A New Authentication Mechanism and Key Agreement Protocol for SIP Using Identity-based Cryptography

Jared Ring Kim-Kwang Raymond Choo Ernest Foo
Mark Looi

Information Security Institute
Queensland University of Technology
GPO Box 2434, Brisbane, QLD 4001, AUSTRALIA
{j.ring,k.choo,e.foo,m.looi}@qut.edu.au

Abstract

The Session Initiation Protocol (SIP) protocol is commonly used to establish Voice over IP (VoIP) calls. IETF SIP standards do not specify a secure authentication process thus allowing malicious parties to impersonate other parties or to charge calls to other parties. This paper proposes an extension to the SIP protocol that uses an identity-based authentication mechanism and key agreement protocol. These extensions provide stronger cryptographic assurances for VoIP authentication and enable provably secure key agreement between users. The use of ID based cryptography means that a large Public Key Infrastructure (PKI) is not required thus making this protocol viable for large scale implementation.

1 Introduction

Voice Over IP (VoIP) is growing dramatically in Australia and worldwide, e.g., a West Australian based ISP signed up 10,000 customers within 3 months of their service going live [14]. The term VoIP is a generic term given to any technology that enables voice communication over the Internet, there are many competing technologies beyond those standardised by the IETF. Some of these other technologies are Skype [21], H.323 [23], and voice aware IM software.

The most commonly used protocol is the combination of SIP for signalling and RTP for transmitting voice and/or video. These protocols are standardised by the IETF and perform their required tasks efficiently. However, due to the lack of basic security features no fundamental proof of identity, no protection from man in the middle or replay attacks, and no assurance of privacy is provided. These can manifest themselves as caller ID spoofing attacks, fraudulent billing, and eavesdropping (with the ability to replay) either the signalling messages or the media itself.

This paper introduces a new authentication mechanism and key agreement protocol for SIP using ID-based cryptography that will provide cryptographic assurances to VoIP communication. These assurances are a step towards addressing the lack of security mentioned above. The new key agreement protocol utilises the modified protocol 3 (with a tighter security definition) of Chen & Kudla [7].

1.1 VoIP: SIP and RTP

IETF standardised VoIP is a combination of four predominate standards: RFC3261: Session Initiation Protocol (SIP) [18], RFC2327: Session Description Protocol (SDP) [12], RFC3550: Real-time Transport Protocol (RTP) [19], and later RFC3711: The Secure Real-time Transport Protocol (SRTP) [3].

SIP is a text based protocol with similar formatting to HTTP capable of operating on TCP or UDP and handles all the signalling requirements of a VoIP session, it is analogous to the SS7 [22] protocol in traditional telephony. The role of SIP is to establish streaming connection between hosts using two primary messages exchanges; `INVITE` consisting of a four way handshake (INVITE, RINGING, OK, and ACK) and `REGISTER` consisting of (REGISTER, Unauthorised, and OK). The UMTS standard [11] uses a modified version of SIP to enable multimedia calls between users on 3G networks.

SDP is a descriptive language used to describe the attributes of a media session being established or reconfigured. SDP messages are attached to the `INVITE` and `OK` messages during a SIP call establishment. The message is made up of a number of key value pairs called attributes. These attributes include what codecs are available and the IP addresses and port numbers of stream endpoints.

RTP on the other hand is a UDP based streaming protocol capable of using arbitrary profiles and parameters. It handles buffering, jitter correction and is reliant upon SIP to know which profile and codecs to use and which ports to utilise for the media stream. SRTP is a later extension to RTP which provides cryptographic support for privacy and integrity.

Unfortunately, SIP and RTP both lack basic security features which have consequences as negligible as receiving nuisance calls, up to the complete loss of privacy. This lack of security features fail to provide users and customers with the assurances they have come to expect from communication using the traditional PSTN. With VoIP becoming more and more widespread and with some end consumers now using VoIP exclusively, a provably secure authentication and key agreement mechanism is required for SIP.

1.2 Research Problem

SIP authentication typically uses HTTP Digest [10] authentication. However, digest authentication is vulnerable to many forms of attack as noted in RFC2617 (HTTP Authentication: Basic and Digest Access Authentication), the RFC even refers to Digest as weak, but an improvement over Basic authentication. It is unfortunate that the IETF adopted Digest authentication as the default authentication mechanism, although it is possible this was done intentionally to keep SIP as HTTP-like as possible, which was a stated design goal [18].

SIP allows sections of the messages to be encrypted using S/MIME, however S/MIME is dependent upon a Certificate Authority (CA) and accompanying Public Key Infrastructure (PKI), and therefore limited by the adoption of such a system. Also, it is possible that S/MIME is likely to be too heavy for resource constrained handsets.

Secure RTP (SRTP) and the associated Secure Real Time Control Protocol (SRTCP) provide cryptographic and integrity checks to the media stream through the use of the Advanced Encryption System (AES) in Counter Mode (CM). However, the master key that is required by SRTP has no means of being established between two previously unknown parties. While SIP allows keying material to be sent to the called party by way of an attribute in the SDP, this is sent in the clear and therefore does not provide an assurance that a call will be secure from eavesdropping. The IETF is in the process of standardising MIKEY (Multimedia Internet KEYing) however this is dependent upon a working, globally available PKI. Additionally, SIP does not provide a mechanism to authenticate each party in a SIP establishment without the use of a global PKI, which opens SIP to Caller-ID impersonation and spoofing attacks.

These core problems stem from the fact that the default authentication mechanism is Digest, which does not perform mutual authentication (introducing Man-in-the-Middle and session hijack attacks), and the fact that the available means for providing cryptographic protection require a PKI and related CA's. Even with a deployed PKI, there currently exists no means for the establishment of a shared session key to provide cryptographic protection for the media session.

1.3 Our Solution

This paper proposes the use of ID-based cryptography [20] as a solution to the authentication and key agreement problems that exist in SIP. This new SIP authentication mechanism and key agreement protocol provides mutual authentication and provably secure key agreement between previously unknown parties. ID based cryptographic schemes are a current and active area of research by the cryptographic community and offer the benefits of public key cryptography without the need for an expansive PKI. This solution fits neatly in the SIP protocols as described in RFC3261 and is intended to be operable in the case of intermediary proxies being unaware of the additions.

1.4 Outline of Paper

Section 2 presents some background information on the SIP standard, RFC3261, and how authentication and key agreement are currently performed. In Section 3, ID-based cryptography is briefly introduced, followed by a presentation of the improved provably secure protocol 3 of Chen & Kudla [7], which the new SIP key agreement protocol utilises. In Section 4 the new SIP authentication mechanism and key agreement protocol is shown. This is followed by a discussion of the proposed solution and observed limitations in Section 5. Finally, Section 6 concludes with an exploration of possible avenues for future work.

2 Current SIP Authentication and Key Agreement

As mentioned previously SIP is a text based protocol similar in formatting to HTTP, the two most commonly used SIP exchanges are the `REGISTER` and `INVITE` exchanges, respectively used to connect to the network, and establish a call.

2.1 Authentication

SIP has two authentication dialogs: 401 - Unauthorized and 407 - Proxy Authentication Required. 401 responses are mainly used during `REGISTER`, while 407 responses are used during call establishment with intermediary SIP proxies (predominately during `INVITE`). SIP presently uses HTTP Digest authentication, with an option to use certificates and a PKI.

Digest authentication is a challenge response authentication scheme that ensures the password is never sent in cleartext, and also provides protection against replay attacks through the use of a nonce. Digest however, does not provide mutual authentication. Furthermore, Digest depends upon a preestablished relationship between the requester and responder.

2.2 Key Agreement

As mentioned previously, SRTP can be used to provide confidentiality and integrity for the media stream, but lacks a mechanism for establishing a master session key. SRTP requires keying information to be present on both hosts prior to the session being established. Keys can be sent as part of the SDP message in the `INVITE` sequence however this information is sent in the clear. Protecting the SIP exchanges by use of TLS or S/MIME is suggested, however these options are limited by the necessity of a global PKI.

Another alternative for providing confidentiality and integrity for both SIP and RTP is the use of IPSec [15] tunnels between endpoints and using IPSec's associated key agreement protocols (IKEv1 [13], IKEv2 [9]). However, IPSec has a heavy establishment cost which impacts on the agility of a mobile device in a heterogeneous wireless environment.

3 Identity-based Cryptography

ID-based cryptosystems were first introduced by Shamir in 1984 [20]. The advantage of these systems is in the easy derivation of an entity's public key – a function of the entity's identity. The entity's private key can then be calculated by a Trusted Authority (TA), also known as a Private Key Generator (PKG). TA services would commonly be provided by a users service provider, employer, or institution, or any body with which the user would have an out-of-band relationship with. As pointed out in a recent survey [6], the ID-based public key cryptosystem is viewed as an alternative for certificate-based PKI as it greatly reduces the problems with key management on a per individual basis (the obstacle faced in PKI).

In recent years, the potential application of elliptic curve pairings to realise cryptographic structures has resulted in renewed interest in the use of ID-based cryptography to solve the problem of constructing non-interactive key distribution schemes [6]. To illustrate how elliptic curve pairings can be used to build novel cryptographic schemes with interesting properties, including ID-based key establishment protocols, the reader is referred to the pairing-based crypto lounge [1].

3.1 Bilinear Maps from Elliptic Curve Pairings

In this section, the mathematical preliminaries required to understand the ID-based protocol presented in the next section are introduced. Using the notation of Boneh & Franklin [5], let $\mathbb{G}_1$ be an additive group of prime order q and $\mathbb{G}_2$ be a multiplicative group of the same order q. Assume the existence of a map $\hat{e}$ from $\mathbb{G}_1 \times \mathbb{G}_1$ to $\mathbb{G}_2$. Typically, $\mathbb{G}_1$ will be a subgroup of the group of points on an elliptic curve over a finite field, $\mathbb{G}_2$ will be a subgroup of the multiplicative group of a related finite field and the map $\hat{e}$ will be derived from either the Weil or Tate pairing on the elliptic curve. The mapping $\hat{e}$ must be efficiently computable and has the following properties.

Bilinearity. For $Q, W, Z \in \mathbb{G}_1$, both

$$\hat{e}(Q, W+Z) = \hat{e}(Q,W) \cdot \hat{e}(Q,Z) \quad \text{and} \quad \hat{e}(Q+W, Z) = \hat{e}(Q,Z) \cdot \hat{e}(W,Z).$$

Non-Degeneracy. For some elements $P, Q \in \mathbb{G}_1$, such that $\hat{e}(P,Q) \neq 1_{\mathbb{G}_2}$.

Computability. For some elements $P, Q \in \mathbb{G}_1$, an efficient algorithm exists to compute $\hat{e}(P,Q)$.

A bilinear map, $\hat{e}$, is said to be an *admissible* bilinear map if it satisfies all three properties. Since $\hat{e}$ is bilinear, the map $\hat{e}$ is also symmetric.

3.2 An Improved Provably-Secure ID-based Protocol

This paper employs the provably-secure protocol of Chen & Kudla [7] proven secure in a restrictive adversarial model of Bellare & Rogaway (hereafter referred to as the BR93 model) [4]. In the BR93 model, there exist an all-powerful adversary, $\mathcal{A}$, which is in control of all the communications that take place between all protocol principals. $\mathcal{A}$ does this by interacting with a set of *oracles*, each of which represents an instance of a principal in a specific protocol run. Each principal has an identifier U and oracle Π_U^s represents the actions of principal U in the protocol run indexed by integer s.

Interactions with the adversary are called oracle *queries*. Each interaction is now described informally.

$\mathsf{Send}(U, s, m)$ This query allows $\mathcal{A}$ to make the principal U run the protocol normally. The oracle Π_U^s will return to $\mathcal{A}$ the same next message that an honest principal U would if sent message m according to the conversation so far.

$\mathsf{Reveal}(U, s)$ If a session key K_s has previously been accepted by Π_U^s then it is returned to $\mathcal{A}$. An oracle is called *opened* if it has been the object of a Reveal query.

$\mathsf{Corrupt}(U, K)$ The query returns the oracle's internal state and sets the long-term key of U to be the value K chosen by $\mathcal{A}$. $\mathcal{A}$ can then control the behaviour of U with Send queries. A principal is called *corrupted* if it has been the object of a Corrupt query.

Protocol 3 of Chen & Kudla [7] was proven secure under the restriction that the adversary is not allowed to make any Reveal query. However, such a restriction means that the protocol is not secure against known (session) key attacks[1] [8].

A revised (provably-secure) protocol in the "full" BR93 model (i.e., without restricting the adversary from asking the Reveal queries) is obtained by

- using recent results of [8, 16], a small change to the way that session keys are constructed[2] in the ID-based protocol of Chen & Kudla is made, and
- deploying the GAP assumption introduced by Okamoto & Pointcheval [17].

The revised protocol is described in Figure 1. In the protocol, there are two trusted authorities, say TA_1 and TA_2, which have public/private key pairs ($s_1P = P_{Pub(TA_1)} \in \mathbb{G}_1, s_1 \in \mathbb{Z}_q$) and ($s_2P = P_{Pub(TA_2)} \in \mathbb{G}_1, s_2 \in \mathbb{Z}_q$) respectively, where P, $\mathbb{G}_1$ and $\mathbb{G}_2$ are globally agreed (e.g., recommended by an international standards body). There are two entities in the protocol, namely an initiator, A, and a responder, B. A registers with TA_1 and gets her private key $S_A = s_1Q_A$ and B registers with a different TA, TA_2 and gets his private key $S_B = s_2Q_B$, where $Q_A = \mathcal{H}(ID_A)$, $Q_B = \mathcal{H}(ID_B)$, and $\mathcal{H}$ denote some secure cryptographic hash function.

The protocol begins by having A randomly select a random challenge, $a \in_R \mathbb{Z}_q^*$, compute the ephemeral public keys, $T_A = aP$, and then send it to B with whom she desires to communicate. Upon receiving the message T_A from A, B also randomly selects a random challenge, $b \in_R \mathbb{Z}_q^*$, computes the ephemeral public keys, $T_B = bP$, and sends T_B back to A. Note that both a and b are the ephemeral private keys of A and B respectively. At the end of the protocol execution, both A and B compute a shared session key SK_{AB} and SK_{BA} independently. Let $||$ denote the concatenation of messages; $\mathcal{H}_1$ denote some secure cryptographic hash function independent of $\mathcal{H}$; and $\mathcal{T}r_A$ and $\mathcal{T}r_B$ denote a record of the transcript of messages sent and received by both parties, which should be identical in the presence of a benign adversary.

4 A New Authentication and Key Agreement Protocol for SIP

The previous section introduced elliptic curve cryptography and the application of curve pairings to an ID-based cryptosystem. In particular the improved Chen & Kudla [7] protocol 3 key agreement is described which is utilised in the new SIP key agreement proposed below. This section describes two independent extensions to the SIP protocol, the first being an ID-based authentication mechanism for use by the two authentication

[1] We advocate that such an attack is realistic in a real world setting as it is normal to assume that a host can establish several concurrent sessions with one or more different parties. Sessions are specific to both the communicating parties. In the case of key establishment protocols, sessions are specific to both the initiator and the responder principals, where every session is associated with a unique session key. Therefore, learning session keys from any session different from the one under attack should not enable the adversary to learn anything about the session key associated with the session under attack.

[2] We note that such a key derivation function method is recently included in the special publication by NIST (SP 800-56A – Recommendation for Pair-Wise Key Establishment Schemes Using Discrete Logarithm Cryptography) (`http://csrc.nist.gov/publications/nistpubs/index.html`), March 2006.

A		B
$a \in_R \mathbb{Z}_q^*$	$\xrightarrow{T_A = aP \in \mathbb{G}_1}$	$b \in_R \mathbb{Z}_q^*$
$K_{AB} = \hat{e}(S_A, T_B) \cdot \hat{e}(Q_B, as_2P)$	$\xleftarrow{T_B = bP \in \mathbb{G}_1}$	$K_{BA} = \hat{e}(S_B, T_A) \cdot \hat{e}(Q_A, bs_1P)$
$\mathcal{T}r_A = T_A \| T_B$		$\mathcal{T}r_B = T_A \| T_B$
$SK_{AB} = \mathcal{H}_1(A\|B\|\mathcal{T}r_A\|K_{AB})$		$SK_{BA} = \mathcal{H}_1(A\|B\|\mathcal{T}r_B\|K_{BA})$
	$SK_{AB} = SK_{BA}$	

Figure 1: The improved provably-secure ID-based protocol 3 of Chen & Kudla

types in SIP. The second is an extension to the SDP messaging format to cater for an ID-based key agreement protocol implementing the improved Chen & Kudla [7] protocol 3 shown in Figure 1.

The goals of the extended SIP authentication and key agreement protocols were to achieve an assurance of privacy that was equal to or better than what could be expected from the traditional PSTN. More specifically, the extensions need to provide protection from Caller-ID spoofing, resistance to Man-in-the-Middle and session hijack attacks, non-repudiation, proof of identity and protection from eavesdropping. To support these goals, the extension will also require mutual authentication, protection from replay attacks as well as a key escrow facility for law enforcement.

Using ID-based cryptography allows the use of the user's SIP identity (user@domain) as the user's public key (Q), when hashed with some designated hash function, $\mathcal{H}$. The service provider ensures that the client is correctly identified using a mechanism such as the 100 points ID check and then issue the corresponding private key to the user. The TA's public key ($P_{Pub_{TA}}$) and the users public/private key pair (Q and S) are stored securely by the user, either on the device or some tamper evident storage mechanism such as the widely used SIM card in current mobile telephony.

4.1 A New Authentication Mechanism for SIP

Alice		Server
	401: Unauthorised, Sign(nonce,realm,opaque,time,Alice) $\longleftarrow$	
	Authorization, Sign(nonce,realm,opaque,Alice,response) $\longrightarrow$	

Figure 2: Example ID-based authentication

SIP has two types of authentication distinguished by the response codes 401 and 407. These are generally required during an `INVITE` or `REGISTER` handshake and refer to service authentication and proxy authentication respectively. The method of authentication is the same in each case and the SIP extension for authentication provides that same functionality.

The new SIP authentication mechanism proposed here provides mutual authentication, prevention from replay attacks, and non-repudiation while operating within the semantics required by RFC2617 HTTP Authentication.

4.1.1 Assumptions

The correct and secure operation of this protocol depends upon a number of realistic assumptions. It is assumed that all parties have globally agreed upon elliptic curve parameters: P, $\mathbb{G}_1$, and $\mathbb{G}_2$, such as would be defined by an appropriate standards body. Furthermore it is assumed that all private keys remain private and secure, and that nonces will be a function of the realm and time, and will not be reused. Lastly, it is assumed that the TA's of security domains can be trusted to provide a true and accurate account of subscribers, and that TA's are accessible by all entities in the system. Informally, the system remains secure in the event of a malicious TA provided that TA's do not collude.

4.1.2 Description

The proposed ID-based authentication handshake is similar to the challenge response handshake of Digest, and as stated can operate without changing the semantics of RFC2617 HTTP Authentication. The handshake is described here:

1. Client makes request of a SIP service requiring authentication (eg `REGISTER` or `INVITE`).

2. Server responds with `401 Unauthorized` or `407 Proxy Authentication Required` as appropriate. This response is a challenge consisting of: a realm string, nonce, opaque string, current time, and claimed identity of the Client, signed with that security domains long term private key, s_j. The realm string is a human readable identifier of the security domain, such as the name of the service provider or network. An opaque string is used as a session identifier.

3. Client verifies the server response against the signature using $P_{Pub_{TA}}$, and calculates the nonce using the supplied time and realm and compares with the nonce sent from the Server. The addition of the clients claimed identity in the server response protects against a man in the middle attack. Requiring the client to generate the nonce and compare it with the server supplied nonce allows the client to protect against nonce reuse as the client can disallow a nonce generated from a time outside a preconfigured threshold.

4. Client prepares a response consisting of a secure hash of the username, realm, nonce and opaque string signed with their private key, S_i. The Client sends the signed response along with their username, realm, nonce and opaque string in cleartext to the entity requesting authentication.

5. Server verifies the response with Q_i respectively. Again, due to the nature of identity based encryption, Q_i is a function of the clients identity (ie their SIP identity).

6. Server responds with appropriate error message or grants access.

```
WWW-Authenticate: IdentityBased
  realm = "qut.com",
  nonce = "dcd98b7102dd2f0e8b11d0f600bfb0c093",
  opaque = "5ccc069c403ebaf9f0171e9517f40e41",
  time = "12-12-2005 03:24:12",
  signature = "base64(s(realm|nonce|opaque|time|ClientID,s_j)"
```

Figure 3: ID-based Authentication Challenge

```
Authorization: IdentityBased
  username = "client@qut.com",
  realm = "qut.com",
  nonce = "dcd98b7102dd2f0e8b11d0f600bfb0c093",
  opaque = "5ccc069c403ebaf9f0171e9517f40e41",
  response = "base64(e(username|realm|nonce|opaque,S_i)"
```

Figure 4: ID-based Authentication Response

4.1.3 Changes to Protocol

A new HTTP authentication dialogue will be introduced, a simplified example is seen in Figure 2. Note, for the sake of illustration, not all information is displayed in the diagram. A full ID-based authentication challenge-response is shown in Figures 3 and 4 respectively.

An attacker observing the handshake will be able to replay the server challenge to the client however a client should ignore nonces that have been used previously for this security domain. Also, an attacker could not impersonate a server by issuing a new challenge as an attacker would not have knowledge of the private key, s_j. Therefore, the attacker could not issue a correctly signed challenge. Furthermore, an attacker observing the client response will be able to decrypt the response, however this does not gain the attacker any advantage due to the asymmetric nature of ID-based cryptography.

4.2 A New Key Agreement Protocol for SIP

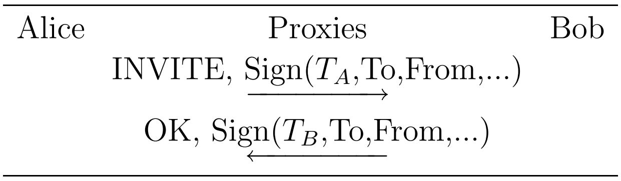

Figure 5: Example ID-based Key Agreement

Support for cryptographic protection of the VoIP media session is possible using SRTP but SRTP requires a preestablished shared secret. A key agreement protocol is required to establish this shared secret, however no such facility exists within SIP. Additionally, ID-based cryptography allows a key agreement to occur where the only knowledge required is the entities public identity, in this case, their SIP identity.

This new key agreement protocol for SIP uses the provably secure implementation of the improved CK key agreement protocol 3 and provides implicit authentication of both parties at the end of the handshake, an ability to provide key escrow for law enforcement and a non-repudiation quality, whilst ensuring that the operation remains consistent with the requirements of RFC3261: Session Initiation Protocol.

4.2.1 Additional Assumptions

In addition to the assumptions outlined in Section 4.1.1 it is assumed that each entity has a secure means of randomly choosing a and b.

4.2.2 Description

Every SIP subscriber is identifiable by their SIP address, eg alice@a.com, and that their public key is a secure hash of that identity. To describe the implementation of the modified Chen–Kudla protocol using SIP messaging, a worked example is explained here to identify the messages being sent. This example consists of two users, Alice and Bob, who are members of separate security domains, with SIP identities alice@a.com and bob@b.com respectively. Each security domain has a proxy server through which all SIP messaging will pass, and a Trusted Authority (TA) which issues ID-based private keys. Furthermore, a Registrar server will also exist within the security domain to provide `REGISTER` services, it is likely that the Proxies and Registrar will share a common TA so that the same credentials are used by both services.

Our example will involve Alice making a call to Bob, each with no prior knowledge of the other, other than Alice knowing Bob's SIP identity. For simplicity, Alice will not authenticate to her security domains proxy, although in a real implementation the authentication steps explained in the preceding section would occur.

Step One Alice chooses ephemeral private key a and calculates $T_A = aP$, which is her contribution to the exchange. Alice then prepares a standard SIP `INVITE` message and adds T_A to the SDP message using a third party attribute. To ensure integrity of the message, Alice then computes a signature over the entire message (sans Request-URI, Via, Record-Route, Route, Max-Forwards, and Proxy-Authorization) using S_A, and places this at the end of the SDP message. This message is then sent to Alice's configured Proxy for routing to Bob's security domain.

Step Two Bob receives the `INVITE` from his Proxy, and verifies the integrity of the message after calculating Q_A from Alice's identity located in the `From:` header and then sends the appropriate `RINGING` message to Alice via appropriate proxies. It is possible that an adversary M could change the `From:` header and recompute the signature using a new private key S_M however the change would be apparent to Bob in that Bob would be under the impression he was talking to M not Alice. Since the perceived partner of Bob is M and not Alice, therefore, the session keys established by both Alice and Bob at the end of the protocol execution will not be the same (recall that the keying materials of the session key comprises identities of the perceived partners)

Step Three If Bob accepts the call, he chooses an ephemeral private key b and calculates $T_B = bP$ which he sends to Alice in the `OK` message in the same way as Alice sent T_A in Step One, signed the same way as in Step One.

Step Four Alice receives Bob's `OK` message and verifies the integrity of the message after calculating Q_B from Bob's SIP identity. If the signature is correct, Alice responds to Bob with an `ACK` message. This signals that both parties are in agreement to begin the key establishment, which is the computationally expensive part of the protocol.

Step Five Alice can now compute

$$\begin{aligned} K_{AB} &= \hat{e}(S_A, T_B)\hat{e}(Q_B, aP_{Pub(TA_2)}) \\ MK_A &= H(A||B||\mathcal{T}r_A||K_{AB}) \end{aligned}$$

and Bob can compute

$$\begin{aligned} K_{BA} &= \hat{e}(S_B, T_A)\hat{e}(Q_A, bP_{Pub(TA_1)}) \\ MK_B &= H(A||B||\mathcal{T}r_B||K_{BA}) \\ &= MK_A \end{aligned}$$

so therefore

$$MK_B = MK_A = MK$$

Note that $\mathcal{T}r_A$ and $\mathcal{T}r_B$ denoted a concatenation of the messages sent between the two parties. When this was introduced earlier in Figure 1, these messages were only T_A and T_B, however, our implementation uses a concatenation of all the messages that were sent and received during the handshake (with the exception of the headers that will be changed by intermediary proxies). This creates an implicit integrity check and provides additional assurances that the messages have not been tampered with or that a session hijacking has occured.

Finally both parties can now utilise MK with the standard SRTP protocol to derive the four keys necessary to encrypt and integrity check the media session and control session, after which encrypted streaming can begin.

4.2.3 Changes to Protocol

A simplified `INVITE` handshake is shown in Figure 5, note that for the sake of illustration, that some messages have been removed that are not relevant to the key agreement process. Inserting two additional attributes into the SDP message facilitates the transmission of each parties ephemeral keys T_A and T_B, and to allow transport of the necessary signatures. Following is an example of the two additional attributes:

```
a=idaka: base64(TA or TB)
a=signature: base64(signature)
```

As RFC2327 defines the `a` attribute for use in describing session attributes, as well as allowing zero or more instances of that attribute, it was a reasonable choice to carry key agreement messages. However, it is possible that a new attribute type could be introduced to support authenticated key agreement, but doing so would not change the algorithm, just the formatting within the message.

5 Discussion

The new ID-based SIP authentication mechanism and key agreement protocol proposed here meets the goal and requirements stated above. The cryptographic primitives used to provide the assurances are provably secure in the adversarial model of the BR93 model which assumes an all-powerful adversary that is in control of all communication links with the ability to manipulate any message sent and impersonate any other party. That is, the adversary is capable of performing Man-in-the-Middle and session hijacking attacks (i.e., known key security), as well as spoofing and identity-based attacks. Since the proposed protocol is proven secure in the BR93 model, it is capable of withstanding these attacks.

The new authentication mechanism proposed in this paper introduces a new dialog for authentication, but only introduces one additional field into the existing authentication message, namely, the field to hold a signature. The additional changes involve how existing fields are interpreted by the parties involved and remain consistent in form with the appropriate RFC. Furthermore, the new key agreement protocol recommended in this paper does not introduce any new fields into the SDP message but makes use of existing fields in such a manner that is still within the bounds of their scope according to the RFC. It is expected that both of these changes will still be functional in the event that intermediary proxies have no understanding of the new additions.

Non-repudiation, protection against replay and session hijacking attacks, and mutual authentication are by-products from the use of ID-based cryptography and signatures. Provided that the assumptions of the private keys remain secure hold true, then a high level of assurance can be provided to the involved parties that all calls are genuine, and that all parties are truthfully represented. An additional benefit that the new key agreement protocol provides is a key escrow mechanism for use by law enforcement with the appropriate procedural oversight at the expense of users' privacy. Law enforcement can obtain the derived session keys by observing the key agreement handshake and "colluding" with the relevant TA's to compute the master key.

It has been shown that SIP lacks basic security constructs and the new authentication mechanism and key agreement protocol introduced in this paper is a step towards addressing these concerns. However, a number of limitations are apparent in the use of ID-based cryptography to provide this solution.

5.1 Limitations

Firstly, this new scheme is suited to a security domain environment, such as would exist within a company, institution, or customers with a service provider. It is not suited to provide protection for peer-to-peer type connections as would be experienced if Alice was to contact Bob directly, irrespective of their various security domains. This is because no Trusted Authority exists to issue private keys. Because SIP is required to work in a P2P manner this restriction means only a subset of SIP's functionality is possible. However, real world implementations of SIP are unlikely to use the P2P method of operation beyond casual communication amongst parties who are likely to have some out of band relationship, which could be used to distribute keying material securely. Also, related to this limitation is the support for conference calling. At present the use of the new solution when more than two parties are present in a call has not been investigated. This is an area for future work.

As explained in 4.1.1 the system relies on the assumption that TA's are trustworthy and would not collude. As a TA is incorporated within the entity that provides SIP services to a user it is reasonable to assume an existing trust relationship with that service provider. Nothing exists to prevent TA's colluding within the protocol, but such an action would have economic and legal ramifications to the TA. It was stated earlier that law enforcement with the appropriate oversight can compel a TA's to "'wilfully collude"' so that a wiretap can occur, but this would occur within an appropriate legislative framework. It could be argued that a communication system that prevents legitimate wiretaps is preferable, but such a system while technically possible would not be publicly viable due to the wiretap requirements of the various Telecommunications Acts.

Another limitation is VoIP to PSTN calls; that is calls, that originate within a VoIP network but are to be terminated in the PSTN. Calls of this nature (and PSTN to VoIP calls for that matter) make use of Media Gateways. There is a potential vulnerability in the new scheme when access to the PSTN is required as the authentication and key agreement is being performed by the Media Gateway and not the PSTN user. Therefore, no assurance of identity can be given to either party. It is likely that traditional telephony users will have a relationship with a PSTN service provider that provides access to a Media Gateway. Assurance can then be given to the VoIP user to the fact that there is a contractually binding agreement between a PSTN address (+61730005000@phonecompany.com.au) and the owner of that phone number, however, the security of such an assurance depends upon the assumption the phone company will act honestly.

5.2 Performance

The assurances provided by the new authentication mechanism and key agreement protocol do have a performance cost. Even though elliptic curve cryptosystems have performance benefits rising from a reduced key size when compared with RSA-based schemes, the operations required by ID-based elliptic curve cryptography has a significant performance impact as explained below.

The proposed solution requires the use of three different elliptic curve operations; addition, multiplication, and pairings. In addition, a secure hash function, and the ability to randomly choose points from an appropriate finite field are also required. Hashing, and choosing points are computationally inexpensive operations, point addition is somewhat more expensive than hashing, but still relatively inexpensive. Point multiplication is more expensive again (as its based upon point addition), and finally, pairings are considerably expensive and should be avoided where possible due to the fact one pairing requires at least ten times more multiplications in the underlying finite field than one multiplication [2]. Unfortunately, elliptic curve pairings are what makes ID-based cryptography possible and cannot be avoided by an ID-based cryptosystem.

Entity authentication involves two elliptic curve pairings and two hashes in addition to the operations required by RFC3261. As the protocol is symmetric the cost is the same to both initiator and responder.

The key agreement process performed during the `INVITE` sequence has two distinct phases, the first phase ensures each party possesses the necessary information while the second phase performs the actual key agreement. During establishment the caller and callee are both required to choose a point on the curve, perform one point multiplication, two hashes, and two curve pairings. It should be noted that the hashes and pairings are required for the message signatures, and not necessary for the key agreement itself. The final calculation to calculate the shared key requires two additional curve pairings, two multiplications, and one hash to be performed by both the caller and callee.

All operations are symmetric, however in practical implementations it is most likely that the caller will need to perform at least one authentication during a call establishment, therefore making the real world protocol slightly asymmetric with the caller being required to perform an additional two curve pairings and two hashes.

While the operations to be performed are extensive, it is only performed during call establishment. Users will notice a delay from the call being answered and media streaming beginning. Once the media stream begins however no call degradation will occur beyond what is already experienced in VoIP telephony.

6 Conclusions

Introducing current state of the art cryptographic research in ID-based cryptosystems to the authentication and key agreement problems that exist in SIP has been discussed. This solution requires only minimal changes to the standard and achieve a much higher standard of security and assurance for users and service providers alike and utilises the provably secure improved ID-based protocol of Chen & Kudla. The new authentication mechanism and key agreement protocol presented here address the lack of secure authentication support in SIP as well as the lack of a key agreement protocol by providing a provably secure solution using ID-based cryptography.

A number of avenues for future work have also been identified. Exploring avenues to increase the performance of the protocol is an important piece of work yet to be done, as the work presented here has only just begun to explore the possibilities. Three possible solutions are: improved algorithms to perform elliptic curve operations, hardware assisted calculations, and the use of a third party to perform the calculations during protocol runs.

It is essential that an implementation of the proposed protocol be developed. This implementation can be used to obtain comparative performance metrics on a variety of hardware and platforms can be observed. This is an immediate area of future work and its acknowledged that such observations will be crucial in analysing the utility of the ID-based authentication and key agreement protocols. It is evident that the operation of this protocol in its entirety would be considerably processor intensive and is likely to be unsuitable for mobile handsets at this point in time due to the expense of calculating curve pairings. However, hardware designed and optimised for performing elliptic curve operations has been developed and the adoption of this by handset manufacturers will mitigate against these issues. An ideal solution would be offloading the expensive curve pairings to a high powered third party without affecting the security of the protocol.

Another aspect of future work is further extending the protocol to support roaming environments as would be experienced by a wireless device. This would involve multiple security domains and trust relationships as well as requiring a secure method for remuneration and billing between service providers who provide network access to customers of other providers.

Acknowledgements

The authors would like to thank all five anonymous referees for their invaluable and insightful feedback.

References

[1] P. S. L. M. Barreto. The Pairing-Based Crypto Lounge, 2005. `http://paginas.terra.com.br/informatica/paulobarreto/pblounge.html`.

[2] P. S. L. M. Barreto, H. Y. Kim, B. Lynn, and M. Scott. Efficient Algorithms for Pairing-based Cryptosystems. In *Crypto 2002*, pages 354–368. Springer-Verlag, 2002. Vol. 2442/2002 of LNCS.

[3] M Baugher, D McGew, M Nasland, E Carrara, and K Norman. The Secure Real-time Transport Protocol (SRTP). Request For Comments 3711, Internet Engineering Task Force, March 2004.

[4] M. Bellare and P. Rogaway. Entity Authentication and Key Distribution. In *Crypto 1993*, pages 110–125. Springer-Verlag, 1993. Vol. 773/1993 of LNCS.

[5] D. Boneh and M. Franklin. Identity-Based Encryption from the Weil Pairing. *SIAM Journal on Computing*, 32(3):585–615, 2003.

[6] C. Boyd and K.-K. R. Choo. Security of Two-Party Identity-Based Key Agreement. In *Mycrypt 2005*, pages 229–243. Springer-Verlag, 2005. Vol. 3715/2005 of LNCS.

[7] L. Chen and C. Kudla. Identity Based Authenticated Key Agreement Protocols from Pairings (Corrected version at `http://eprint.iacr.org/2002/184/`). In *CSFW 2003*, pages 219–233. IEEE Computer Society Press, 2003.

[8] K.-K. R. Choo, C. Boyd, and Y. Hitchcock. On Session Key Construction in Provably Secure Protocols (Extended version available from `http://eprint.iacr.org/2005/206`). In *Mycrypt 2005*, pages 116–131. Springer-Verlag, 2005. Volume 3715/2005 of LNCS.

[9] C. Kaufman (Ed). Internet Key Exchange (IKEv2) Protocol. Request for Comments RFC4306, Internet Engineering Task Force, December 2005.

[10] J Franks, P Hallam-Baker, J Hostetler, S Lawrence, P Leach, A Luotonen, and L Stewart. HTTP Authentication: Basic and Digest Access Authentication. Request for Comments 2617, Internet Engineering Task Force, 1999.

[11] M Garcia-Martin. Input 3rd Generation Partnership Project (3GPP) Release 5 Requirements on the Session Initiation Protocol (SIP). Informational RFC RFC4083, Internet Engineering Task Force, May 2005.

[12] M Handley and V Jacobson. Session Description Protocol. Request for Comments RFC2327, Internet Engineering Task Force, April 1998.

[13] D. Harkins and D. Carrel. The Internet Key Exchange (IKE). RFC RFC2409, Internet Engineering Task Force, November 1998.

[14] iiNet Ltd. iiNet Surpasses 10,000 VoIP Subscribers. Press Release, November 2005. Available at `http://www.iinet.net.au/about/investor/voip\_10000.pdf`.

[15] S. Kent and K. Seo. Security Architecture for the Internet Protocol. RFC RFC4301, Internet Engineering Task Force, December 2005.

[16] C. Kudla and K. G. Paterson. Modular Security Proofs for Key Agreement Protocols. In *Asiacrypt 2005*, pages 549–565. Springer-Verlag, 2005. Vol. 3788/2005 of LNCS.

[17] T. Okamoto and D. Pointcheval. The Gap-Problems: a New Class of Problems for the Security of Cryptographic Schemes. In *PKC 2001*, pages 104–118. Springer-Verlag, 2001. Vol. 1992/2001 of LNCS.

[18] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M Handley, and E. Schooler. SIP: Session Initiation Protocol. Request for Comments 3261, Internet Engineering Task Force, June 2002.

[19] H Schulzrinne, S Casner, R Frederick, and V Jacobson. RTP: A Transport Protocol for Real-Time Applications. Request for Comments RFC3550, Internet Engineering Task Force, July 2003.

[20] A. Shamir. Identity-Based Cryptosystems and Signature Schemes. In *Crypto 1984*, pages 47–53. Springer-Verlag, 1984. Vol. 196/1985 of LNCS.

[21] Skype Technologies. Skype Website. `http://www.skype.com/` [Last Accessed: 12-01-2006].

[22] International Telecommunications Union. ITU-T Recommendation Q.700: Introduction to CCITT Signalling System 7. Recommendation Q.700, International Telecommunications Union, March 1993.

[23] International Telecommunications Union. ITU-T Recommendation H.323: Packet-based Multimedia Communications Systems Version 5. ITU-T Recommendation H.323, International Telecommunications Union, July 2003.