Table of Contents

Transport Layer Security Protocol For Internet Of Things

Illya Gerasymchuk
illya@iluxonchik.me

Instituto Superior Tcnico
Supervisors: Ricardo Chaves, Aleksandar Ilic

Abstract. Transport Layer Security (TLS) is, by far, the most used communication security protocol, it is however, not suitable in the context of Internet Of Things (IoT). The resource-limited nature of a big part of IoT devices does not allow for the use of computationally complex and memory demanding operations present in a standard TLS implementation. Most of previous work focused entirely on Datagram TLS (DTLS) and can not be easily integrated with existing deployments. This work focuses on how TLS and the extension mechanism can be used to define a framework to adapt the protocol to specific needs. This is the approach that will be followed in the second part of the work. Having an adaptable and easy to use solution is crucial for its adaptation in IoT, where security might have been completely foregone otherwise.

Keywords: TLS, DTLS, SSL, IoT, cryptography, protocol, lightweight cryptography

1 Intoduction

The Internet of Things (IoT) is a network of devices, from simple sensors to smartphones and wearables which are connected together. In fact, it can be any other object that has an assigned IP address and is provided with the ability to transfer data over a network. Even a salt shaker[**?**] can now be part of the global network.

The IoT technology provides many benefits, from personal comfort to transforming entire industries, mainly due to increased connectivity and new sources for data analysis. The technological development, however, tends to focus on innovative design rather than on privacy and security. IoT devices frequently connect to networks using inadequate security and are hard to update when vulnerabilities are found.

This lack of security in the IoT ecosystem has been exploited by the the *Mirai* botnet[**?**] when it overwhelmed several high-profile targets with massive Distributed Denial-Of-Service (DDoS) attacks. This is the most devastating attack involving IoT devices done to date. However, the *Reaper* botnet[**?**] could be even worse if it is ever put to malicious use. Similar attacks will inadvertently come in the future.

TLS is one of the most used security protocols in the world, allowing two peers to communicate securely. It is designed to run on top of a reliable, connection-oriented protocol, such as TCP. Datagram TLS (DTLS) is the version of TLS that runs on top of an unreliable transport protocol, such as UDP. Most IoT devices have very limited processing

power, storage and energy. Moreover, the performance of TCP is known to be inefficient in wireless networks, due to its congestion control algorithm. This situation is worsened with the use of low-power radios and lossy links found in sensor networks. Therefore, the use of TCP with IoT is usually not the best option. For this reason, DTLS, which runs on top of UDP, is used more frequently in such devices. The work that will be done in the context of this dissertation, can however, be applied to either one of them, so even though mostly TLS will be mentioned, almost everything can also be applied to DTLS. This is a consequence of DTLS being just an adaption of TLS over unreliable transport protocols, with no changes done to the core protocol.

The problem in using (D)TLS in IoT is that it is not lightweight, since it has not been designed for such environments. An IoT device may only have 256 KB of RAM and needs to conserve the battery, while sending and receiving a large amount of small information constantly. For example, consider the case of a temperature sensor that sends temperature measures every 30 seconds to a server. In this case it just needs to send a few bytes of data and do it with minimal overhead, to conserve RAM and battery. If that sensor is going to use (D)TLS 1.2, it will need two extra roundtrips before it can send any data. This can result in an overhead of several hundreds of milliseconds. Besides that, it will need to perform heavy mathematical operations involved in cryptography, using even more energy and taking even more time. Given this, there is a clear need for a more lightweight (D)TLS for the IoT.

The goal of this work is to develop a lightweight version of (D)TLS that is fully backwards compatible and does not require any third-party entities, in order to simplify its deployment process. The solution will be developed for (D)TLS version 1.2, while also bearing in mind the new 1.3 version. The idea is to make it customizable, depending on the security requirements and the context of its usage.

In the process of the work on this dissertation, we have already made several contributions to the TLS 1.3 specification, being recognized as contributors[**?**].

The document is organized as follows: Section 2 describes the background. It introduces some of the concepts that will be used throughout the document. Section 3 describes the TLS and DTLS protocol versions 1.2 and 1.3, with a focus on the version 1.2 since it is the latest and the most used version of the protocol (version 1.3 is still in draft mode). Section 4 describes all of the related work done in the area and the current state of the art. Section 5 provides an architecture of the solution that will be developed in the second part of the dissertation, and describes how the results will be evaluated and presents a general work plan. Finally, the conclusion of the work is done in Section 6.

2 Background

TLS is a complex protocol that relies on various concepts to provide security. The most relevant ones will be described here.

In a typical scenario, TLS uses Asymmetrical Cryptography (AC) for peer authentication and Symmetrical Cryptography (SC) for bulk data encryption and integrity protection, for this reason this topic will be covered in Section **??**. Section **??** covers the most common way of peer authentication: public key certificates. Authenticated Encryption With Additional Data (AEAD) ciphers offer various advantages in the context of IoT, particularly less computational and spacial overhead. Furthermore, they are the only type

of ciphers that can be used in TLS 1.3. For those reasons, they're covered in Section **??**. When compared to other public key cryptography approaches, Elliptic Curve Cryptography (ECC) offers shorter keys, lower processing requirements and lower memory usage for equivalent security strength, being heavily used in TLS. An overview of ECC in presented in Section **??**.

2.1 Symmetric vs Asymmetric Cryptography

AC is more expensive than SC in terms of performance. There are two main reasons for this. First, larger key sizes are required for an AC system to achieve the same level of security as in a SC system. Second, CPUs are slower at performing the underlying mathematical operations involved in AC, namely exponentiation requires $O(loge)$ multiplications for an exponent e. The 2016 NIST report [**?**] suggests that an AC algorithm would need to use a secret key with size of 15360 bits to have equivalent security to a 256-bit secret key for a SC algorithm. This situation is ameliorated by ECC, which requires keys of 512 bits, but it is still slower than using SC. The 2017 BSI report [**?**] (from the German federal office for information security) suggests similar numbers.

Another argument for avoiding the use of AC algorithms as much as possible, is that they require additional storage space. This can be a problem for many IoT devices, like class 1 devices according to the terminology of constrained-code networks[**?**] which have approximately 10KB of RAM and 100KB of persistent memory. We measured and compared the resulting size of the *mbedTLS 2.6.0* library[**?**] binary when it was compiled with and without the Rivest-Shamir-Adleman (RSA) module (located in the rsa.c file). The conclusion is that that using the rsa.c module adds an overhead of about 32KB.

2.2 Public Certificates and Certificate Chains

A public key certificate, also known as a digital certificate, is an electronic document used to prove the ownership of a public key. This allows other parties to rely upon assertions made by the private key that corresponds to the public key that is certified. In the context of (D)TLS, certificates serve as a guarantee that the communication is done with the claimed entity and not someone impersonating it.

A Certification Authority (CA) is an entity that issues digital certificates. There are two types of CAs: the **root CAs** and the **intermediate CAs**. An intermediate CA is provided with a certificate with signing capabilities signed by one of the root CAs. A **certificate chain** is a list of certificates from the root certificate to the end-user certificate, including any intermediate certificates along the way. In order for a certificate to be trusted by a device, it must be directly or indirectly issued by a CA trusted by the device.

In (D)TLS, the certificates are in the X.509 format, defined in RFC 5280[**?**].

2.3 Authenticated Encryption With Associated Data (AEAD) Ciphers

Authenticated Encryption (AE) and AEAD are forms of encryption which simultaneously provide confidentiality, integrity and authenticity guarantees on the data. An AE cipher takes as input a key, a nonce and a plaintext and outputs the pair (ciphertext, MAC), if it is encrypting and does the inverse process, while also performing the Message Authentication Code (MAC) check if it is decrypting.

AEAD is nothing more than a variant of AE, which comes with an extra input parameter that is additional data, that is **only authenticated, but not encrypted**. Some AEAD ciphers have shorter authentication tags (*i.e.* shorter MACs), which makes then more suitable for low-bandwidth networks, since the messages to be sent are smaller in size.

2.4 ECC

public key cryptography is based on the use of one-way math functions. Such functions make it easy to compute the answer given an input, but hard to compute the input given the answer. For example, RSA uses factoring as the one one way function: it is easy to multiply large numbers, but it is hard to factor them.

ECC is based on elliptic curves, which are set of points (x, y) that are solutions to the equation $y^2 = x^3 + ax + b$, where $4a^3 + 27b^2 \neq 0$. Depending on the value of a and b, elliptic curves assume different shapes on the plane.

The security of ECC is based on the elliptic curve discrete logarithm problem. It states that scalar multiplication is a one way function. To exemplify, given a curve $E(\mathbb{Z}/p\mathbb{Z})$ and points Q and P on that curve $Q, P \in E(\mathbb{Z}/p\mathbb{Z})$, where Q is a multiple of P, the elliptic curve discrete logarithm problem states that finding the integer k, such that $Q = kP$ is a very hard problem.

3 The TLS Protocol

TLS is a **client-server** protocol that runs on top a **connection-oriented and reliable transport protocol**, such as **TCP**. Its main goal is to provide **privacy** and **integrity** between the two communicating peers. Privacy implies that a third party will not be able to read the data, while integrity means that a third party will not be able to alter the data.

In the TCP/IP Protocol Stack, TLS is placed between the **Transport** and **Application** layers. It is designed to simplify the establishment and use of secure communications from the application developer's standpoint. The developer's task is reduced to creating a "secure" connection (*i.e.* socket), instead of a "normal" one.

A secure communication established using TLS has two phases. In the first phase, the communicating peers authenticate one to another and negotiate the parameters, such as the secret keys and the encryption algorithm. In the second phase, they exchange cryptographically protected data under the previously negotiated parameters. The first phase is done under the Handshke Protocol and the second under the Record Protocol. In order to achieve its goals, during the Handshake Protocol the client and the server exchange various messages. The message flow is depicted in Figure **??** and described in more detail in Section **??**.

TLS provides the following **security services**:

- **authentication** - both, **peer entity** and **data origin** (or **integrity**) authentication.
  - **peer entity authentication** - a peer has a guarantee that it is talking to certain entity, for example, `www.google.com`. This is achieved thought the use of AC, also known as Public Key Cryptography (PKC), (*e.g.* `RSA` and `DSA`) or **symmetric key cryptography**, using a Pre-Shared Key (PSK).

- **confidentiality** - the data transmitted between the communicating entities (the client and the server) is encrypted. Symmetric cryptography is used for data encryption (*e.g.*, AES).
- **integrity** (also called **data origin authentication**) - a peer can be sure that the data was not modified or forged, *i.e.*, there is a guarantee that the received data is coming from the expected entity. For example, a peer can be sure that the `index.html` file that was sent to when it connected to `www.google.com` did, in fact, come from `www.google.com` and it was not tampered with by an attacker (**data integrity**). This is achieved either through the use of a keyed MAC or an AEAD cipher.
- **replay protection** (also known as **freshness**) - a peer can be sure that a message has not been replayed. This is achieved through the use of sequence numbers. Each TLS record has a different sequence number, which is incremented. If a non-AEAD cipher is used, the sequence number is a direct input of the MAC function. If an AEAD cipher is used, a nonce derived from the sequence number is used as input to that cipher.

Despite using PKC, TLS does **not** provide **non-repudiation services**: neither **non-repudiation with proof of origin**, which addresses the peer denying the sending of a message, nor **non-repudiation with proof of delivery**, which addresses the peer denying the receipt of a message. This is due to the fact that instead of using **digital signatures**, either a keyed MAC or an AEAD cipher is used, both of which require a secret to be **shared** between the peers.

It is not required to use all of the tree security services every situation. In this sense, TLS is like a framework that allows to select which security services should be used for a communication session. As an example, certificate validation might be skipped, which means that the **authentication** guarantee is not provided. There are some differences regarding this claim between TLS 1.2[**?**] and TLS 1.3. For example, while in the first there is a `null` cipher (no authentication, no confidentiality, no integrity), in the latter this is not true, since it deprecated all non-AEAD ciphers in favor of AEAD ones.

The terms Secure Sockets Layer (SSL) and TLS are often used interchangeably, but one is a predecessor of another - SSL 3.0[**?**] served as the basis for TLS 1.0[**?**].

Section **??** will begin with a brief overview of the various sub-protocols that compose TLS. The TLS Record Layer will be described in sufficient detail for the TLS Handshake Protocol description that follows in Section **??**. The way each record is processed when sending and receiving data is covered in Section **??**. The symmetric keys involved in cryptographic operations that provide confidentiality and security are described in Section **??**. Section **??** explains how those keys are generated in TLS 1.2. There are various methods that the client and the server can use to exchange keys, those will be covered in Section **??**. The TLS Extension mechanism will be covered in Section **??**. There are various differences from TLS 1.2 to 1.3 and those that were not covered in the previous sections will be in Section **??**. This section ends with an outline of the main differences from DTLS to TLS in Section **??**.

3.1 TLS (Sub)Protocols

TLS is composed of several protocols, which are illustrated in Figure **??** and briefly described below:

- **TLS Record Protocol** - the lowest layer in TLS. It takes messages to be transmitted, fragments the data into manageable blocks, optionally compresses them, encrypts them and transmits the result. When the data is received, the reverse process is done. The TLS Record Protocol is located directly on top of **TCP/IP** and it serves as an **encapsulation for the remaining sub-protocols** (`4` in case of TLS 1.2 and `3` in case of TLS 1.3). To the **Record Protocol**, the remaining sub-protocols are what `TCP/IP` is to `HTTP`. A TLS Record is comprised of 4 fields, with the first 3 comprising the TLS Record header. The first field is a 1-byte record `type` specifying the type of record that is encapsulated (ex: value `0x16` for the handshake protocol). The second is a 2-byte `TLS version` field. The third is a 2-byte `length` field specifying the length of the data in the record, excluding the header itself (this means that TLS has a maximum record size of `16384` bytes). The fourth is a `fragment` field, containing `length` bytes of data that is transparent to the Record layer and should be dealt by a higher-level protocol. That higher-level protocol is specified by the `type` field. This is illustrated in Figure **??**.
- **TLS Handshake Protocol** - the core protocol of TLS. It allows the communicating peers to **authenticate** one to another and to negotiate the connection state. In TLS 1.2 a **cipher suite** and a **compression** method are negotiated. In TLS 1.3, a **cipher suite** and a **key exchange** algorithm are negotiated. The agreed upon **cipher suite** is used to provide the previously described security services. In TLS 1.2, a **cipher suite** consists of a **cipher spec**, a **key exchange** algorithm and a Pseudo-Random Function (PRF), which is used for key generation. In TLS 1.2, **cipher spec** defines the message encryption algorithm and the message authentication algorithm. In TLS 1.3, the term **cipher spec** is no longer present, since the **ChangeCipherSpec** protocol has been removed. The concept of **cipher suite** has been updated to define the pair consisting of an AEAD algorithm and a hash function to be used with HMAC-based Extract-and-Expand Key Derivation Function (HKDF). In TLS 1.3 the **key exchange** algorithm is negotiated via extensions.
- **TLS Alert Protocol** - allows the communicating peers to signal potential problems.
- **TLS Application Data Protocol** - used to transmit application data messages securely using the security parameters negotiated during the **Handshake Protocol**. The messages are treated as transparent data to the record layer.
- **TLS Change Cipher Spec Protocol** (removed in TLS 1.3) - used to activate the initial **cipher spec** or change it during the connection.

3.2 TLS 1.2 Handshake Protocol

The Handshake Protocol is responsible for negotiating a **session**, which will then be used in a **connection**. There is a difference between a TLS session and a TLS connection:

- **TLS session** - association between two communication peers that is created by the **TLS Handshake Protocol**, which defines a set of negotiated parameters (cryptographic and others, such as the compression algorithm, depending on the TLS version) that are used by the **TLS connections associated with that session**. A single **TLS session** can be shared among multiple **TLS connections** and its main purpose is to avoid the expensive negotiation of new parameters for each **TLS connection**. For

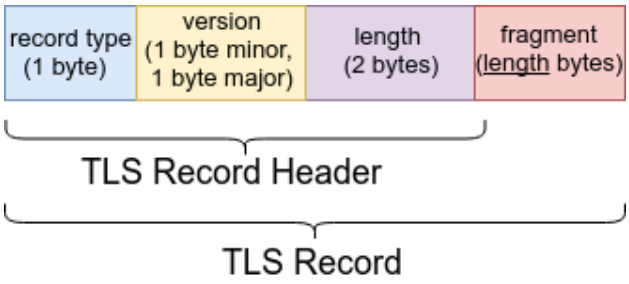


Fig. 1. TLS Record header

Fig. 2. TLS (Sub)protocols and Layers

example, let us say that a Hypertext Markup Language (HTML) page is being downloaded over the Hypertext Transfer Protocol Secure (HTTPS) and that page references some images from that same server using HTTPS links. Instead of the web browser negotiating a new TLS session for every single image again, it can re-use the the one it has established to download the HTML page, saving time and computational resources. Session resumption can be done using various approaches, such as **session identifiers**, described throughout `Section 7.4` of `RFC 5246`[**?**] and **session tickets**, defined in `RFC 5077`[**?**].

– **TLS connection** - used to actually transmit the cryptographically protected data. For the data to be cryptographically protected, some parameters, such as the secret keys used to encrypt and authenticate the transmitted data need to be established; this is done when a **TLS session** is created, during the **TLS Handshake Protocol**.

In the handshake phase the client and the server agree on which version of the TLS protocol to use, authenticate one to another and negotiate session state items like the cipher suite and the compression method. Figure **??** shows the message flow for the full TLS 1.2 handshake. `*` indicates situation-dependent messages that are not always sent. `ChangeCiperSpec` is a separate protocol, rather than a message type.

As already mentioned, every TLS handshake message is encapsulated within a TLS record. The actual handshake message is contained within the `fragment` of a TLS record. The record type for a handshake message is `0x16`. The handshake message has the following structure: a 1-byte `msg_type` field (specifies the Handshake message type), a 2-byte `length` field (specifies the length of the `body`) and a `body` field, which contains a structure depending on the `msg_type` (similar to `fragment` field in a TLS record).

A typical handshake message flow will be described next, with only the most important fields of each message mentioned.

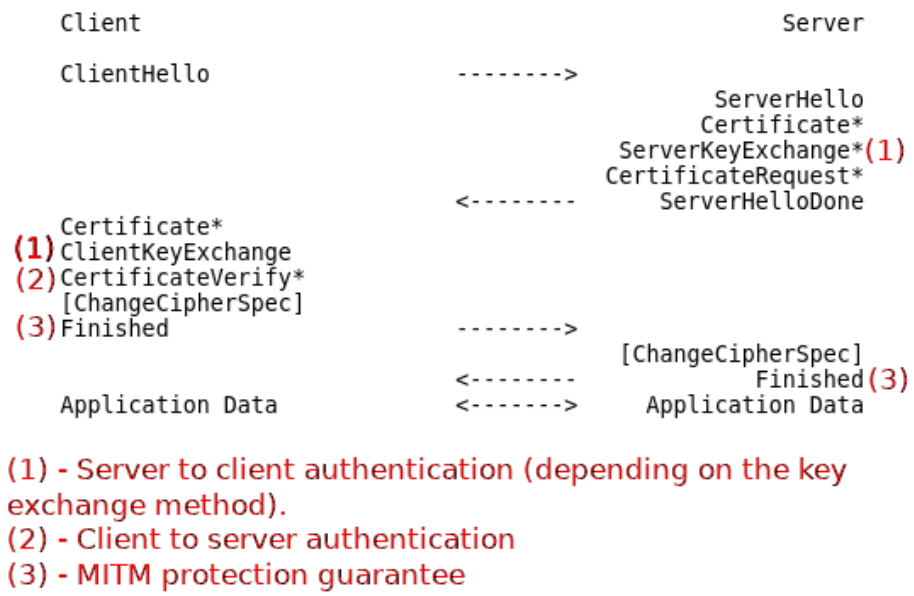


Fig. 3. TLS 1.2 message flow for a full handshake

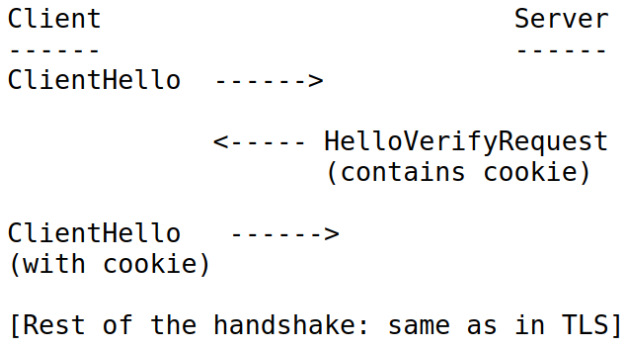


Fig. 4. DTLS handshake with HelloVerifyRequest containing the cookie

The TLS handshake starts with the client sending a `ClientHello`, containing `random`, `cipher_suites` and `compressison_methods`, among other fields.
`cipher_suites` contains a **list** of cipher suites and `compression_methods` contains a **list** of compression methods that the client supports, **ordered by preference**, with the most preferred one appearing first. The TLS record contains a 2-byte `version` field which indicates the highest version supported by the client.

The server responds to the `ClientHello` with a `ServeHello`. This message is similar, but contains the chosen `cipher_suite` and `compression_method` from the list sent by the client. Just like in the client's case, a `random` is present. The `version` field in the TLS record indicates the TLS version chosen by the server, which will be the one used for that connection.

TLS requires cryptographically secure pseudorandom numbers to be generated by both of the parties independently. Those random numbers (or *nonces*) are essential for freshness (protection against replay attacks) and session uniqueness. To provide those properties, both of the random values are required. Those two random values are inputs to the PRF when the master secret is generated, meaning that a new keying material will be obtained with every new session. If the output of the pseudorandom number generator can be predicted by the attacker, he can predict the keying material, as described in "A Systematic Analysis of the Juniper Dual EC Incident"[**?**]. The `32-byte` random value is composed by concatenating the `4-byte` GMT UNIX time with `28` cryptographically random bytes. Note that, in TLS 1.3, the random number structure has the same length, but is generated in a different manner: the client's `32 bytes` are all random, while the server's last `8 bytes` are fixed when negotiating TLS 1.2 or 1.3.

Next, the server sends a `Certificate` message, which contains a list of public key certificates: the server's certificate, every intermediate certificate and the root certificate, *i.e*, a certificate chain. The certificate's contents will depend on the negotiated cipher suite and extensions. The same message type occurs later in the handshake, if the server requests the client's certificate with the `CertificateRequest` message. In a typical scenario, the server will not request client authentication.

The `ServerKeyExchange` message follows, containing additional information needed by the client to compute the premaster secret. This message is only sent in some key exchange methods, namely `DHE_DSS`, `DHE_RSA` and `DH_anon`. For non-anonymous key exchanges, this is the message that authenticates the server to the client, since the server sends a digital signature over the client and server randoms, as well as the server's key exchange parameters. Note that this is not the only place where the server can authenticate itself to the client. For example, if `RSA` key exchange is used, the server authentication is done indirectly when the client sends the premaster secret encrypted with the public RSA key provided in the server certificate. Since only the server knows the corresponding private key, if both of the sides generate the same keying material, then the server must be who it claims to be. In TLS 1.3 this message is non-existent and a similar functionality is taken by the `key_exchange` extension.

The `ServerHelloDone` is sent to indicate the end of `ServerHello` and associated messages. Upon the receipt of this message, the client should check if the server provided a valid certificate. This message is not present in TLS 1.3.

With the `ClientKeyExchange` message the premaster secret is set. This is done either by direct transmission of the secret generated by the client and encrypted with the server's public RSA key (thus, authenticating the server to the client) or by the transmission of

Diffie-Hellman (DH) parameters that will allow each side to generate the same premaster secret independently. In TLS 1.3 this message is non-existent and a similar functionality is taken by the `key_exchange` extension.

The `CertificateVerify` message is sent by the client to verify its certificate. This message is only sent if client authentication is used and if the client's certificate has signing capability (*i.e.* all certificates except for the ones containing fixed DH parameters).

The `ChangeCipherSpec` is its own protocol, rather than a type of handshake message. It is sent by both parties to notify the receiver that subsequent records will be protected under the newly negotiated `cipher spec` and keys. This message is not present in TLS 1.3.

The `Finished` message is an essential part of the protocol. It is the first message protected with the newly negotiated algorithms, keys and secrets. Only after both parties have sent and verified the contents of this message they can be sure that the Handshake has not been tampered with by a Man In The Middle (MITM) and begin to receive and send application data. Essentially, this message contains a keyed hash with the master secret over the hash of all the data from all of the handshake messages not including any `HelloRequest` messages and up to, but not including, this message. The other party must perform the same computation on its side and make sure that the result is identical to the contents of the other party's `Finished` message. If at some point a MITM has tampered with the handshake, there will be a mismatch between the computed and the received contents of the `Finished` message.

At any time after a session has been negotiated, the server may send a `HelloRequest` message, to which the client should respond with a `ClientHello`, thus beginning the negotiation process anew.

At any point in the handshake, the Alert protocol may be used by any of the peers to signal any problems or even abort the process through the use of an appropriate message type.

Besides the full handshake, TLS 1.2 also defines an abbreviated handshake mechanism, which can be used to either resume a previous session, or duplicate one, instead of negotiating new security parameters. This requires state to be maintained by both peers. The advantage of this mechanism is that the handshake is reduced to `1 RTT`, instead of the usual `2 RTT`, as it is the case in the full handshake.

In order to perform an abbreviated handshake, the client and the server must have established a session previously, by the means of a full handshake. In its `ServerHello` phase, the server generates and sends a `session_id`, which will be associated with the newly negotiated session.

To resume a session, in its `ClientHello` phase the client includes the `session_id` of the session it wants to resume. It is up to the server to decide if it will resume that session. In the positive case, the server responds with a `ServerHello` containing the same `session_id` value as the one sent by the client. In the negative case, the `ServerHello` will contain a different `session_id` value, thus triggering a new session negotiation process.

The keying material, such as the bulk data symmetric encryption keys and the MAC keys are formed by hashing the new client and server random values with the master secret. Therefore, provided that the master secret has not been compromised and that the secure hash operations are, in fact, secure, the new connection will be secure and independent from previous ones. The TLS 1.2 spec, suggests and upper limit of 24 hours for `session`

ID lifetimes, since an attacker which obtains the master secret may be able to impersonate the compromised party until the corresponding session ID is retired.

3.3 TLS Record Processing

A TLS record must go through some processing before it can be sent over the network. This processing is done by the **TLS Record Protocol** and involves the following steps (1-4 for TLS 1.2 and 1, 3-4 for TLS 1.3):

1. **Fragmentation** - the **TLS Record Layer** takes arbitrary-length data and **fragments** it into manageable pieces: each one of the resulting fragments is called a TLSPlaintext. Client message boundaries are not preserved, which means that multiple messages of the same type may be placed into the same fragment or a single message may be fragmented across several records.
2. **Compression** (removed in TLS 1.3) - the **TLS Record Layer** compresses the TLSPlaintext structure according to the negotiated compression method, outputting TLSCompressed. Compression is optional. If the negotiated compression method is null, TLSCompressed is identical to TLSPlaintext.
3. **Cryptographic Protection** - in TLS 1.2, either an AEAD cipher or a separate encryption and MAC functions transform a TLSCompressed fragment into a TLSCipherText fragment. In the case of TLS 1.3, the TLSPlaintext fragment is transformed into a TLSCipherText by applying an AEAD cipher, since all non-AEAD ciphers have been removed.
4. Append the TLS Record Header - encapsulate TLSCipherText in a TLS Record.

The process described above, as well as the structure names are depicted in Figure **??**. The compression step is not present in TLS 1.3. The structure names are exactly as the appear in the TLS specifications.

3.4 TLS Keying Material

In TLS, the confidentiality and integrity guarantees are achieved through the use of SC. Consequently, the communicating peers need to **share** a **set of keys**. In TLS they are derived independently by the client and the server, during the TLS Handshake Protocol.

The keys appear with different names in TLS 1.2 and 1.3 specs, but they serve the same purpose. Additionally, more more keys can be found in TLS 1.3, for reasons that will be covered in Section **??**. In TLS 1.2, the peers agree on the following set of keys:

- client write key - used by the client to encrypt the data to be sent
- client read_key - used by the client to decrypt the incoming data from the server
- server write key - used by the server to encrypt the data to be sent
- server read key - used by the server to decrypt the incoming data from the server
- client write IV - used by the client for implicit nonce techniques with AEAD ciphers
- server_write_IV - used by the server for implicit nonce techniques with AEAD ciphers
- client write MAC key (TLS 1.2 only) - used by the client to authenticate the data to be sent

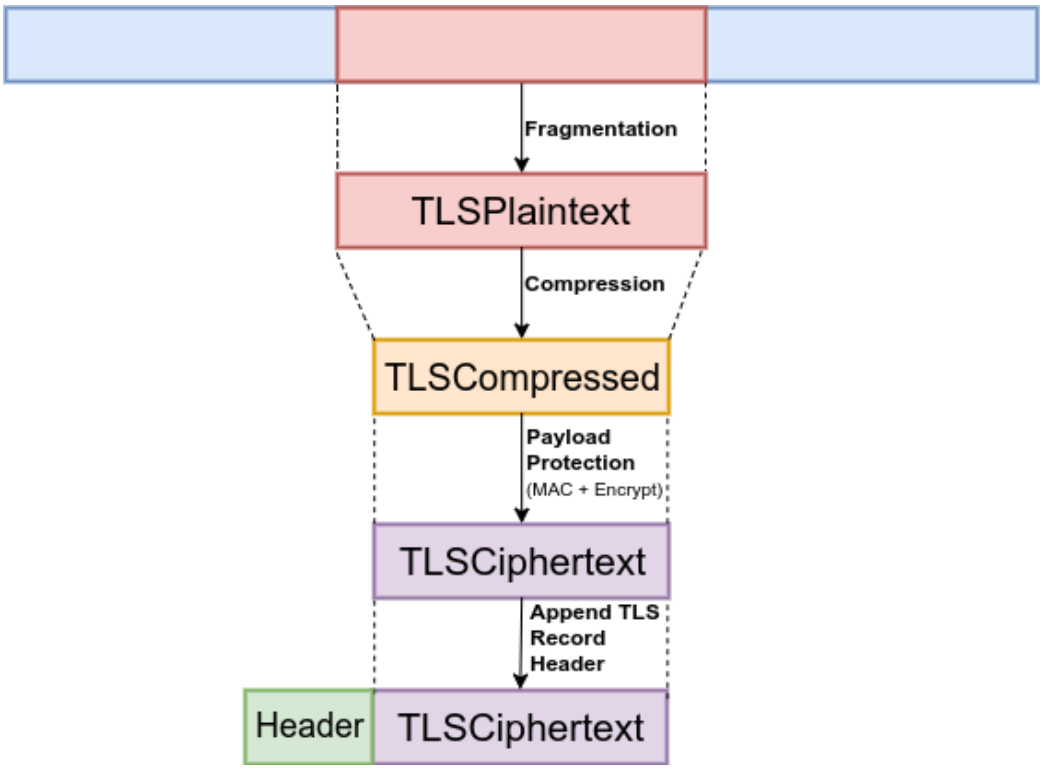


Fig. 5. TLS 1.2 Record Processing

– `client write MAC key` (TLS 1.2 only) - used by the client to authenticate the data to be sent

When communicating with one another, the client uses one key to encrypt the data that it sends to the server and another key, different from the first one, to decrypt the data that it receives from the server, and vice-versa. This implies that the following relationships must hold: `client write key == server read key` and `server write key == client read key`.

3.5 TLS 1.2 Keying Material Generation

The generation of secret keys, used for various cryptographic operations involves the following steps, in order:

1. Generate the **premaster secret**.
2. From the **premaster secret** generate the **master secret**.
3. From the **master secret** generate the various secret keys, which will be used in the cryptographic operations.

The derivation of the keying material needed for a connection is done using the TLS PRF. It is defined as `PRF(secret, label, seed) = P_hash(secret, label + seed)`. The `P_hash(secret, seed)` function is an auxiliary data expansion function which uses a single cryptographic hash function to expand a `secret` and a `seed` into an arbitrary quantity of output. Therefore, it can be used to generate anywhere from 1 to an infinite number of bits of output. `PRF(secret, label, seed)` is used to generate as many bits of output as needed. When generating the master secret, the `secret` input is the `premaster secret`. When generating the key block, from which the final keys will be obtained, the `secret` input is the `master secret`.

The cryptographic hash function used in `P_hash(secret, label, seed)` is the hash function that is implicitly defined by the cipher suite in use. All of the cipher suites defined in the TLS 1.2 base spec use `SHA-256` and any new cipher suites must explicitly specify a the same hash function or a stronger one.

3.6 TLS 1.2 Key Exchange Methods

The way the peremaster secret is generated depends on the key exchange method used. This is the only phase of the keying material generation phase that is variable for a fixed cipher suite, since a cipher suite defines the PRF function that will be employed. Neither the derivation of the shared keys are impacted by the key exchange method.

There are many key exchange methods to choose from. Some of them are defined in the base spec (`RFC5246`[**?**]), while others in separate Request For Comment (RFC)s. For example, the ECC based key exchange, specified in `RFC4492` [**?**]).

The base spec specifies four key exchange methods, one using RSA and three using DH:

– static RSA (`RSA`; removed in TLS 1.3) - the client generates the premaster secret, encrypts it with the server's public key (which it obtained from the server's `X.509`

certificate) and sends it to the server. The server then decrypts it using the corresponding private key and uses it as its premaster secret. Perfect Forward Secrecy (PFS) is a property that preserves the confidentiality of past interactions even if the long-term secret is compromised. This key exchange method offers authenticity, but does not offer PFS.
- anonymous DH (DH_annon; removed in TLS 1.3) - each run of the protocol, uses different pubic DH parameters, which are generated dynamically. This results in a different, **ephemeral** key being generated every time. Since the exchanged DH parameters are **not authenticated**, the resulting key exchange vulnerable to MITM attacks. TLS 1.2 spec states that cipher suites using DH_annon **must not** be used, unless the application layer explicitly requests so. This key exchange offers PFS, but does not offer authenticity.
- fixed/static DH (DH; removed in TLS 1.3) - the server's/client's public DH parameter is embedded in its certificate. This key exchange method offers authenticity, but does not offer PFS.
- ephemeral DH (DHE) - the DH protocol is used, identically to DH_annon, but the public parameters are digitally signed in some way, usually using the sender's private RSA (DHE_RSA) or Digital Signature Algorithm (DSA) (DHE_DSS) key. This key exchange offers both, authenticity and PFS.

When either of the DH variants is used, the value obtained from the exchange is used as the premaster secret. Usually, only the server's authenticity is desired, but client's can also be achieved if it supplies the server with its certificate. Whenever the server is authenticated, it is secure against MITM attacks. Table **??** summarizes the security properties offered by each key exchange method.

Table 1. Key exchange methods and security properties

Key Exch Meth	Authentication	PFS
RSA	X	
DH_anon		X
DH	X	
DHE	X	X

In TLS 1.3, static RSA and DH ciphersuites have been removed, meaning that all public key exchange mechanisms now provide PFS. Even though anonymous DH key exchange has been removed, unauthenticated connections are still possible, by either using raw public keys[**?**] or not verifying the certificate chain and any of its contents.

The use of ECC-based key exchange (Elliptic Curve Diffie-Hellman (ECDH) and Elliptic Curve Diffie-Hellman Ephemeral (ECDHE)) and authentication (Elliptic Curve Digital Signature Algorithm (ECDSA)) algorithms with TLS is described in RFC4492[**?**]. The document introduces five new ECC-based key exchange algorithms, all of which use ECC to compute the premaster secret, differing only in whether the negotiated keys are ephemeral (ECDH) or long-term (ECDHE), as well as the mechanism (if any) used to authenticate them. Three new ECDSA **client authentication** mechanisms are also defined, differing in the algorithms that the certificate must be signed with, as well as the key exchange

algorithms that they can be used with. Those features are negotiated through TLS extensions.

3.7 TLS Extensions

TLS extensions were originally defined in RFC 4366[?] and later merged into the TLS 1.2 base spec. Each extension consists of an extension type, which identifies the particular extension type, and extension data, which contains information specific to a particular extension.

The extension mechanism can be used by TLS clients and servers; it is backwards compatible, which means that the communication is possible between a TLS client that supports a particular extension and a server that does not support it, and vice versa. A client may request the use of extensions by sending an extended ClientHello message, which is just a normal ClientHello with an additional block of data that contains a list of extensions. The backwards compatibility is achieved based on the TLS requirement that the servers that are not extensions-aware must ignore the data added to the ClientHellos that they do not understand (section 7.4.1.2 of RFC 2246[?]). Consequently, even servers running older TLS versions that do not support extensions, will not break.

The presence of extensions can be determined by checking if there are bytes following the compression_methods field in the ClientHello. If the server understands an extension, it sends back an extended ServerHello, instead of a regular one. An extended ServerHello is a regular ServerHello with an additional block of data following the compression_method, containing a list of extensions.

An extended ServerHello message can only be sent in a response to an extended ClientHello message. This prevents the possibility that an extended ServerHello message could cause a malfunction of older TLS clients that do not support extensions. An extension type must not appear in the extended ServerHello, unless the same extension type appeared in the corresponding extended CleintHello, and if this happens, the client must abort the handshake.

3.8 TLS 1.3

Due to limited space, TLS 1.3[?] will not be described in detail. The focus was on TLS 1.2 instead, because TLS 1.3 is still in draft mode and 1.2 is the latest and the recommended to use version. Despite the protocol name not suggesting it, TLS 1.3 is very different from TLS 1.2. It should have probably been called TLS 2.0 instead.

Numerous differences from TLS 1.3 to 1.2 have been mentioned throughout the document. Various characteristics found in TLS 1.3 make it more suitable for the context of IoT than TLS 1.2. Some of them were already mentioned previously, and in this section a additional ones will be outlined.

The first important difference is that the use of extensions is required in TLS 1.3. This can be explained by the fact that some of the functionality has been moved into extensions, in order to preserve backwards-compatibility with the ClientHellos of the previous versions. The way a server distinguishes if a client is requesting TLS 1.3 is by checking the presence of the supported_versions extension in the extended ClientHello.

In TLS 1.3 more data is encrypted and the encryption begins earlier. For example, at the server-side there is a notion of "encrypted extensions". The EncryptedExtensions

message, as the name suggests, contains a list of extensions that are encrypted under a symmetric key. It contains any extensions that are not needed for the establishment of the cryptographic context.

One of the main problems with using TLS in IoT is that while IoT traffic needs to be quick and lightweight, TLS 1.2 adds two additional round trips (2 RTT) to the start of every session. TLS 1.3 handshake has a lower latency, and this is extremely important in the context of IoT. The full TLS 1.3 handshake is only 1 RTT. TLS 1.3 even allows clients to send data on the first flight (known as **early client data**), when the clients and servers share a PSK (either obtained externally or via a previous handshake). This means that in TLS 1.3 0-RTT data is possible, by encrypting it with a key derived from a PSK. Session resumption via identifiers and tickets has been obsoleted in TLS 1.3, and both methods have been replaced by a PSK mode. This PSK is established in a previous connection after the handshake is completed and can be presented by the client on the next visit.

Keying material generation is more complex in TLS 1.3 than in TLS 1.2, since different keys are used to encrypt data throughout the Handshake protocol. This can be explained by the fact that in TLS 1.3 the encryption begins earlier. Other Handshake messages besides Finished are encrypted. As a result, multiple encryption keys are generated and used to encrypt different data throughout the handshake.

The way the keying material is derived is also different. The PRF construction described above has been replaced. In TLS 1.3, key derivation uses the HKDF function [**?**] and its two components: HKDF-Extract and HKDF-Expand. This new design allows easier analysis by cryptographers due to improved key separation properties.

3.9 DTLS

As already mentioned, DTLS is an adaption of TLS that runs on top of an unreliable transport protocol, such as UDP. The design of DTLS is deliberately very similar to TLS, in fact, its specification is written in terms of differences from TLS. This similarity allows to both, minimize new security invention, and maximize the amount of code and infrastructure reuse. The changes are mostly done at the lower level and don not affect the core of the protocol. Even extensions defined before DTLS existed can be used with it. The latest version of DTLS is 1.2 and it is defined in RFC 6347[**?**]. There is a draft of DTLS 1.3 [**?**] that is currently under active development.

Since DTLS operates on top of an unreliable transport protocol, such as UDP, it must explicitly deal with the absence of reliable and ordered assumptions that are made by TLS. The main differences from DTLS 1.2 to TLS 1.2 are:

- two new fields are added to the record layer: an explicit 2 byte sequence number and a 6 byte epoch. The DTLS MAC is the same as in TLS, however, rather than using the implicit sequence number, the 8 byte value formed by concatenation of the epoch number and the sequence number is used.
- stream ciphers must not be used with DTLS.
- a stateless cookie exchange mechanism has been added to the handshake protocol in order to prevent Denial-of-Service (DoS) attacks. To accomplish this, a new handshake message, the HelloVerifyRequest has been added. After the ClientHello, the server responds with a HelloVerifyRequest containing a cookie, which is returned back to the server in another ClientHello that follows it. After this, the handshake proceeds

as in TLS. This is depicted in Figure **??**. Although optional for the server, this mechanism highly recommended, and the client must be prepared to respond to it. DTLS 1.3 follows the same idea, but does it differently, namely, the `HelloVerifyRequest` message has been removed, and the cookie is conveyed to the client via an extension in a `HelloRetryRequest` message.
– the handshake message format has been extended to deal with message reordering, fragmentation and loss by addition of three new fields: a message sequence field, a fragment offset field and a fragment length field.

4 Related Work

Lightweight cryptography is an important topic in the context of IoT security, due to the resource-limited nature of the devices. This section will begin with the description of the work done in this area.

Biryukov *et al*[**?**] explore the topic of lightweight symmetric cryptography, providing a summary of the lightweight symmetric primitives from the academic community, the government agencies and even proprietary algorithms which have been either reverse-engineered or leaked. All of those algorithms are listed in the paper, alongside relevant metrics. The list will not be included herein due to the lack of space. The authors also proposed to split the field into two areas: ultra-lightweight and IoT cryptography.

The paper systematizes the knowledge in the area oe lightweight cryptography in order to define "lightweightness" more precisely. The authors observed that the design of lightweight cryptography algorithms varies greatly, the only unifying thread between them being the low computing power of the devices that they are designed for.

The most frequently optimized metrics are the memory consumption, the implementation size and the speed or the throughput of the primitive. The specifics depend on whether the hardware or the software implementations of the primitives are considered.

If the primitive is implemented in hardware, the memory consumption and the implementation size are lumped together into its gate area, which is measured in Gate Equivalents (GE), a metric quantifying how physically large a circuit implementing the primitive is. The throughout is measured in *bytes/sec* and it corresponds to the amount of plaintext processed per time unit. If a primitive is implemented in software (typically for use in micro-controllers), the relevant metrics are the RAM consumption, the code size and the throughput of the primitive, measured in *bytes/CPU cycle*.

To accommodate the limitations of the constrained devices, most lightweight algorithms are designed to use smaller internal states with smaller key sizes. After analysis, the authors concluded that even though at least `128 bit` block and key sizes were required from the AES candidates, most of the lightweight block ciphers used only `64-bit` blocks, which leads to a smaller memory footprint in both, software and hardware, while also making the algorithm better suited for processing of smaller messages.

Even though algorithms can be optimized in implementation: whether it is a software or a hardware, dedicated lightweight algorithms are still needed. This comes down mainly to two factors: there are limitations to the the extent of the optimizations that can be done and the hardware-accelerated encryption is frequently vulnerable to various Side-Channel Attack (SCA)s. An example of such an attack is the one done on the Phillips light bulbs [**?**], where the authors were able to recover a secret key used to authenticate updates.

It is more difficult to implement a lightweight hash function than a lightweight block cipher, since standard hash functions need large amounts of memory to store both: their internal states, for example, `1600 bits` in case of SHA-3, and the block they are operating on, for example, `512 bits` in the case of SHA-2. The required internal state is acceptable for a desktop computer, but not for a constrained device. Taking this into consideration, the most common approach taken by the designers is to use a sponge construction with a very small bitrate. A sponge function is an algorithm with an internal state that takes as an input a bit stream of any length and outputs a bit stream of any desired length. Sponge functions are used to implement many cryptographic primitives, such as cryptographic hashes. The bitrate decides how fast the plain text is processed and how fast the final digest is produced. A smaller bitrate means that the output will take longer to be produced, which means that a smaller capacity (the security level) can be used, which minimizes the memory footprint at the cost of slower data processing. A capacity of `128 bits` and a bitrate of `8 bits` are common values for lightweight hash functions.

Another trend in the lightweight algorithms noticed by the authors is the preference for *ARX*-based and *bitsliced-S-Box* based designs, as well as simple key schedules.

Finally, a separation of the "lightweight algorithm" definition into two distinct fields has been proposed:

- **Ultra-Lightweight Crypto** - algorithms running on very cheap devices **not connected to the internet**, which are easily replaceable and have a limited life-time. Examples: *RFID* tags, smart cards and remote car keys.
- **IoT Crypto** - algorithms running on a low-power device, **connected to a global network**, such as the internet. Examples: security cameras, smart light bulbs and smart watches.

Considering the two definitions above, this the work of this dissertation focuses on **IoT Crypto** devices. A summary of differences between the both categories is summarized in table **??**.

Table 2. A summary of the differences between ultra-lightweight and IoT crypto

	Ultra-Lightweight	**IoT**
Block Size	64 bits	128 bits
Security Level	80 bits	128 bits
Relevant Attacks	low data/time complexity	same as "regular" crypto
Intended Platform	dedicated circuit (ASIC, RFID...)	micro-controllers, low-end CPUs
SCA Resilience	important	important
Functionality	one per device, e.g. authentication	encryption, authentication, hashing...
Connection	temporary, only to a given hub	permanent, to a global network

While there is a high demand for lightweight public key primitives, the required resources for them are much higher than for symmetric ones. As a paper by Katagi *et al*[**?**] concluded, there are no promising primitives that have enough lightweight and security properties, compared to the conventional ones, such as RSA and ECC. Further research on this topic, as part of the work on this dissertation, lead to the same conclusion.

Lightweight cryptography is an important topic this work and there are papers detailing various algorithms. In order to provide a good overview of it while staying succinct, recent papers that provide a summary of the area, rather than focusing on specific implementations, were chosen. The remainder of this section will focus on the work done on the (D)TLS protocol in the context of IoT.

The "Scalable Security With Symmetric Keys"[?] paper proposes a key management architecture for resource-constrained devices, which allows devices that have no previous, direct security relation to use (D)TLS using one of two approaches: shared symmetric keys or raw public keys. The resource-constrained device is a server that offers one or more resources, such as temperature readings. The idea in both approaches is to introduce a third-party `trust anchor (TA)` that both, the client and the server use to establish trust relationships between them.

The first approach is similar to Kerberos[?], and it does not require any changes to the original protocol. A client can request a PSK `Kc` from the `TA`, which will generate it and send it back to the client via a secure channel, alongside a `psk_identity` which has the same meaning and use as in `RFC 4279`[?]. When connecting to the server, the client will send to the server the `psk_identity` that it received in a previous handshake. Upon its receipt, the server will derive the `Kc`, using the `P_hash()` function defined in `RFC 5246`[?].

The second approach consists in requesting an Authorized Public Key (APK) from the `TA`. The client includes his Raw Public Key (RPK) in its request, which is used for authorization. The TA creates an authorization certificate, protects it with a MAC and sends it to the client alongside the server's public key. The client then sends this APK (instead of the RPK) when connecting to the server, which verifies it (to authorize the client) and proceeds with the handshake in the RPK mode, as defined in `RFC 4279` [?]. To achieve this, a new certificate structure is defined, alongside a new `certificate_type`. The new certificate structure is just the `RFC7250` [?] structure, with an additional MAC.

The hash function used for key derivation is SHA256. The authors evaluated the performance of their solution with and without SHA2 hardware acceleration and concluded that while it had significant impact on key derivation, it had little impact on the total handshake time (`711.11 ms` instead of `775.05 ms`), since most of the time was spent in sending data over the network and other parts of the handshake, the longest one being the `ChangeCipherSpec` message which required a processing time of `17.79ms`.

6LoWPAN[?] is a protocol that allows devices with limited processing ability and power to transmit information wirelessly using the `IPv6` protocol. The protocol defines IP Header Compression (IPHC) for the IP header, as well as, Next Header Compression (NHC) for the IP extension headers and the UDP header in `RFC 6282`[?]. The compression relies on the shared context between the communicating peers.

The work proposed in [?] uses this same idea, but with the goal of compressing DTLS headers. 6LoWPAN does not provide ways to compress the UDP payload and layers above. A proposed standard[?] for generic header compression for 6LoWPANs that can be used to compress the UDP payload, does exit, however. The authors propose a way to compress DTLS headers and messages using this mechanism.

Their work defines how the DTLS Record header, the DTLS Handshake header, the `ClientHello` and the `ServerHello` messages can be compressed, but notes that the same compression techniques can be used to compress the remaining handshake messages. They explore two cases for the header compression: compressing both, the Record header and the Handshake header and compressing the Record header only, which is useful after the

handshake has completed and the fragment field of the Record layer contains application data, instead of a handshake message.

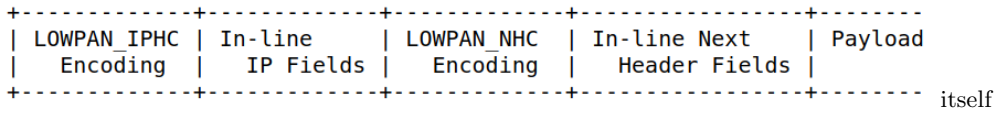


Fig. 6. IPv6 Next Header Compression

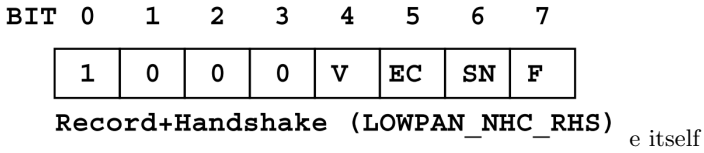


Fig. 7. LOWPAN_NHC_RHS structure

Each DTLS fragment is carried over as a UDP payload. In this case, the UDP payload carries a header-like payload (the DTLS record header). Figure **??** shows the way IPv6 next header compression is done. The authors use the same value for the `LOWPAN_NHC Encoding` field (defined in `RFC 6282`[**?**]) as in `RFC7400` and define the format of the `In-line Next Header Fields` (also defined in [**?**]), which is the compressed DTLS content. The `LOWPAN_IPHC Encoding` and `In-Line IP Fields` fields are used in the IPv6 header compression and are not in the scope of the paper.

All of the cases follow the same basic idea, for this reason only one of them will be exemplified: the case where both, the Record and the Handshake headers are compressed. In this case `LOWPAN_NHC Encoding` will contain the `LOWPAN_NHC_RHS` structure (depicted in figure **??**), which is the compressed form of the Record and Handshake headers. The parts that are not compressed will be contained in the `Payload` part. The first four bits represent the ID field and in this case they are fixed to `1000`, that way, the decompressor knows what is being compressed (*i.e* how to interpret the structure that follows the ID bits). If the `F` field of the `LOWPAN_NHC_RHS` structure contains the bit `0`, it means that the handshake message is not fragmented, so the `fragment_offset` and `fragment_length` fields are elided from the Handshake header (common case when a handshake message is not bigger than the maximum header size), meaning that they are not going to be sent at all (*i.e.* they are not going to be present in the `Payload` part). If the `F` bit has the value `1`, the `fragment_offset` and `fragment_length` fields are carried inline (*i.e.* they are present in the `Payload` part). The remaining two fields define similar behavior for other header fields (some of them assume that some default value is present, when a field is elided). The `length` field in the Record and Handshake headers are always elided, since they can be inferred from the lower layers.

The evaluation showed that the compression can save a significant number of bits: the Record header, that is included in all messages can be compressed by 64 bits (*i.e.* by 62%).

There is also a proposal for TCP header compression for 6LoWPAN[**?**], which if adopted, in many cases can compress the mandatory 20 bytes TCP header into 6 bytes. This means that the same ideas can be applied to TCP and TLS as well.

Later, in 2013, Raza *et al.* proposed a security scheme called Lithe[**?**], which is a lightweight security solution for Constrained Application Protocol (CoAP) that uses the same DTLS header compression technique as in [**?**] with the goal of implementing it as a security support for CoAP. CoAP[**?**] is a specialized *REST*ful Internet Application Protocol for constrained devices. it is designed to easily translate to HTTP, in order to simplify its integration with the web, while also meeting requirements such as multicast support and low overhead. CoAP is like "HTTP for constrained devices". It can run on most devices that support UDP or a UDP-like protocol. CoAP mandates the use of DTLS as the underlying security protocol for authenticated and confidential communication. There is also a CoAP specification running on top of TCP, which uses TLS as its underlying security protocol currently being developed[**?**].

The authors evaluated their system in a simulated environment in *Contiki OS*[**?**], which is an open-source operating system for the IoT. They obtained significant gains in terms of packet size (similar numbers to the ones observed in [**?**]), energy consumption (on average 15% less energy is used to transmit and receive compressed packets), processing time (the compression and decompression time of DTLS headers is almost negligible) and network-wide response times (up to 50% smaller RTT). The gains in the mentioned measures are the largest when the compression avoids fragmentation (in the paper, for payload size of 48 bytes).

Angelo *et al.* [**?**] proposed to integrate the DTLS protocol inside CoAP, while also exploiting ECC optimizations and minimizing ROM occupancy. They implemented their solution in an off-the-shelf mote platform and evaluated its performance. DTLS was designed to protect web application communication, as a result, it has a big overhead in IoT scenarios. Besides that, it runs over UDP, so additional mechanisms are needed to provide the reliability and ordering guarantee. With this in mind, the authors wanted to design a version of DTLS that both: minimizes the code size and the number of exchanged messages, resulting in an optimized Handshake protocol.

In order to minimize the code size occupied by the DTLS implementation, they decided to delegate the tasks of **reliability** and **fragmentation** to CoAP. This means that the code responsible for those functionalities, can be removed altogether from the DTLS implementation, thus reducing ROM occupancy. This part of their work was based on an informational RFC draft[**?**], in which the authors profiled DTLS for CoAP-based IoT applications and proposed the use of a *RESTful* DTLS handshake which relies on CoAP block-wise transfer to address the fragmentation issue.

To achieve this they proposed the use of a *RESTful* DTLS connection as a CoAP resource, which is created when a new secure session is requested. The authors exploit the the CoAPs capability to provide connection-oriented communication offered by its message layer. In particular, each Confirmable CoAP message requires an Acknowledgement message (page 8 of RFC 7252 [**?**]), which acknowledges that a specific Confirmable message has arrived, thus providing reliable retransmission.

Instead of leaving the fragmentation function to DTLS, it was delegated to the block-wise transfer feature of CoAP[**?**], which was developed to support transmission of large payloads. This approach has two advantages: first, the code in the DTLS layer responsible for this function can be removed, thus reducing ROM occupancy, and second, the fragmentation/reassembly process burdens the lower layers with state that is better managed in the application layer.

The authors also optimized the implementation of basic operations on which many security protocols, such as ECDH and ECDSA rely upon. The first optimization had to do with modular arithmetic on large integers. A set of optimized assembly routines based on [**?**] allow the improved use of registers, reducing the number of memory operations needed to perform tasks such as multiplications and square roots on devices with `8-bit` registers.

Scalar multiplication is often the most expensive operation in Elliptic Curve (EC)-based cryptography, therefore optimizing it is of high interest. The authors used a technique called *IBPV* described in [**?**], which is based on pre-computation. of a set of discrete log pairs. The mathematical details have been purposefully omitted, since they are not relevant for this description. The *IBPV* technique was used to improve the performance of the ECDSA signature and extended to the ECDH protocol. In order to reduce the time taken to perform an ECDSA signature verification, the *Shamir Trick* was used, which allows to perform the sum of two scalar multiplications (frequent operation in EC cryptography) faster than performing two independent scalar multiplications.

The results showed that the ECC optimizations outperform the scalar multiplication in the state of the art class 1 device platforms, while also improving the the network lifetime by a factor of up to 6.5 with respect to a standard, non-optimized implementation. Leaving reliability and fragmentation tasks to CoAP, reduces the DTLS implementation code size by approximately 23%.

`RFC 7925`[**?**] describes a TLS and DTLS 1.2 profile for IoT devices that offer communication security services for IoT applications. In this context, "profile" means available configuration options (ex: which cipher suites to use) and protocol extensions that are best suited for IoT devices. The document is rather lengthy, only its fundamental parts will be summarized. A number of relevant RFCs will also be described.

`RFC 7925` explores both cases: constrained clients and constrained servers, specifying a profile for each one and describing the main challenges faced in each scenario. The profile specifications for constrained clients and servers are very similar. Code reuse in order to minimize the implementation size is recommended. For example, an IoT device using a network access solution based on TLS, such as EAP-TLS[**?**] can reuse most parts of the code for (D)DTLS at the application layer.

For the credential types the profile considers 3 cases:

- PSK - authentication based on PSKs is described in `RFC 4249`[**?**]. When using PSKs, the client indicates which key it wants to use by including a PSK identity in its `ClientKeyExchange` message. A server can have different PSK identities shared with different clients. An identity can have any size, up to a maximum of `128 bytes`. The profile recommends the use of shorter PSK identities and specifies `TLS_PSK_WITH_AES_128_CCM_8` as the only mandatory-to-implement cipher suite to be used with PSKs, just like CoAP does. If a PFS cipher suite is used, ephemeral DH keys should not be reused over multiple protocol exchanges.

- RPK - the use of RPKs in (D)TLS is described in RFC 7250[?]. With RPKs, only a subset of the information that is found in typical certificates is used: namely the SubjectPublicKeyInfo structure, which contains the necessary parameters to describe the public key (the algorithm identifier and the public key itself). Other PKIX certificate[?] parameters are omitted, making the resulted RPK smaller in size, when compared to the original certificate and the code to process the keys simpler. In order for the peers to negotiate a RPK, two new extensions have been defined: one for the client indicate which certificate types it can provide to the server, and one to indicate which certificate types it can process from the server. To further reduce the size of the implementation, the profile recommends the use of the TLS Cached Information extension[?], which enables TLS peers to exchange just the fingerprint (a shorter sequence of bytes used to identify a public key) of the public key. Identical to CoAP, the only mandatory-to-implement cipher suite to be used with RPKs is TLS_ECDHE_ECDSA_WITH_AES_128_CCM_8.
- certificate - conventional certificates can also be used. The support for the Cached Information extension[?] and the TLS_ECDHE_ECDSA_WITH_AES_128_CCM_8 cipher suite is required. The profile restricts the use of named curves to the ones defined in RFC 4492[?]. For certificate revocation, neither the Online Certificate Status Protocol (OCSP)[?], nor the Certificate Revocation List (CRL)[?] mechanisms are used, instead this task is delegated to the software update functionality. The Cached Information extension does not provide any help with caching client certificates. For this reason, in cases where client-side certificates are used and the server is not constrained, the support for client certificate URLs is required. The client certificates URL extension[?] allows the clients to point the server to a URL from which it can obtain its certificate, which allows constrained clients to save memory and amount of transmitted data. The Trusted CA Indication[?] extension allows the clients to indicate which trust anchors they support, which is useful for constrained clients that due to memory limitation posses only a small number of CA root keys, since it can avoid repeated handshake failures. If the clients interacts with dynamically discovered set of (D)TLS servers, the use of this extension is recommended, if that set is fixed, it is not.

The signature algorithms extension[?] allows the client to indicate to the server which signature/hash pairs it supports to be used with digital signatures. The client must send this extension to indicate the use of SHA-256, otherwise the defaults defined in [?] are used. This extension is not applicable when PSK-based cipher suites are used.

The profile mandates that constrained clients must implement session resumption to improve the performance of the handshake since this will lead to less exchanged messages, lower computational overhead (since only symmetric cryptography is used) and it requires less bandwidth. If server is constrained, but the client is not, the client must implement the Session Resumption Without Server-Side State mechanism[?], which is achieved through the use of tickets. The server encapsulates the state into a ticket and forwards it to the client, which can subsequently resume the session by sending back that ticket. If both, the client and the server are constrained, both of them should implement RFC 5077[?].

The use of compression is not recommended for two reasons. First, RFC7525[?] recommends disabling (D)TLS level compression, due to attacks such as CRIME[?]. RFC7525 provides recommendations for improving the security of deployed services that use TLS

and DTLS and was published as a response to the various attacks on (D)TLS that have emerged over the years. Second, for IoT applications, the (D)TLS compression is not needed, since application-layer protocols are highly optimized and compression at the (D)TLS layer increases the implementation's size and complexity.

RFC6520[?] defines a heartbeat mechanism to test whether the peer is still alive. The implementation of this extension is recommended for server initiated messages. Note that since the messages sent to the client will most likely get blocked by middleboxes, the initial connection setup is initiated by the client and then kept alive by the server.

Random numbers play an essential role in the overall security of the protocol. Many of the usual sources of entropy, such as the timing of keystrokes and the mouse movements, will not be available on many IoT devices, which means that either alternative ones need to be found or dedicated hardware must be added. IoT devices using (D)TLS must be able to find entropy sources adequate for the generation of quality random numbers, the guidelines and requirements for which can be found in RFC4086[?].

Implementations compliant with the profile must use AEAD ciphers, therefore encryption and MAC computation are no longer independent steps, which means that neither encrypt-then-MAC[?], nor the truncated MAC[?] extensions are applicable to this specification and must not be used.

The Server Name Indication (SNI) extension[?] defines a mechanism for a client to tell a (D)TLS server the name of the server that it is contacting. This is crucial in case when multiple websites are hosted under the same IP address. The implementation of this extension is required, unless the (D)TLS client does not interact with a server in a hosting environment.

The maximum fragment length extension[?] lowers the maximum fragment length support of the record layer from 2^14 to 2^9. This extension allows the client to indicate the server how much of the incoming data it is able to buffer, allowing the client implementations to lower their RAM requirements, since it does not not need to accept packets of large size, such as the 16K packets required by plain (D)TLS. For that reason, client implementations must support this extension.

The Session Hash Extended Master Secret Extension[?] defines an extension that binds the master secret to the log of the full handshake, thus preventing MITM attacks, such as the triple handshake[?]. Even though the cipher suites recommended by the profile are not vulnerable to this attack, the implementation of this extension is advised. In order to prevent the renegotiation attack[?], the profile requires the TLS renegotiation feature to be disabled.

With regards to the key size recommendations, the authors recommend symmetric keys of at least 112 bit, which corresponds to a 233-bit ECC key and to a 2048 DH key. Those recommendations are made conservatively under the assumption that IoT devices have a long expected lifetime (10+ years) and that those key recommendations refer to the long-term keys used for device authentication. Keys that are provisioned dynamically and used for protection of transactional data, such as the ones used in (D)TLS cipher suites, may be shorter, depending on the sensitivity of transmitted data.

Even though TLS defines a single stream cipher: *RC4*, its use is no longer recommended due to its cryptographic weaknesses described in RFC 7465[?].

RFC 7925[?] points out that designing a software update mechanism into an IoT system is crucial to ensure that potential vulnerabilities can be fixed and that the functionality can be enhanced. The software update mechanism is important to change configuration

information, such as trust anchors and other secret-key related information. Although the profile refers to LM2M[?] as an example of protocol that comes with a suitable software update mechanism, there has been new work done in this area since the release of this profile. There is a document specifying an architecture for a firmware update mechanism for IoT devices[?] currently in Internet-Draft state.

5 Solution

5.1 Things That Might Be Useful To Include Somewhere

Not all IoT devices are limited to the point of not being able to use public key cryptography altogether. For some of them, the use of RPKs, which is considered the first entry point into the area of public key cryptography, is acceptable, while others are powerful enough to take advantage of certificates and Public Key Infrastructure (PKI), at least up to a point.

5.2 Evaluation

In order to evaluate the quality of the work, both, the original protocol implementation and the one provided as part of the solution, will be profiled and compared. The relevant profiling metrics are power consumption, RAM usage, storage usage, CPU cycles elapsed and time taken. The profiling will be done over various simulated scenarios, which emulate real-life usage, such as connecting to the server multiple times over a short time period and transferring small quantities of data, and connecting to the server and transferring a large amount of data, all at once.

Due to limited time and the fact that TLS 1.3 still lacks stable implementations, most likely, only the solution under TLS 1.2 will be implemented and evaluated.

5.3 Planning

During the month of February, up until mid-March 2018, the solution will be defined precisely. Due to the nature of the solution, it will have many different versions, depending on the target device and required security services. Most likely, there will be no time to implement and evaluate every possible scenario, so only a subset of them will be chosen.

From mid-March until mid-April 2018, a version that uses existing configuration options and protocol extensions to best support the IoT environment will be developed. In essence, this would involve incorporating a lightweight profile (like RFC 7925 does) into the solution. The system will be implemented in code, by modifying the *mbedTLS 2.6.0* library[?].

From mid-April until June 2018, the customized part will be implemented. This might involve custom cipher suites, key exchange methods and changes in the Handshake Protocol.

From June until July 2018, the work will be focused around PSK solutions. This might involve adapting the existing PSK configurations or creating new ones.

From July until August 2018, the work will be evaluated. The most important evaluation metrics will be chosen and the related profiling code set up. Testing scenarios will be designed and implemented.

From mid-July until mid-September 2018, the focus will be on writing the dissertation's text. Some minor improvements and additions might also be done during this period of time.

6 Objectives

IoT devices have limited resources. Those resources are processing speed, memory and power. Communication security is a desirable property in the context of interconnected devices. There are many protocols that can be used to provide communication security. (D)TLS is one of the most used protocols for this purpose.

Due to the constrained nature of the IoT devices, typical (D)TLS configurations cannot be used in many cases.

(D)TLS is a complex protocol with numerous possible configurations. Each configuration implies a certain security level and resource usage. In fact, it's almost always a tradeoff between these two. When configured properly, (D)TLS can run on constrained devices. Such a configuration might imply foregoing some of the security serivices, or using a lower security level.

A (D)TLS configuration consists of a key exchange algorithm, an encryption algorithm, a hash function and the associated key sizes. There are numerous choices for each, which leads to numerous possible configurations. As an example, the *mbedTLS 2.7.0* library has a total of 161 possible configurations, without taking into account the asmectric cryptography key sizes. Existing work does not explore the costs of the various configurations. It also fails to establish a relationship between the security services, security level and their associated costs in the context of (D)TLS. Developers wishing to deploy the (D)TLS protocol in constrained environments do not have a tool that would help them to select a (D)TLS configuration appropriate to the environment's needs and limitations.

The majority of existing work proposes a solution that is either tied to a specific protocol, such as CoAP, or requires an introduction of a third-party entity, such as the trust anchor in the case of the S3K system[**?**] or even both. This has two main issues. First, a protocol-specific solution cannot be easily used in an environment where (D)TLS is not used with that protocol. Second, the requirement of a third-party introduces additional cost and complexity, which will be a big resistance factor in adopting the technology. This is specially true for developers working on personal projects or projects for small businesses, leaving the communications insecure in the worse case scenario. The goal of this work is to design a solution that can be used out of the box and is not tailored towards any specific protocol, while fully backwards compatible with the original protocol, that can be used with both, TLS and DTLS.

Another topic that existing work fails to explore with enough detail is TLS optimization. Most of the work has been centered around DTLS and not all of it can be applied to TLS, since it Herein we want to further explore TLS optimization. There is clearly a need for that, specially with CoAP over TCP and TLS standard being currently developed. The mentioned standard does not explore any TLS optimizations, and since any IoT device using it in the future would benefit from them, this is an important area to explore.

The objective of this work is to provide a means of assisting application developers who wish to include secure communications in their applications to make security level/resource usage tradeoffs, according to the environment's needs and limitations. In order to achieve

this goal, the costs of each individual security service will be evaluated. With this information, the programmer will be able to choose a configuration that meets his security requirements and device constraints. If the limitations of the device's hardware do not allow to meet the requirements, he can decide on an alternative configuration, possibly with a loss of some security services and a lower security level, or forgo using (D)TLS altogether.

7 Methodology

Things to cover:

- local machine specs
- why CPU cycles is a good measure (how CPU cycles relate to time taken; say that cryptography is CPU-bound)
- why collecting on a local machine makes sense (should be similar on others)
- which tools I develped any why I did (too many configurations to evaluate manually)

In (D)TLS the key the authentication algorithm, the encryption algorithm, the data integrity algorithm, as well as the associated key sizes for each are all defined in a *cihersuite*. A ciphersuite defines the security properties of a (D)TLS connection. For this reason, the terms *ciphersuite* and *configuration* will be used interchangeably.

A (D)TLS ((D)TLS) connection consists of two main phases:

1. The peers authenticate one to another, agree on the data encryption and integrity algorithms that they will use and establish the shared keys. This part is known as the handshake protocol.
2. The peers exchange the data securely, using the algorithms and keys negotiated in the previous step. This part is known as the record protocol.

The relative cost of each phase depends on the chosen algorithms, as well as the amount of data transferred. For this reason, it is important to evaluate the costs of both of them.

(D)TLS has numerous possible configurations. Each one of those configurations is defined in an RFC. Each ciphersuite is assigned a unique identification number. Internet Assigned Numbers Authority (IANA) is responsible for maintaining the full list of them. At the moment of this writing, there are over 300 ciphersuites defined for (D)TLS [**?**].

mbedTLS implements a subset of those ciphersuites. As of version 2.7.0, *mbedTLS* has a total of 161 ciphersuites [**?**]. Manual cost evaluation and data analysis would greatly limit the scope of obtained results, as it would be very time consuming and error-prone. For this reason, we developed tools that would automate the profiling and collection of results.

In our work, we evaluated the *mbedTLS's* implementation of the TLS protocol. The obtained metrics reflect the algorithm's implementations used within the library.

7.1 Evaluated Metrics

7.2 Limitations

My cache:

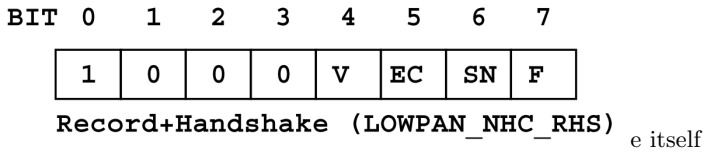
L1d cache: 32K L1i cache: 32K L2 cache: 256K L3 cache: 6144K

* what is valgrind * how does valgrind work * what is callgrind * how does callgrind work * explain how this limits the accuracy of results * explain why I went with those instead of actual measures (speed, time limit, etc) – put here or other section? – * with which options I profiled the resutls with * why I chose those options

In order to estimate the number of executed instructions, we used *valgrind*, more specifically its *callgrind* tool. *valgrind* runs the application on a synthetic CPU. While running the code in that synthetic environment, it is able to insert instructions to do profiling and debugging.

In essence, *valgrind* is a virtual machine, using just-in-time (JIT) compilation techniques. One of the most notable techniques is the dynamic recompilation, which is a feature where some part of the program is recompiled during execution.

The *valgrind* tool consists of two parts, the *valgrind core* and the *tool plugin*. The *valgrind core* transforms the machine code into a simpler form called Intermediate Representation (IR). The IR code is then passed to the *tool plugin*, which modifies the IR code as needed. This modified IR code is then passed back to the *valgrind core*, which transforms it back into machine code, which will run on the host CPU (the JIT step). This process is illustrated in Figure **??**.



e itself

Fig. 8. LOWPAN_NHC_RHS structure

In our case, the *tool plugin* is *callgrind*. Among other metrics, *callgrind* collects the number of executed instructions and can optionally simulate cache and/or branch prediction. Those metrics can then be used to estimate the number of executed instructions. *KCachegrind* is a GUI tool that can be used for such purpose.

The number of executed instructions presented by *KCachegrind* is an estimate, which might not correspond to the real value. Although undocumented, we found the formula for executed instructions estimation in the *KCachegrind*'s source code **??**. In order to estimate the number of executed CPU cycles, *cachegrind* uses the following formula: $CEst = Ir + 10 * Bm + 10 * L1m + 20 * Ge + 100 * L2m + 100 * LLm$, where

- $CEst$ - estimated CPU cycles
- Ir - instruction fetches
- Bm - mispredicted branches (direct and indirect)
- Ge - number of global bus events
- $L1m$ - total L1 cache misses (instruction fetch, data read and data write)
- $L2m$ - total L2 cache misses (instruction fetch, data read and data write)
- LLm - total shared cache misses (instruction fetch, data read and data write)

Since *callgrind* only simulates L1 and LL caches, $L2m = 0$, so the actual formula used by *KCachegrind* to estimate the number of CPU cycles used is: $CEst = Ir + 10 * Bm + 10 * L1m + 20 * Ge + 100 * LLm$. Ge is a useful metrics when synchronization primitives are present, since it counts the number of atomic instructions executed. For example, on the *x86* and $x86_64 architectures, these are instructions using the lock prefix. In our evaluation, we used single-threaded code only, fo$

If the cachegrind tool is run with disabled cache and branch prediction metrics, $Bm = 0$ *and.*

In order to estimate the number of CPU instructions executed, we used a formula based on the one used by KCachegrind.

8 Results and Data Analysis

TODO

9 Discussion

TODO

10 Further Work

TODO

11 Conclusion

The lack of security in IoT is a serious issue that can lead to a high monetary costs, when botnets infect the devices. Recent attacks clearly show that serious damage can be caused. An old saying attributed to the US National Security Agency (NSA) states that "Attacks always get better; they never get worse". Combined with the fact that the number of IoT devices is growing at a high pace, without any major improvements to their security, makes it clear that it is fundamental for this issue to be addressed.

While there are well established security solutions, not all of them can be used with IoT devices, due their constrained nature. One such example is the (D)TLS protocol, that because of its heavyweight nature is not suitable for a large part of IoT devices. With the proposed work, we want to contribute to this area, by designing a solution that is suitable for the IoT devices, transparent to the programmer and provides security services adaptable to the specific context needs.