a call, the user's willingness to disclose sensitive information via media.

Such practices can also affect automatic call or message handling at the UAS. For example, if the UAS access a white list that allows calls from a particular domain, or a particular user in a domain, or if the UAS is programmed to forward calls from a particular domain or a particular user in a domain, a manipulated From URI will fail to match, causing incorrect handling.

These collective problems currently prevent the deployment of SIP Identity, except in some relatively simple environments (e.g., within a single domain) where arguably PAI is as good. For proper end-to-end (or end-domain-to-end-domain) authenticated identity in multi-domain environments, SIP Identity is needed, but current practices (in terms of signature breaking and changing the From URI) prevent this working. Notwithstanding these problems, SIP Identity (perhaps in modified form) seems to be the longer term solution to inter-domain authenticated identity.

RECOMMENDATION 5: Suppliers of enterprise networks should consider migration towards use of SIP Identity [14] [19], because of its superior security properties for inter-domain traffic compared to P-Asserted-Identity, subject to the finding solutions to or work-arounds for its known problems.

7.5.4 Authenticated response identity

Authenticated identity in a SIP response can give some assurance that the request reached the intended target or, if not, what entity was reached. In the case of an INVITE request, an identity in the 200 response would be the connected identity, i.e., the user who answered the call. A request can reach an unexpected target for legitimate reasons (e.g., the intended target wanted the request to be forwarded), by accident (e.g., misrouting due to a provisioning error) or because of malicious intervention.

PAI can be used in a response, with all the weaknesses associated with PAI in a request. In particular, PAI in a response is unlikely to be useful in detecting malicious intervention, since the entity that misroutes can also forge PAI in the response. Another issue is how a SIP intermediary can authenticate a UAS, in order to be able to assert an identity in a response, since SIP digest authentication is not available in responses. Receipt of a response over a TLS session over which a UA has previously been digest-authenticated is not necessarily sufficient, as discussed in 7.2.3. Notwithstanding these issues, P-Asserted-Identity is frequently used in responses, these issues seemingly being ignored.

AIB is equally applicable to requests and responses.

RFC 4474 does not make provision for SIP Identity in a response, because of the difficulty authenticating a UAS. In the absence of this, the work-around (in the case of the INVITE method only) is to send a new

request on the same dialog in the reverse direction and apply SIP Identity to that request, as specified in RFC 4916. This does not work for methods that do not result in a dialog (e.g., MESSAGE) and for negative (3xx/4xx/5xx/6xx) responses.

There is a need for a standardised means of achieving authenticated response identity, which requires some way of authenticating the UAS.

RECOMMENDATION 6: As part of migration towards use of SIP Identity, enterprise networks should also consider the use of RFC 4916 as a means of achieving authenticated connected identity, in the absence of authenticated response identity.

7.6 NGN considerations

The risks that lead to the need to secure signalling apply equally, or even to a greater extent, when enterprise communications use an NGN, either for private network traffic (between NGCN sites or for NGN-hosted enterprise users) or for public network traffic (for communicating with the outside world). Risks from unsecured signalling hops can be greater, since the enterprise is not in charge of the transport infrastructure. Furthermore the need for end-to-end security is greater because the enterprise is not in charge of SIP intermediaries.

The NGN security architecture is specified in [34]. As far as signalling is concerned, this is based on IMS security, including IMS access security as specified in [31], for securing the access of a User Equipment (UE) to the IMS, and IP network layer security as specified in [32], for securing signalling between IMS entities. The basis for IMS security is IPsec, together with either IMS Authentication and Key Agreement (AKA) or SIP digest authentication at the access. In addition, recent versions of [31] specify the use of TLS at the access, as an alternative to IPsec for use with access networks not specified by 3GPP, and therefore this can be used with NGN access networks. Also [31] allows the use of TLS for signalling between IMS and a non-IMS SIP proxy. An authentication framework for network domain security is specified in [33], which addresses the use of certificates in support of inter-domain security using TLS or IPsec.

This can impact enterprise networks in a number of ways:

1. Public network traffic that is carried over NGN to or from the enterprise is subject to IMS signalling security, and hence may be secured by IPsec rather than TLS. According to RFC 3261, this prohibits the forwarding of SIPS requests, but in practice IPsec provides an equivalent degree of security.

2. For interconnection of NGCN to NGN using peering-based business trunking, the use of TLS to secure signalling is possible according to [31] (the NGCN being treated as a non-IMS SIP Proxy), although it depends on the willingness of NGN operators to support TLS in this situation. Thus there may be a need in the short to medium term to secure signalling using IPsec rather than TLS.

3. For interconnection of NGCN to NGN using subscription-based business trunking, the use of TLS to secure signalling is possible according to [31], although it depends on the willingness of NGN operators to support TLS. Thus there may be a need in the short to medium term to secure signalling using IPsec rather than TLS and to use either IMS AKA or SIP digest for authentication. Subscriber Identity Module (SIM) cards in support of IMS AKA are unlikely to be feasible for NGCN SIP entities, and therefore a software-based key storage will be required, Whichever authentication method is chosen, this would authenticate the NGCN (as opposed to any individual NGCN user) to the NGN.

4. The hosting of enterprise networks on NGN infrastructure will involve IMS security, and hence will probably involve the use of IPsec rather than TLS for signalling security and the use of IMS AKA for authenticating hosted UAs.

5. Roaming enterprise users need to know that their communication with the enterprise network is secure.

6. Enterprises need to be sure nobody is masquerading as a roaming enterprise user.

The last two points are mobility-related and require separate consideration in that context.

TLS and SIP digest authentication are the mechanisms specified by the IETF for securing SIP signalling, and therefore will be used within NGCN domains and when peering with domains other than NGN. In the interests of consistency and because of the considerations in 7.2.2, the use of TLS, with SIP digest authentication where appropriate, should be preferred for interworking with NGN.

REQUIREMENT 1: An NGN shall support TLS with server authentication and SIP digest authentication for subscription-based business trunking.

REQUIREMENT 2: An NGN shall support TLS with mutual authentication for peering-based business trunking.

Concerning end-to-end security, NGN support for SIP Identity is questionable. Even transport of an Identity header field across NGN between enterprise domains is unlikely to work at present, because of the issues described in 7.4.2, and also because of the practice of modifying E.164-based From URIs. Therefore the use of SIP Identity in its present form to achieve authenticated identity when interworking with NGN is questionable. NGN relies mainly on P-Asserted-Identity, which suffers from limitations described in 7.5.1. As observed in 7.4.2 and 7.5.3, enterprises will need to migrate towards using SIP Identity (subject to finding solutions to or work-arounds for its known problems), particularly for inter-domain working, and this should include interworking with NGNs. Hence there is an enterprise need for NGNs to support SIP Identity.

Another issue with NGN use of P-Asserted-Identity concerns the extent of a trust domain. If an NGN does not consider an NGCN to be within its trust domain, it will either not accept a P-Asserted-Identity header field from the NGCN or it will only accept known identifiers for the NGCN site concerned. This has problems for calls from other parts of the enterprise network, or calls from outside the enterprise forwarded to the NGN. If the P-Asserted-Identity header field value is not accepted by the NGN, the only way to provide caller identity to the remote user is using the From header field URI, which means SIP Identity is needed for authentication (subject to finding solutions to or work-arounds for the known problems). Similar considerations apply to connected identity.

REQUIREMENT 3: An NGN shall support SIP Identity in accordance with RFC 4474 [14] in order to provide enterprise networks with authenticated identity and near end-to-end authentication and integrity protection of SIP requests, subject to the finding solutions to or work-arounds for the known problems.

Signalling security is an important consideration during contractual negotiations with an NGN, as discussed in 6.5.

7.7 Public Switched Telephony Network (PSTN) interworking

When a SIP UA is involved in a call to or from a PSTN via a gateway, nothing can be known about signalling security within or beyond the PSTN (e.g., in another IP-based network beyond the PSTN). Even if the SIPS URI scheme is used to ensure hop-by-hop security as far as the gateway, wire-tapping or lawful interception in the PSTN or the absence of secure signalling in IP-based networks beyond the PSTN will go undetected.

Furthermore, a calling or connected party identity received from the PSTN may not be wholly trustworthy, because of the transitive trust between PSTN networks and lack of knowledge of procedures used by PSTN networks to screen received identifies. If the UA receives PAI, it suffers from all the stated weaknesses of PAI. If the UA receives AIB or SIP Identity, it provides authentication only as far as the gateway's domain, and the accuracy of any telephone number in the user part cannot be known.

There is no explicit indicator in SIP that a request or response is from a PSTN gateway. The only hint a user might receive is that the domain part in a received URI represents the domain of a gateway.

STANDARDISATION GAP 2. There is no standardised means of indicating in SIP that a call is to/from a PSTN gateway and that any security used within the SIP network terminates at the gateway and therefore is not end-to-end.

8 Media security

Media security is important for several reasons. First, it is important to authenticate the source and check the integrity of a media stream, to prevent unauthorised substitution or modification. Secondly, media needs to be protected against eavesdropping.

Although there may be protection provided at the transport level, this is very much dependent on the particular network segment, its type and security measures. A locally secure network infrastructure (e.g., security over a WLAN segment) does not guarantee security over the entire media path. Media security needs to be end-to-end.

Whilst media security can be important for communication within a single domain, or perhaps even within a multi-domain enterprise network, it becomes even more important for communications outside the enterprise. As more domains are involved, the threats increase, and the need for media security is much greater. It can be regarded as essential when transmitting media over the public Internet without any underlying security such as a VPN tunnel.

Media security is discussed here in terms of security of real-time media (voice and video) transported over RTP. Security is achieved by using Secure RTP (SRTP) [7] and an appropriate key management protocol. Depending on the key management protocol, separate provision may need to be made for authentication.

8.1 SRTP

Media transmitted over RTP can be secured by means of SRTP, which extends RTP by allowing media to be encrypted and by adding information for authentication and integrity checking. Secure RTCP (SRTCP) provides similar capabilities for RTP Control Protocol (RTCP).

Because RTP normally flows end-to-end between UAs, SRTP likewise is end-to-end. Intermediate entities may, however, be involved to provide specific services (e.g., transcoding, conferencing), in which case SRTP will operate between each UA and the intermediate entity (which itself appears as a UA to each of the other UAs).

The use of SRTP and SRTCP, as opposed to RTP and RTCP, is negotiated using SDP.

8.2 Key management for SRTP

To enable the use of SRTP (and SRTCP), a key management capability is required for providing the two UAs with a secret master key from which session keys can be derived, as well as agreeing various security parameters such as algorithms and key lengths. The key management capability must ensure that:

- the secret master key (or information from which to derive it) is delivered to or accepted from only a known, authenticated entity; and
- the secret master key (or information from which to derive it) cannot be eavesdropped en route;
- the various security parameters cannot be modified by a man-in-the middle, with the potential of downgrading the level of security negotiated.

Currently there are several ways of performing key management, either published as RFCs or in draft form. Some use the signalling path to exchange information between the UAs, whereas others run protocols on the media path, prior to starting SRTP. The particular method is negotiated using SDP.

8.2.1 Key management on the signalling path

There are two basic means of performing key management on the signalling path: Multimedia Internet KEYing (MIKEY) [8] and Security Descriptions [16]. In each case key management information is embedded in SDP within SIP messages. At best, the SIP signalling path is only secured hop-by-hop, and therefore SIP intermediaries can inspect or modify any information exchanged. MIKEY provides a number of ways of

securing key management information, but the MIKEY null option (see below) and Security Descriptions rely on separate measures being taken.

8.2.1.1 MIKEY

RFC 3830 [8] specifies the basic MIKEY protocol and several negotiable modes of operation (options). Further options are specified in RFC 4650 [17] and RFC 4738 [18]. These options differ according to whether a key exchange or a Diffie-Hellman key agreement is used. Furthermore they differ according to the method used to secure the key exchange or key agreement and to provide authentication. Security is based either on shared secrets or on private keys and certificates. There is also an option with null encryption and null authentication, i.e., an insecure option whereby key exchange is carried out in the clear. This null option is suitable for use when SDP is secured using S/MIME or when it is know that all signalling hops are secured and the SIP intermediaries can be trusted. MIKEY requests are carried in SDP offers and responses are carried in SDP answers in accordance with RFC 4567 [15].

The MIKEY key exchange or key agreement results in an agreed traffic encryption key generation key, from which each side can compute SRTP master keys and salts for each direction of transmission.

The limited scalability of shared secrets, coupled with the difficulties deploying a PKI for providing private keys and certificates to UAs, has prevented significant deployment of most of these options. The null option has seen some deployment, without S/MIME and therefore relying on hop-by-hop signalling security.

8.2.1.2 Security descriptions

RFC 4568 [16] specifies security descriptions, which is a simple method of carrying security parameters and keys in SDP offers and answers. SRTP master keys and salts are carried in the clear, one in each direction. Although intended for use when SDP is secured using S/MIME, in practice deployment has been without S/MIME, and thus relying on hop-by-hop signalling security.

Although Security Descriptions is similar to the MIKEY null option, in that they both exchange keys without any integrity protection or encryption, the two methods are incompatible. This is because Security Descriptions exchanges SRTP master keys and salts in each direction (in the offer and answer respectively), whereas MIKEY exchanges a single key (in the offer) and each end computes SRTP master keys and salts for the two directions. Therefore the two cannot interwork without decrypting and re-encrypting media.

8.2.1.3 Common aspects of MIKEY and security descriptions

Current MIKEY deployments (null option, without S/MIME) and security descriptions deployments (without S/MIME) exchange keys in the clear and rely on underlying hop-by-hop signalling security and the integrity of SIP intermediaries. Therefore they do not provide a proper end-to-end security solution for media.

Both MIKEY and security descriptions are capable of providing end-to-end security, by using other MIKEY options or by securing SDP with S/MIME. Reluctance to provide a PKI capable of issuing private keys and certificates to large numbers of SIP UAs is a prime reason for these solutions not having been deployed.

Another issue is SIP retargeting, whereby a SIP request carrying an SDP offer can end up at a UA different from that expected and with different credentials. Hence S/MIME or MIKEY encryption based on the known public key of the expected target can result in the actual target being unable to decrypt the information. Some MIKEY options take this into account, either by using Diffie-Hellman (which does not require encryption) or by sending the encrypted key in the SDP answer instead.

Yet another issue is forking, where with security descriptions and some MIKEY options non-answering branches will nevertheless have received the key and can potentially listen in to the call at the answering branch. This might necessitate immediate re-keying.

Finally, key management information is additional information that needs to be passed through SIP intermediaries during call establishment and during mid-call re-keying. This information increases the size of SIP messages, and also introduces additional SIP messaging for re-keying.

RECOMMENDATION 7: Security descriptions and the MIKEY null option should be seen as interim solutions to key management for media security, because of their various limitations. More advanced MIKEY options, or the use of S/MIME with either security descriptions or the MIKEY null option, are not seen as easy to deploy. Enterprises should consider the use of key management on the media path for the medium to long term.

8.2.2 Key management on the media path

In view of the difficulties that have limited the deployment of MIKEY and security descriptions, and hence limited the deployment of SRTP, the IETF has considered alternative approaches. Operating a key management protocol on the media path rather than the signalling path was considered to offer advantages, and the solution adopted by the IETF for further work is based on Datagram TLS (DTLS): DTLS-SRTP. Another solution being pursued by some companies outside the IETF, also with key management on the media path, is known as ZRTP [28].

8.2.2.1 DTLS-SRTP

DTLS [13] is an adaptation of TLS to operate over UDP rather than TCP. Thus the authentication, integrity protection and confidentiality that TLS provides for TCP connections can also be used for UDP-based communications. All the security algorithms of the well-proven TLS apply.

DTLS-SRTP [25] uses DTLS to establish a security session for SRTP. When a media path has been established between the calling and called UAs (i.e., addresses and ports have been exchanged via SDP and any NAT/firewall traversal has been accomplished), DTLS handshake is started to establish the security session. One UA becomes the client and the other the server for DTLS. Extensions to the DTLS handshake 'hello' messages allow the client and server to negotiate the SRTP protection profile to be used. The DTLS key derivation function provides SRTP master keys and master salts for the two SRTP sessions (one in each direction). When the handshake is complete, the transmission of SRTP packets is started. The media path is therefore shared between DTLS and SRTP, i.e., the two types of packet are received on the same port. However, packets are easily distinguishable and can therefore easily be directed to the appropriate stack. The use of DTLS-SRTP is negotiated using SDP offer answer by means of special tokens in the media descriptions concerned.

The DTLS handshake requires the use of certificates, but these can be self-signed, relying on the signalling path to achieve authentication (see 8.3.2).

The number of DTLS-SRTP sessions needed between a calling UA and a called UA depends on a number of factors. First, each medium requires its own DTLS-SRTP session. Second, unless symmetric RTP [20] is used by both UAs for a given medium (i.e. each UA transmits and receives SRTP on the same port), separate DTLS-SRTP sessions are required for each direction of SRTP. Third, unless RTP/RTCP multiplexing [23] is used (i.e., each UA multiplexes SRTP and SRTCP on the same port), separate DTLS-SRTP sessions are required for SRTP and SRTCP. Therefore the minimum requirement is one DTLS-SRTP session per medium, rising to a maximum of four per medium. However, where more than one DTLS-SRTP session is used, DTLS session resumption can be used, whereby public key cryptography operations need to be performed only for the first session, the results being re-used by the other sessions.

Where a call is forked, DTLS-SRTP will be needed between the calling UA and each called UA that potentially could answer the call, e.g., each called UA that alerts the user. This ensures that separate keys are negotiated with each called UA, so that the eventual call between the calling UA and the UA that answers cannot be intercepted by one of the other called UAs.

Clearly there may need to be a number of DTLS-SRTP sessions, and hence handshakes, but the number can be kept to a minimum by using symmetric RTP (more or less standard practice in the industry already) and RTP/RTCP multiplexing. The overhead of these handshakes might seem large compared with, say, using MIKEY at the session level, but DTLS-SRTP has the advantage of operating on the media path, thereby by-passing SIP intermediaries and potentially being able to operate significantly faster. The limiting factor is more likely to be the number of cryptographic transforms a UA needs to perform, particularly a calling UA in a forking situation.

8.2.2.2 ZRTP

ZRTP [28] uses a Diffie-Hellman key agreement to generate a shared key, from which SRTP master keys and salts can be derived. It operates on the media path, prior to SRTP. There are similar considerations to DTLS-SRTP concerning the number of instances of ZRTP. A significant difference from other approaches is its mechanism for authentication (see 8.3).

8.3 Authentication

A call may pass through several domains, and signalling may pass through one or more SIP intermediaries in each of those domains. Media, although it by-passes these SIP intermediaries, is generally steered along a path through the network infrastructure of each of those domains. However many domains and SIP intermediaries are involved on the signalling path, and however many domains the media pass through, the media must be secured end-to-end. Moreover, the user must be given some assurance that this is the case. This means somehow binding media security to the authenticated identities of the communicating users. This is best illustrated by example.

Consider a call from Alice (sip:alice@example1.com) to Bob (sip:bob@example2.com). Alice and Bob wish to have secure audio communication. What this means is that Alice needs some assurance that the secure audio path terminates at a UA that is being used by Bob, and vice versa. To give Alice this assurance, Alice's UA needs cryptographic evidence that the remote end of the secure audio path is a UA having the credentials of sip:bob@example2.com. Similarly Bob's UA needs cryptographic evidence that the remote end of the secure audio path is a UA having the credentials of sip:alice@example1.com.

Now, supposing the call was established via a third party SSP, provider.net. Cryptographic evidence of audio security as far as some media-handling entity (e.g., SBC) in the provider.net network is insufficient to satisfy the requirements of either Alice or Bob. Although media might be secured beyond provider.net, there is no mechanism for obtaining cryptographic evidence of this. Also there can be no cryptographic evidence that the entity in provider.net that terminates media security is not intentionally or accidentally disclosing the audio to an unauthorised party or allowing unauthorised substitution.

What that means is that there needs to be proof that whatever key management protocol is used between the two UAs, it really is those UAs that are engaging in that protocol and agreeing keys to be used for media security, and not some unwanted intermediary. The means of achieving this depends on the key management protocol.

8.3.1 Authentication with key management on the signalling path

Security Descriptions and the MIKEY null option provide no intrinsic means of authentication. However, if the SDP body is authenticated using S/MIME, the certificate used for S/MIME can be the basis for authentication of the entities engaging in key management for SRTP, and hence the basis for authentication of SRTP. Alternatively, SIP Identity (see 7.4.2 and 7.5.3), because is signs all body parts (including SDP) as well as some header fields, binds the key management protocol to the authenticated user of signalling, and hence achieves near end-to-end authentication of SRTP.

Other MIKEY options provide authentication based on shared secrets or certificates.

The need to provide UAs with certificates or shared secrets has prevented the deployment of MIKEY options other than null, and likewise has prevented the deployment of S/MIME. SIP Identity seems to provide a suitable solution for Security Descriptions or the MIKEY null option, but, for reasons discussed in 7.4.2 and 7.5.3, there are some deployment problems with SIP Identity. If SIP Identity is not available, hop-by-hop signalling security based on SIPS would still give some measure of protection.

Another issue is that the MIKEY options based on certificates in the UAs require the certificate to identify the user. If some other signalling mechanism is used to provide authenticated identification (i.e., AIB or SIP Identity) and the certificate for that is not the same as the certificate used for MIKEY, it leaves the peer UA with a problem of correlating media with signalling.