

Efficient In-Network Adaptation of Encrypted H.264/SVC Content [★]

Hermann Hellwagner,^a Robert Kuschnig,^a Thomas Stütz,^b
Andreas Uhl^b

^a*Institute of Information Technology, Klagenfurt University
Universitätsstraße 65–67, 9020 Klagenfurt, Austria*

^b*Department of Computer Sciences, University of Salzburg
Jakob-Haringer-Straße 2, 5020 Salzburg, Austria*

Abstract

This paper addresses the efficient adaptation of encrypted scalable video content (H.264/SVC). RTP-based in-network adaptation schemes on a media aware network element (MANE) in an IPTV and VoD scenario are considered.

Two basic alternatives to implement encryption and adaptation of H.264/SVC content are investigated: *(i)* full, format-independent encryption making use of Secure RTP (SRTP); *(ii)* SVC-specific encryption that leaves the metadata relevant for adaptation (NAL unit headers) unencrypted.

The SRTP-based scheme *(i)* is straightforward to deploy, but requires the MANE to be in the security context of the delivery, i.e., to be a trusted node. For adaptation, the content needs to be decrypted, scaled, and re-encrypted. The SVC-specific approach *(ii)* enables both full and selective encryption, e.g., of the base layer only. SVC-specific encryption is based on own previous work, which is substantially extended and detailed in this paper. The adaptation MANE can now be an untrusted node; adaptation becomes a low-complexity process, avoiding full decryption and re-encryption of the content.

This paper presents the first experimental comparison of these two approaches and evaluates whether multimedia-specific encryption can lead to performance and application benefits. Potential security threats and security properties of the two approaches in the IPTV and VoD scenario are elementarily analyzed. In terms of runtime performance on the MANE our SVC-specific encryption scheme significantly outperforms the SRTP-based approach. SVC-specific encryption is also superior in terms of induced end-to-end delays. The performance can even be improved by selective application of the SVC-specific encryption scheme. The results indicate that efficient adaptation of SVC-encrypted content on low-end, untrusted network devices is feasible.

Key words: Scalable Video Coding (H.264/SVC), In-network Adaptation, RTP/RTSP MANE, Video Encryption, Format Compliance

1 Introduction

Today, multimedia content is accessible on diverse end devices over various networks. Content providers have to offer multimedia content tailored to a wide variety of possible usage contexts in order to maximize the Quality of Experience (QoE) of the individual content consumer. Scalable representations of the content facilitate the adaptation of the content to the usage context in a highly efficient manner. Apart from maximizing the QoE of the content consumer, protection of the content is a major interest of the content providers. In this work, the issue of scalable content encryption and adaptation is discussed on the basis of two encryption schemes for H.264/SVC (Scalable Video Coding), the most recently standardized video compression format [17].

SVC is the scalable extension of H.264/AVC and has been designed to enable efficient adaptation to the preferred usage context of each individual consumer, i.e., to satisfy the requirements of modern video transmission and storage systems, which are characterized by a wide range of connection qualities and receiving devices. SVC offers a scalable bitstream which can be adapted in an efficient and flexible fashion. The scalability of the SVC bitstream allows adaptation in any of its supported scalability dimensions (temporal, spatial and quality). This adaptation can be conducted in the compressed domain by simply removing parts of the SVC bitstream.

Scalability is considered to be a major advantage in network scenarios, as bitrate adaptations can be conducted efficiently. In order to enable the efficient transmission of non-scalable H.264 video content, a Real-time Transport Protocol (RTP) payload format has been developed [45]. An analogous specification is currently being developed for SVC [46]. This payload format enables the efficient adaptation of the SVC bitstream [32]. For many applications, not only efficient transmission, but also security/confidentiality of the video data during transmission is mandatory. For RTP, the default solution is the Secure Real-time Transport Protocol (SRTP) [5]. However, SRTP does not take the special properties of the RTP payload, the multimedia (SVC) data, into account, which potentially introduces overhead for content adaptation. We thus introduce an SVC-specific encryption scheme that permits efficient content adaptation within the network.

* This work was supported by the Austrian Science Fund (FWF) under project “Adaptive Streaming of Secure Scalable Wavelet-based Video (P19159)” and by the EC in the context of the ENTHRONE II project (IST-1-507637).

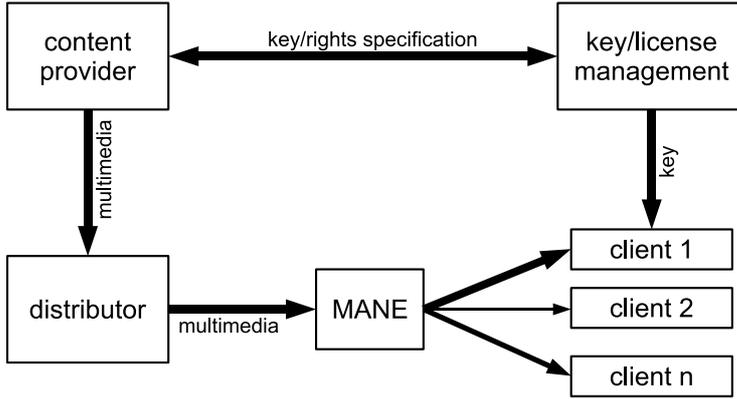


Figure 1. Application scenario

1.1 Motivation

The secure and adaptive transmission of video data is important for a broad range of real-world application scenarios: digital TV broadcasting, e.g., over IP (IPTV), video on demand (VoD), video conferencing, and surveillance.

In the following, we introduce the major application and design considerations behind our work and present the main contributions of this paper.

1.1.1 Application Scenario: IPTV and VoD

IPTV and VoD are important applications in the context of digital video. The basic setup of the scenario assumed in this paper targets those applications and is illustrated in Figure 1. The model differentiates between a content provider and a distributor and further introduces a media-aware network element (MANE) [45]. The content provider owns the content and provides the content to the distributor. Note that the content is already compressed by the content provider. The distributor is solely responsible for distributing the compressed content to the clients. The distributor will also be referred to as server in this work. In the IPTV and VoD application scenario, the MANE is assumed to be a small device close to the clients (e.g., a home gateway or a wireless network access point) that performs content adaptation according to the network conditions and the clients' demands. Both the distributor and the MANE are desired *not* to be within the security context of the content provider and the clients. Thus, key management (i.e., key exchange) shall only happen between the clients and the content provider. Key management is not addressed in this work; we assume that key exchange takes place over a secure channel.

1.1.2 In-Network Adaptation

In-network adaptation by a MANE [45] close to the clients has several advantages:

- The video content needs to be transmitted to a MANE only once in a scalable format. The MANE distributes the content to the clients (Figure 1). Thereby bandwidth in the distribution network is saved. For IPTV the MANE may be located at a telephone company's end office, distributing the content to thousands of clients and/or at the customer's home distributing the content to just a few clients.
- The content can be readily adapted to heterogeneous client devices, e.g., SD and HD TV sets and a smartphone. Typically, the devices would register with the MANE and provide their device/capabilities profiles in order to obtain tailored content.
- Bitrate adaptations can be conducted rapidly (and judiciously, exploiting the layered encoding) by the MANE to cope with fluctuating access network conditions, especially in a wireless network.

In [20], we describe H.264/SVC adaptation on an off-the-shelf wireless router, demonstrating the feasibility of such a MANE.

An alternative to the MANE approach would clearly be server-side adaptation. In the IPTV/VoD application scenario the server is equivalent to the distributor. However, using adaptation on the server, many of the advantages listed cannot be realized, or at least would incur increased server load and increased reaction time to changes in the access network. There are further practical considerations. Compared to the well-known Receiver-driven Layered Multicast (RLM) [26] technique that delivers multiple streams, the MANE approach minimizes the number of firewall pinholes (to only one). In RLM, the base layer and the enhancement layers of a scalable bitstream are multicast in separate streams. If the bitstream is sent with RTP [46], multiple sessions, i.e., different sender ports, are employed. Thus port-blocking firewalls pose a problem. Our implementation also does not rely on IP multicast, which is still poorly supported in networks.

1.1.3 Multimedia Encryption

The IPTV and VoD scenario requires that the content is encrypted (pay TV, DRM). Encryption of multimedia data may be implemented at different levels:

- *Transport level:* Encryption is applied regardless of the content; packets or stream segments of the transport layer are encrypted (e.g., using IPsec [18], TLS [33], SRTP [5]).
- *Meta format level:* Encryption is applied within the scope of a meta for-

mat, such as the ISO base media file format [15]. Approaches which employ bitstream descriptions (e.g., MPEG-21 gBSD) and encryption fall into that category [27,13].

- *Codec format level:* Codec-specific encryption is commonly applied to preserve codec specific features, most importantly scalability.

For IPTV and VoD, encryption of multimedia data with a cryptographic cipher on the transport level (e.g., with SRTP) has certain drawbacks since its application in an adaptation scenario leads to

- either a compromise of security (the distributor/MANE must be in the security context, i.e., must have the encryption keys, in order to decrypt the data and perform adaptation),
- an increase of computational complexity (the distributor/MANE has to perform decryption before adaptation and re-encryption after adaptation),
- or a waste of bandwidth or storage capacity (there has to be a stream for each of the diverse requirements of the consumers).

Therefore, specific (multimedia) encryption schemes have been introduced to overcome these disadvantages.

The preservation of scalability is essential for secure adaptation, i.e., the adaptation of the encrypted content without any knowledge of the encryption keys. A MANE, even if it is untrusted and does not have any knowledge about the encryption keys, is then capable of performing adaptation of the encrypted content. Thus the scalability information has to be accessible in the encrypted domain.

1.1.4 *Format Compliance*

The preservation of format compliance is an important property of a multimedia encryption scheme [8,38,40,9]. In the case of H.264/SVC, format compliance means that the encrypted bitstream obeys all the syntactic and semantic requirements specified in [17]. Format-compliant encryption enables easy deployment and integration into existing frameworks, e.g., in this work the same RTP packetization mechanisms as for H.264/SVC [45,46] are applied to the SVC-specifically encrypted bitstreams. In general, format-compliant encryption and the encapsulation in container formats are compatible.

Format-compliant encryption prevents any undetermined decoder behavior (e.g., a decoder freezing or crashing). If a base layer is left in plaintext and the enhancement layers are format-compliantly encrypted, decoding the base layer is still possible with a standard decoder. If conventional encryption is employed, the base layer can not be decoded, as “accidentally” generated syntax elements in the encrypted parts result in a loss of synchronisation and

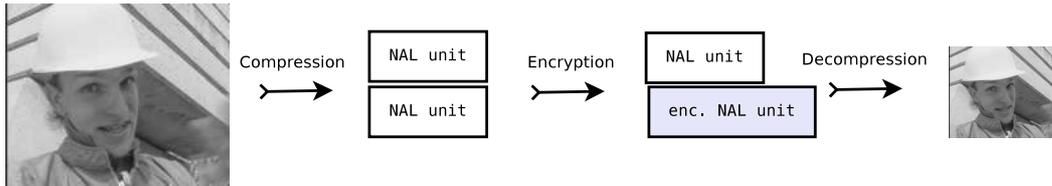


Figure 2. Transparent encryption via conditional access

render the entire bitstream undecodeable. Non-compliant bitstreams will even crash or freeze some decoder implementations. This is highly inconvenient for an end-user at the client when, e.g., its set-top box freezes.

Format-compliance is therefore fundamental transparent/perceptual encryption [23]. Transparent encryption (also called perceptual encryption, predominantly in the area of audio encryption) has been introduced mainly in the context of digital TV broadcasting: a pay TV broadcaster does not always intend to prevent unauthorized viewers from receiving and watching a program, but rather intends to promote a contract with non-paying viewers. This can be facilitated by providing a low quality version of the broadcast program for everyone; but only legitimate (paying) users get access to the full quality visual data. Figure 2 illustrates an example of transparent encryption. Only a smaller resolution video stream is publicly available. The high resolution content is format-compliantly encrypted and thus the video stream is still decodeable.

More generally speaking, format-compliant encryption enables the transparent interleaving of encrypted and plaintext video data. This is an important feature for pay TV broadcasters, who frequently mix encrypted and unencrypted data (e.g., the begin of movie is broadcast to attract customers). In order to decode the plaintext parts of the interleaved and partially encrypted data, no proprietary software nor hardware is needed. Thus deployment of format-compliant encryption is lightweight compared to proprietary encryption solutions, where proprietary software or hardware is required at the clients.

1.2 Contributions

In this work, we compare two fundamentally different approaches to enabling secure transmission and adaptation of SVC content.

One approach employs RTP packaging of the SVC data. Confidentiality is achieved by the application of SRTP. Although the SRTP approach is most likely to be applied in practice (the standards and software are already available), it has certain drawbacks. Most importantly, the MANE has to be within the security context; otherwise adaptation is not possible. Furthermore, for

adaptation purposes, the content has to be decrypted and re-encrypted by the MANE.

The second approach employs specific encryption of SVC, such that format compliance of the encrypted bitstream can be achieved. This approach enables provider–consumer security and secure adaptation on untrusted nodes (MANEs) *without* decryption and re-encryption.

In this context, we answer the question whether the increased flexibility of the SVC-specific encryption scheme has to be traded-off by an increased computational complexity during encryption. Furthermore, the computational complexity of the two schemes during adaptation (on the MANE) is examined. As the SVC-specific approach leaves scalability information in plaintext, the question arises whether a security breach is introduced. This question is answered with respect to the IPTV and VoD application scenario.

The main original contributions of this work are summarized as follows:

- (1) To the best of our knowledge, this is the first work that experimentally compares an SVC-specific encryption scheme with a conventional standardized encryption solution (SRTP) in a streaming scenario. On the basis of our experimental evaluation, we can answer the question whether encryption routines specifically tailored for multimedia data (SVC bitstreams) can lead to performance improvements in an actual *application* implementation (while the advantages of multimedia-specific encryption schemes have been postulated, at least for SVC this has not yet been shown experimentally in a real-world-based set-up).
- (2) The SVC-specific encryption approach enhances and details own previous work [39] by substantial improvements regarding the encryption processes. A more secure encryption mode is applied and the format-compliant encryption routines are discussed in detail.
- (3) Selective encryption modes have been proposed for various reasons. In most cases, performance gains have been postulated by referring to the smaller amount of data to be processed (i.e., encrypted) but the imposed parsing and processing overhead is ignored. We are able to demonstrate actual performance gains for selective SVC encryption schemes in the IPTV/VoD application context.
- (4) We provide a security analysis and evaluation for the two discussed approaches. Special interest is taken in the information leakage caused by the RTP packetization (which strongly depends on the SVC NAL units) and by the SVC-specific encryption. We find both schemes to be not secure regarding the highest level of security, but sufficiently secure for multimedia entertainment applications such as IPTV.

1.3 Paper Organization

A basic introduction to SVC and SVC adaptation is given in Section 2. The two encryption schemes are presented in Section 3 and their application together with in-network adaptation are dealt with in Section 4. Important properties of a multimedia encryption scheme are its security, its influence on compression performance, and its computational complexity. Only a negligible increase of the size of the encrypted bitstream, compared to the original bitstream, is desired. These three properties of a multimedia encryption system are discussed for the two considered encryption systems (SRTP and SVC-specific) in Sections 5, 6, and 7 respectively. In Section 8, we review previous work in the fields of video adaptation, video encryption and secure streaming. Finally, we draw our conclusions in Section 9.

2 H.264/SVC Content Adaptation

H.264/SVC is the scalable extension of H.264/AVC [17]. A major design requirement for SVC has been backward compatibility with the existing H.264/AVC standard. An SVC stream contains an H.264/AVC compatible base layer and one or more enhancement layers, each of which may augment the user experience in one of three dimensions (temporal, spatial, quality). A stream is temporally scalable if it contains substreams with a lower frame rate. Spatial scalability offers different video resolutions, contained in different substreams. Also spatial scalability with arbitrary resolutions is supported in SVC. A stream is quality-scalable if it contains substreams with different qualities (in an SNR sense) but equal resolution. There are two options for quality scalability. Coarse Grained Scalability (CGS) only allows a few distinct quality layers in a bitstream, i.e., only few selected bit-rates are supported. Scalable bitstreams with finer granularity can be achieved via Medium Grained Scalability (MGS). MGS allows changing the quality of each video frame. The coded representation of a frame (also called picture) of the video sequence is called access unit (AU). An AU consists of one or more network abstraction layer (NAL) units. For an in-depth discussion of SVC, the reader is referred to [47].

2.1 SVC Adaptation

H.264/SVC offers scalability in the temporal, spatial, and quality dimensions. In order to identify to which scalability layers a NAL unit belongs, H.264/SVC defines a header extension.

Listing 1. NALU header with SVC extension

0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
F NRI				Type				R I				PRID				N		DID		QID		TID		U D O		RR					

The SVC header extension is not used in the NAL units of the AVC compatible base layer. Thus in front of each AVC NAL unit a Prefix NAL unit signals its scalability information. The SVC NAL unit header (see Listing 1) signals to which specific scalability layer a NAL unit contributes. The most interesting fields for adaptation are the following:

- The *Priority Id (PRID)* can be used in an application-specific manner to signal the importance of a specific NAL unit.
- The *Dependency Id (DID)* signals the spatial or CGS quality layer.
- The *Quality Id (QID)* denotes the quality layer of an MGS NAL unit.
- The *Temporal Id (TID)* provides information on the temporal layer.

Adaptation mechanisms utilize these fields to adapt the video to the requirements of a client. In addition, Supplemental Enhancement Information (SEI) messages can be used to support the adaptation process. The Scalability SEI message supplies aggregated information on the layers of the media stream, such as resolution, frame rate and bit rate.

2.2 In-Network SVC Content Adaptation

H.264/SVC enables easy adaptation: parts of the bitstream can simply be dropped. We showed in previous work [32,20] how in-network SVC adaptation can be implemented using RTSP/RTP [37,36]. The main component is an RTSP signaling-aware RTP mixer [36], in line with the MANE concept defined in [45].

Our SVC adaptation MANE (Figure 3) acts as an RTP mixer, which receives and delivers the video data in a single unicast RTP stream. It receives the RTSP request from the client and creates a new RTSP request for the actual RTSP/RTP server. The server returns a description of the RTP streams utilizing the SDP protocol [11]. Based on the RTSP session and SDP information, a new state is created on the MANE, which is used in the RTP mixing process. The mixing includes full de-packeting of the incoming RTP streams and processing/adaptation on bitstream level. After adaptation, the bitstream is packetized with a new SSRC and delivered to the client. Thus, the actual processing/adaptation is performed on the application layer, not on the network layer. In the following, the components of the architecture are introduced.

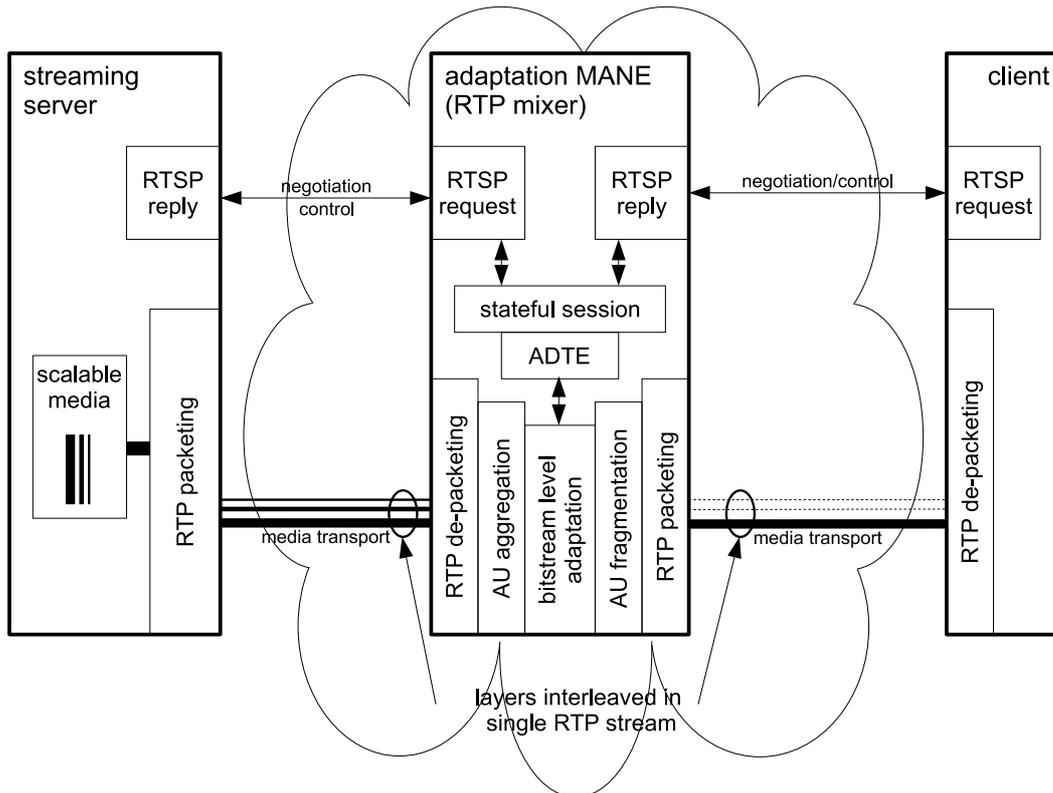


Figure 3. Adaptation-enabled MANE based on RTSP/RTP

The *bitstream level adaptation* component is exchangeable and enables easy replacement of the adaptation mechanism. It is steered by the *adaptation decision taking engine (ADTE)*, which supplies the information how to actually adapt the media bitstream. The *RTP packetizer and de-packetizer* implement the RTP stack by utilizing the payload format for H.264/SVC [46]. The *access unit (AU) aggregator and fragmenter* is needed to be compliant with the RTP marker bit semantic. Only the last packet of an AU should have the marker bit set, so in general it has to be updated after adaptation. Before packetizing the adapted AU, the AU has to be fragmented into pieces that the packetizer understands. In case of H.264 these are NAL units. For more details, the interested reader is referred to [32].

3 Encryption of H.264/SVC

The proposed SVC-specific encryption scheme recommends itself for three reasons: seamless integration into the H.264/SVC standard, potential format-compliance, and simplicity. The proposed scheme integrates well into the H.264/SVC standard. It could even become a part of the standard, simply by employing a reserved instead of an unspecified NAL unit type to signal encrypted data. SVC-specific encryption offers efficient format-compliant en-

encryption of H.264/SVC. As a consequence, transparent encryption can be implemented as well. In general, it is not trivial to design format-compliant encryption schemes. Most format-compliant encryption schemes therefore modify some routines in the compression pipeline. Note that these schemes have the disadvantage that the computationally complex compression has to be conducted for encryption! This is a fundamental drawback if pre-compressed content has to be securely distributed, e.g., in the application scenario of IPTV and VoD. In general, proving format compliance is tedious, as every syntactical and semantical requirement of the standard has to be validated. Even if a scheme is proven to be format-compliant, the uncommon encrypted bitstreams may crash some decoder implementations. As our proposed scheme simply ensures that encrypted NAL units are discarded, it is unlikely to crash decoder implementations. Due to the simplicity of the SVC-specific encryption scheme, implementation and integration are lightweight.

3.1 SVC-Specific Encryption

In order to preserve the scalability of the SVC stream, the SVC NAL unit headers have to be preserved, as these contain information regarding the scalability of the bitstream, i.e., the information to which dependency layer, temporal layer and quality layer a NAL unit contributes.

Unspecified NAL unit types (NUT) can be employed to signal encrypted data [39] format-compliantly, as compliant H.264/SVC decoders have to ignore NAL units with an unspecified NUT value. This behavior is the reason why H.264/SVC bitstreams are still valid H.264/AVC bitstreams (SVC NAL units are simply skipped). As encrypted NAL units with an unspecified NUT have to be ignored, only the remaining set of unencrypted NAL units has to be format-compliant. This method guarantees a strictly defined decoder behaviour. Note that unspecified NAL unit types will never be used within the H.264 standard suite. In [39], a direct mapping to unspecified NAL unit type values is defined for the most frequent NAL unit types (NUTs 1, 5, 14, 20). For all other NAL unit types, the original NAL unit header is preserved as the first payload byte and a specific unspecified NAL unit type is used to signal these encrypted NAL units.

The packetization of RFC 3984 [45] and the draft RFC defining the RTP payload for SVC video [46] assigns all but one (NUT 0) of the unspecified NUT values a specific meaning. Hence, the only possibility to employ unspecified NAL unit types to signal encrypted data is NUT 0, which is done in this paper.

In order to achieve format compliance of the encrypted bitstream, an appropriate set of NAL units has to be selected for encryption. If, for example,

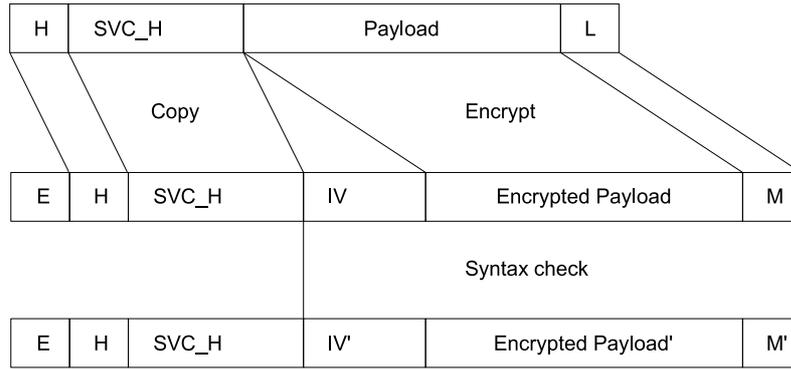


Figure 4. SVC-specific encryption

only a SPS NAL unit (sequence parameter set) is encrypted, the resulting bitstream will not be format-compliant, as the SPS NAL unit is required for decoding. Thus this scheme makes format-compliant encryption possible, but care has to be taken in the selection of the NAL units to be encrypted. As format compliance may be preserved with this scheme, transparent encryption and the interleaving of encrypted and unencrypted parts is possible.

An improvement compared to [39] is the application of AES in Counter mode instead of the ECB mode. We employ a separate initialization vector (IV) for every single NAL unit. Thus every encrypted NAL unit can be decrypted independently of all other NAL units. Independent decryption of the NAL units is crucial, as adaptation or transmission errors may remove NAL units from the bitstream. The issue of IVs is discussed in Section 3.1.2.

The proposed scheme is outlined in Figure 4. In the following the SVC-specific encryption algorithm is described:

- (1) Prefix the encrypted NAL unit with NAL unit header with NUT 0 (labelled E in Fig. 4).
- (2) Copy the original NAL unit header (H).
- (3) If an SVC extension header is present: Copy the SVC extension header (SVC_H).
- (4) Optionally insert an IV.
- (5) Format-compliantly encrypt the NAL unit payload (see Sect. 3.1.1).

3.1.1 Format-Compliant Payload Encryption

Although the NAL units with an unspecified NUT have to be ignored by a compliant decoder, certain syntax requirements have to be met by every NAL unit payload (in order to achieve format compliance). Otherwise the encrypted NAL units may affect the decoding process. E.g., when encrypting NAL units with conventional encryption, start code prefix patterns (within an encrypted

NAL unit) may be generated that indicate the start of a NAL unit in the H.264 byte stream format. The parsing of NAL units will be incorrect and decoding the bitstream impossible. Therefore it is crucial that the encrypted NAL unit data meet the syntax requirements for NAL units.

The syntax requirements for NAL units are given in [17]: within the NAL unit, the following three-byte sequences shall not occur at any byte-aligned position:

- 0x000000
- 0x000001
- 0x000002
- 0x000003

Within the NAL unit, any four byte sequence that starts with 0x000003 other than the following sequences shall not occur at any byte-aligned position:

- 0x00000300
- 0x00000301
- 0x00000302
- 0x00000303

Additionally, the last byte of a NAL unit shall not be 0x00.

The encryption scheme has to ensure that these requirements are met. Therefore, after encryption the procedure for the encapsulation of an SODB (string of data bits) within an RBSP (raw byte sequence payload) [17] has to be applied. For the case of two consecutive 0x00 bytes, this procedure ensures that the NAL unit does not end with a 0x00 byte. In H.264/SVC all NAL units that end with a 0x00 byte also end with two consecutive 0x00 bytes. Straightforwardly encrypted NAL unit payloads do not have this property and the encapsulation procedure will not work.

Thus special care has to be taken with the encryption of the last byte of a NAL unit. In our approach we use AES in Counter mode and treat the last byte with special care. In the following, it is assumed that a keystream is available.

Every cipher byte, except the last one, is the plaintext byte XORed with a keystream byte. The last cipher byte c is derived from the plaintext byte p and a keystream byte k (optimally in the range [0x00,0xfe], which can be ensured by ignoring 0xff bytes from the keystream) in the following way:

$$c = (p - 1 + k) \bmod 0xfe + 1$$

For decryption the following procedure is applied:

$$p = (c - 1 - k) \bmod 0xfe + 1$$

Afterwards the encapsulation procedure is applied to the encrypted NAL unit payload. Thus format-compliance of the NAL unit payload is achieved.

The format-compliant NAL unit encryption algorithm is outlined in the following:

- (1) Initialize the counter with the IV.
- (2) Encrypt the counter.
- (3) For all but the last byte of the NAL unit payload:
 - (a) For every encrypted counter byte:

Xor the next NAL unit payload byte with the encrypted counter byte.
 - (b) Increment and encrypt the counter.
- (4) The last byte is specifically encrypted, namely $c = (p-1+k) \bmod 0\text{xf}\text{e}+1$. In Figure 4 the last NAL unit payload byte is denoted L, the specifically encrypted byte M.
- (5) Apply the encapsulation procedure to the encrypted NAL unit payload. In Figure 4 the syntax-checked and corrected elements are suffixed with a prime.

3.1.2 Initial Vector Construction

For Counter mode a unique IV is needed for every NAL unit which is encrypted in order to be capable of decrypting every NAL unit independently. This is necessary as adaptation may remove NAL units from the bitstream. Even if no adaptation takes place, RTP packets and the contained NAL units may be lost during transmission. The loss of synchronization in the Counter mode renders the remaining encrypted data useless since it cannot be decrypted. A straightforward solution is to explicitly add the IV for every NAL unit in the bitstream. This can be simply done by appending the IV after the preserved NAL unit header. This scheme is the most robust one: every encrypted NAL unit can be decrypted without any side information (except the decryption key). The IV for a NAL unit can be freely chosen at encryption, the only prerequisite is its uniqueness (in combination with the same encryption key). A solution for the structure of the IV is to split the 128 bit of the IV for AES into two parts, the first is used for a NAL unit counter, while the remaining bits are sequentially incremented for every byte. A split into two 64 bit parts allows the generation of a unique IV for 2^{64} NAL units each consisting of a maximum of 2^{64} bytes (exceeding the expected number of NAL units and their expected sizes by far). In Section 6 experimental results show that the negative influence of adding the IVs on compression performance is limited.

Nevertheless, there are more efficient solutions in terms of compression performance, but they are achieved at the expense of decreased robustness. The

construction of adaptation-invariant IVs is not a trivial task. In an access unit, the NAL unit header and the SVC extension header are usually unique for each NAL unit. Thus an adaptation invariant IV can be generated in dependence of the current access unit count (in decoding order), the NAL unit header, and if available the SVC extension header. There may still be NAL units in an access unit for which these data items are not unique, and it is therefore necessary to include a NAL unit counter per access unit for those in the IV construction. The access unit count can be derived from the bitstream, but adaptation and transmission errors (frame dropping) can change the number of access units. If special prediction structures, such as hierarchical B-pictures with a restricted GOP size, are used, the access unit count can still be determined after temporal adaptation. For example, if the GOP consists of an IDR picture, followed by 32 hierarchically predicted non-IDR pictures, the decrypter can assume, if only an IDR picture and 16 non-IDR pictures are received, that the highest temporal level has been removed. The access unit count can still be determined. However, burst errors in the transmission may lead to a loss of synchronization, as entire GOPs may be lost. It is therefore reasonable to add the current access unit count to the bitstream at the start of every GOP.

3.1.3 Selective/Partial Encryption

There are several reasons for applying selective/partial encryption. One reason is enabling transparent/perceptual encryption; another one is improving the runtime performance. In this paper, we evaluate the second aspect, i.e., whether selective encryption may reduce the computational effort. Previous work has shown that the expected performance gains of selective encryption are negligible compared to the overall complexity if compression is taken into account [40,8]. The encryption of the base layer renders the entire bitstream corrupt and no longer trivially decodeable. The encryption of the IDR frames has a similar effect. Even if this data were replaced such that decoding is possible, the resulting video quality might be sufficiently low. Thus, these selective/partial encryption schemes can be considered instead of full SVC-specific encryption. However, the security of these schemes is not subject of in-depth discussion in this work. It is investigated whether further research into the assessment of the security of selective/partial encryption of SVC is reasonable, i.e., if there are possible performance gains (a frequent argument for selective/partial encryption). The performance of two selective/partial encryption schemes (base layer and IDR frame encryption) implemented with the SVC-specific encryption scheme is evaluated in Section 7.

3.2 Secure Real-Time Transport Protocol (SRTP)

SRTP [5] extends the RTP audio/video profile [35] by adding security features such as encryption, authentication, and replay protection. It is located between the RTP application and the transport layer. The sent RTP packets are intercepted by SRTP and transformed into SRTP packets. After receiving and processing the SRTP packets, the resulting RTP packets are forwarded to the application. Secure RTCP (SRTCP) is also handled in this manner. The encryption does not cover the whole packet, because it would render the RTP packet useless. Only the payload data of the RTP packet is encrypted with one of the pre-defined encryption transforms. The authentication comprises the whole packet and allows to identify whether the content really belongs to the RTP session.

As we do not deal with the topics of integrity and authentication in this work, we only apply SRTP encryption. SRTP employs AES for encryption. One of the two modes of operations is Counter mode, the other one is a variant of the Output Feedback mode (OFB). We employ the Counter mode of operation: an IV is generated on the basis of a salting key, the SSRC field and the packet index. The last 16 bits are subsequently incremented to obtain the next counter.

Apart from the IV generation, the encryption routine (AES in Counter mode) is the same for SRTP and the SVC-specific approach (see Section 3.1).

4 In-Network Adaptation of Encrypted H.264/SVC Content

Multimedia adaptation is driven by metadata describing the media characteristics. In case of H.264/SVC, the metadata is contained within the NAL unit and the SVC extension header, as shown in Section 2.1. For in-network adaptation of SVC, these header have to be accessible.

In an in-network adaptation scenario for encrypted SVC content, there are two alternative ways of accessing this information:

- The information on how to adapt the encrypted SVC content has to be supplied in an unencrypted form. This is the case in the SVC-specific encryption scheme presented in Section 3.1, where the NAL unit header and the SVC extension header are available in plaintext.
- The MANE has to be within the security context [25] (in case of full encryption like in SRTP), such that encrypted video data can be decrypted before adaptation.

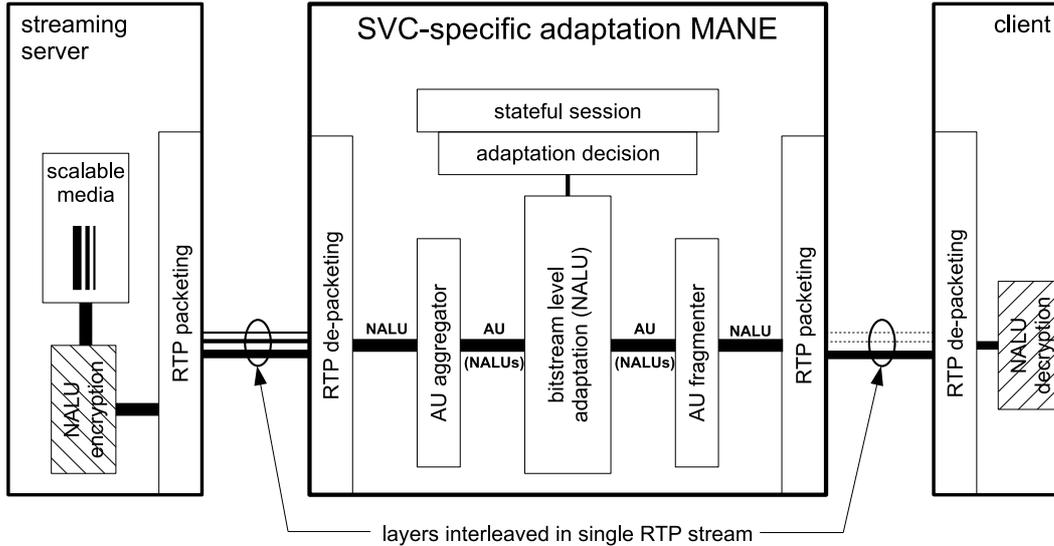


Figure 5. SVC-specific adaptation system

In Sections 4.1 and 4.2, we describe two in-network adaptation systems realizing the two alternative ways:

- An SVC-specific system which implements encryption at the application layer.
- An SRTP-based system, where encryption is located between the RTP stack and the transport layer.

A qualitative comparison of these systems is finally given in Section 4.3.

4.1 SVC-Specific Adaptation System

For our investigations into SVC-specific encryption, we are able to use the adaptation system presented in Section 2.2 with small modifications. On the server, NAL units are specifically encrypted and afterwards packetized in a standard manner (see Figure 5). The adaptation MANE identifies an encrypted NAL unit by its NAL unit type (see Section 3.1) and extracts the original NAL unit header and its extension (if present). This enables us to handle encrypted NAL units in exactly the same way as unencrypted ones, because only the NAL unit header is processed in the adaptation process. At the client, the encrypted NAL units are depacketized and decrypted. Note that in Figure 5 the signaling components were omitted for the sake of clarity.

Our approach to H.264/SVC content adaptation is based on NAL units. The NAL unit stream derived from the de-packetizer is aggregated into access units (AUs). Encrypted NAL units are detected and the original NAL unit headers are extracted. After aggregation, each NAL unit of the AU is adapted.

The adaptation decision process depends on several SVC-specific parameters describing which parts of the bitstream should be kept and which should be filtered out. By simply examining the header fields of a NAL unit, it is possible to adapt the bitstream. NAL units not matching the parameters of the adaptation decision are consecutively removed. After adaptation, the AU fragmenter analyzes the remaining NAL units in such a manner that the integrity of the marker bit is preserved. The resulting NAL units are forwarded to the packetizer and sent to the client.

This solution relies only on the metadata provided by the (always unencrypted) NAL unit header. Hence, no decryption has to be conducted on the MANE, which results in almost no processing overhead.

Pre-encrypted content: Media-aware encryption, like presented in Section 3.1, enables pre-encryption of content. The video needs to be encrypted only once, e.g., by the content provider, but not on the server which packetizes and streams the content to the clients. This is possible because the SVC-specific encryption scheme does not lead to changes in packetization. When utilizing pre-encryption, NAL unit encryption on the streaming server (hatched area in Figure 5) is omitted, which results in a significant reduction of server load. The remaining parts of the adaptation system still work as described above.

4.2 SRTP-Based Adaptation System

In SRTP, each RTP packet is encrypted completely and encapsulated in an SRTP packet. The RTP packets are generated by the packetization of the H.264/SVC bitstream. The complete media stream has to be decrypted before adaptation can take place on the MANE (see Figure 6). The received SRTP packet is decrypted into an RTP packet, which is used by the de-packetizer to reconstruct the NAL units.

After decryption, the adaptation MANE follows the same principles as the MANE described in Section 4.1 (fine-hatched area in Figure 6). The adapted NAL unit stream is packetized into RTP packets and encrypted again with SRTP. At the client, the SRTP packets are decrypted and forwarded to the de-packetizer, which assembles the bitstream, i.e., the NAL units.

4.3 Comparison of SVC-Specific and SRTP-Based Encryption and Adaptation

In the following the major qualitative strengths and weaknesses of these two approaches are summarized.

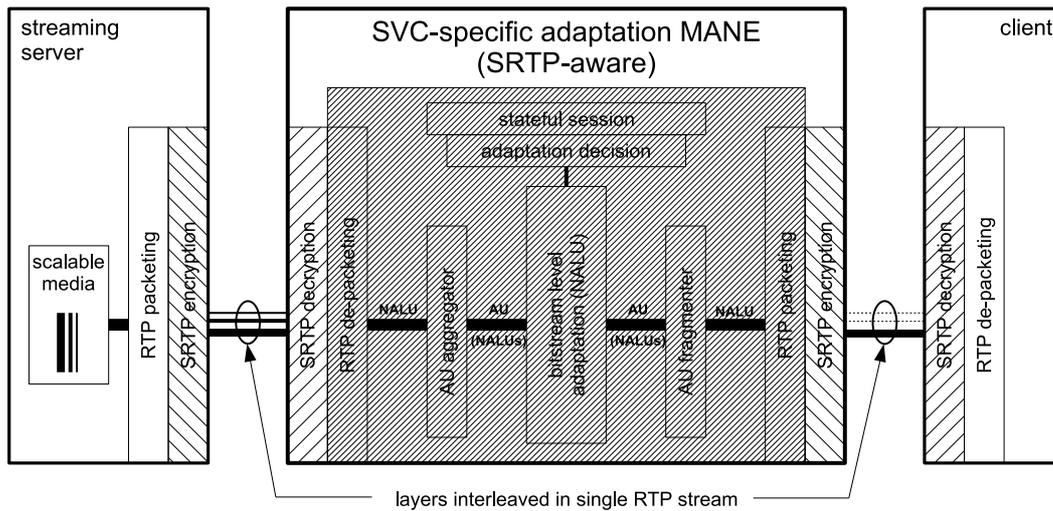


Figure 6. SRTP-based adaptation system

SVC-Specific Encryption and Adaptation

Strengths:

- The encryption is transparent for the transport mechanism.
- The method supports selective encryption, e.g., base layer encryption.
- Streaming of pre-encrypted video content is possible.
- Support of format-compliant encryption is provided, e.g., for transparent encryption.
- The technique enables adaptation without decryption on the MANE.

Weaknesses:

- The technique is specific to H.264/SVC and can not be generalized for other standards.
- Format compliance might introduce overheads in compression performance and encryption effort.
- The NAL unit header and its SVC extension are available in plaintext.

SRTP-Based Encryption and Adaptation

Strengths:

- The encryption is independent of the compression standard. Only the packetization for RTP depends on the compression standard.
- This is a standardized approach, which is deployed in many systems.
- The content is fully encrypted.

Weaknesses:

- The MANE has to be within the security context to enable in-network adaptation.
- High processing effort is induced on the MANE due to full en-/decryption.
- SRTP preserves the RTP packet lengths and thus a fingerprint of the SVC

bitstream.

5 Security Analysis and Evaluation

Although we consider the integrity of the video data as interesting and important, we do not consider security with respect to integrity in this work. Consequently we limit the capabilities of a potential attacker to a read access to the transmission channels anywhere in the distribution chain.

Firstly we point out that for SRTP the MANE has to be within the security context of the delivery. Thus the MANE may leak key information, while for SVC-specific encryption only the content provider and the client need to have access to the key.

Multimedia encryption may have an entirely different aim as opposed to maximum confidentiality or privacy in the context of certain application scenarios, e.g., as in the case of transparent encryption. We can summarize distinct application scenarios and requirements for multimedia encryption as follows:

- Highest Level Security / Cryptographic Security
Applications that require a very high level of security, no information about the plaintext (compressed video stream) shall be deductible from the ciphertext.
- Content Security / Confidentiality
Information of the plaintext may leak, but the video content must not be discernible.
- Sufficient encryption / Commercial application of encryption
The content must not be consumable due to the high distortion (DRM systems).
- Transparent / Perceptual encryption
A preview image needs to be decodeable, but the high quality version has to be hidden. Another application is privacy protection.

In fact the entire notion of security depends on the application context.

If we consider IPTV, the TV program is publicly known. Thus highest level security is not an issue, as meta-information of the content (title, short description, ...) are publicly distributed. In this scenario the indistinguishability of encrypted video streams is not of concern. However, in the case of the VoD application scenario, the privacy of a customer is violated if an attacker can derive information about which video stream is being transmitted by monitoring the network packets. Thus it is of interest to evaluate whether one of the proposed encryption schemes, namely SRTP or SVC-specific encryption,

can meet the requirements of highest level security.

The IPTV scenario does not require the highest level security, content security or sufficient encryption are adequate. As both discussed encryption schemes, the SVC-specific encryption scheme and SRTP, are capable of the encryption of all of the video data, content security is achieved. In both schemes a secure encryption primitive, i.e., AES, is employed. Thus the recovery of the encrypted video data is infeasible. In the SVC-specific encryption scheme, only the NAL unit header, the SVC extension header, and the NAL unit length are preserved. Any reconstruction of the content on the basis of header data is impossible for H.264/SVC. This not the case for all standards, e.g., coarse reconstructions of the image content are possible on the basis of JPEG2000 packet headers [7,8]). Naturally the weaker requirements of sufficient encryption can be met too.

If the SVC-specific encryption is applied in a partial/selective fashion, namely the encryption of the base layer or the IDR frames, it is assumed that these schemes can not provide content security. Although currently no attacks against these two partial/selective encryption schemes are known, specific attacks similar to those outlined in [34,8] are probable. However, sufficient encryption may be in the scope of the partial/selective encryption schemes.

Transparent encryption can only be implemented with the SVC-specific approach. Its implementation is straightforward, as only the enhancement NAL units of the publicly available base layer are encrypted. For the security analysis we have to consider specifically tailored attacks [34] that aim at improving the quality of the publicly available base layer. As all the enhancement information are encrypted, the side-channel information available to an attacker are limited. Thus attacking transparent encryption with the SVC-specific approach is very similar to improving the video quality of the base layer without any side-channel information.

In the following we answer the question whether the SVC-specific and the SRTP-based encryption scheme are secure with respect to confidentiality in a strong cryptographic notion (highest level security). Modern security notions are equivalent to the property of ciphertext indistinguishability under a chosen plaintext attack (IND-CPA) [1]. Given two plaintexts and a ciphertext (randomly chosen from the two plaintexts), an adversary can not identify the plaintext with a probability better than $1/2$. If one can link a plaintext and a ciphertext, the scheme is not secure under IND-CPA. In our VoD application scenario, security with respect to this notion of security ensures that an attacker cannot identify which video sequences are transmitted, e.g., which movies are watched by whom. Therefore, we have analyzed the length of the network packets (packet trace) of the encrypted SVC streams and the plaintext SVC streams. If the packet traces enable us to link a ciphertext SVC

stream and a plaintext SVC stream, the corresponding encryption scheme is not secure under IND-CPA. SRTP preserves the packet length of the RTP packet, a fact pointed out in the corresponding RFC [5]. As RTP packetization as specified in [46] is strongly dependent on the SVC NAL units, a certain fingerprint of a bitstream is preserved in the encrypted domain for both RTP-packetized SVC-specifically encrypted bitstreams and SRTP SVC streams. In the following, we evaluate the discriminative power of this fingerprint.

As it is infeasible to evaluate the fingerprint for a huge number of full-length video sequences, it is evaluated for a reasonably sized set of very short sequences (582 sequences with 8 frames). As the fingerprint of the packet traces is shorter for short sequences, it compensates for the limited number of sequences. The assumption is that if we can link a ciphertext and a plaintext for a limited number of short sequences, it is possible to link ciphertexts and plaintexts for a larger number of longer sequences (e.g., all commercial movies).

5.1 Test Content

For our security evaluation, we have split well-known CIF sequences (Akiyo, Bus, City, Coastguard, Container, Crew, Flower, Football, Foreman, Harbour, Ice, Mobile, News, Silent, Soccer, Tempete, Waterfall) into non-overlapping subsequences of 8 frames. This results in 582 distinct sequences, of which some are very similar, e.g., the subsequences of the Akiyo sequence. The bitstreams were generated using the Joint Scalable Video Model (JSVM) [31] 9.14 software. The encoder configuration has been chosen to meet requirements of a VoD system; the scalable bitstream contains a QCIF substream (compliant to the H.264 baseline profile) and two CIF MGS layers to enable bitrate adaptation. For practical reasons (excessive encoding time), higher resolutions have been omitted. In this work, we employ the same static encoder configuration for all our evaluations (Sections 6 and 7).

5.2 Similarity between Packet Traces

In order to assess the similarity between packet traces, the mean squared error of the packet lengths is considered. Other similarity measures, such as correlation, can be considered; however, the only goal is to link a ciphertext to a plaintext via the similarity measure. If this works, the measure is sufficient. If the number of packets differs between two packet traces, the MSE is calculated for the smaller number of packets. Thus, the difference between two packet

traces pt_1 and pt_2 is defined in the following way:

$$d(pt_1, pt_2) = \sum_{i=1}^{\min(n_1, n_2)} (l_{1i} - l_{2i})^2$$

The overall number of packets for pt_1 is n_1 and the corresponding packet lengths are denoted by l_{1i} and similarly for pt_2 .

5.3 Evaluation

For the different encryption modes (SVC-specific encryption and SRTP) and every compressed subsequence a packet trace has been generated. The question is whether this information is discriminative enough to enable an attacker to identify a subsequence. For a strong notion of security, even a slight advantage is a security breach. In our evaluation, we compared each packet trace with every other packet trace, i.e., $582 \times 2 \times 582 \times 2$ comparisons. It turned out, despite very similar subsequences, that each packet trace was unique for both the SVC-specific and the SRTP encryption. No subsequence that was encrypted with either scheme has the same fingerprint.

Moreover, Figure 7 indicates that this fingerprint is not just unique, but also a weak measure for sequence similarity. In Figure 7 we have evaluated the MSE between the packet trace of the first subsequence from the Akiyo sequence with SVC-specific encryption and all other subsequences, even those encrypted with SRTP. There are 37 very similar subsequences of the original Akiyo sequence that reveal an obviously smaller distance to the reference packet trace. For our VoD application scenario this means that it is very likely that an attacker can identify the streamed and encrypted bitstream if the fingerprint of a sequence is known.

This kind of security breach is inherent in every secure and scalable scheme that chooses the scalable extraction points in dependency of the original source. These data (the scalable truncation points, i.e., the scalability information) are a distinctive fingerprint of the source data (as shown experimentally for SVC in this section). However, this fingerprint is leaked as well for the SRTP-based approach, as the size of the NAL units determines the size of the SRTP packets. These results indicate that schemes that offer both encryption and preservation of scalability do not meet the requirements of the highest level of security. This is due to the fact that the scalability information is mostly unique for a source sequence. It is unique for all the considered sequences.

As the requirements for the highest level security can not be met for both encryption schemes, the application of selective/partial encryption schemes

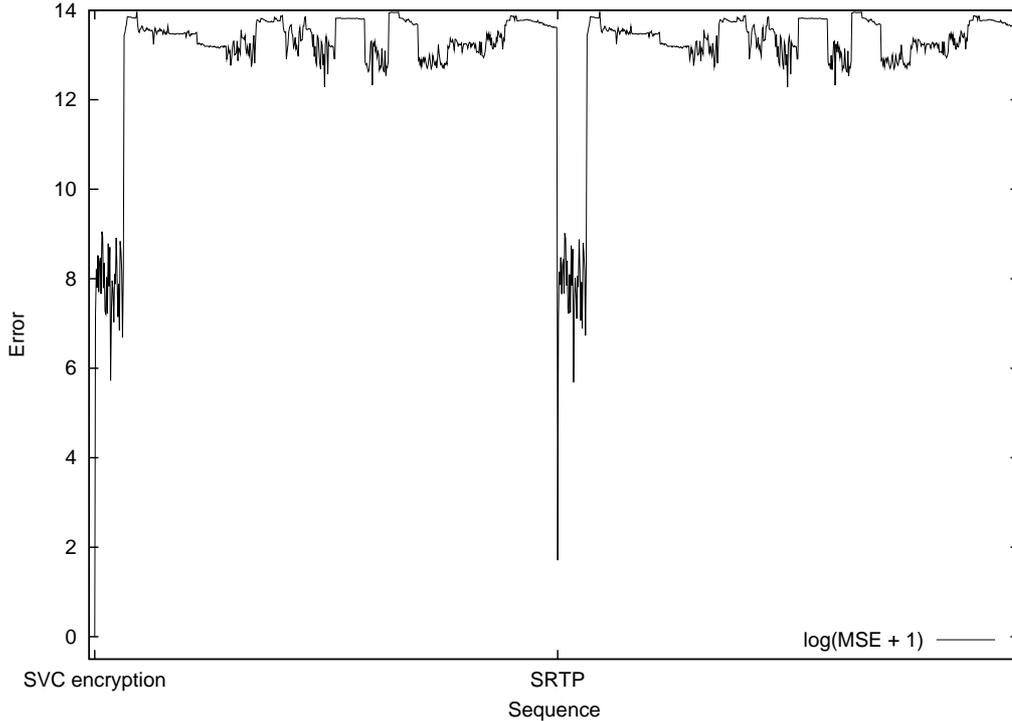


Figure 7. Similarity between packet traces

may be even more tempting. In Section 7, we clarify whether further research is justified due to significant advantages in the application, i.e., considerable performance gains.

6 Compression Evaluation

One aspect of the assessment of a multimedia encryption scheme is its influence on the compression performance. SRTP has no influence on the compression performance as no padding or additional headers are used. For the SVC-specific approach, an additional byte is used for every NAL unit. The resulting compression overhead is negligible, as can be seen in Table 1, which depicts the compression overhead for four well-known sequences (compressed with the same encoder settings as in Sections 5 and 7). Even if we explicitly add the IV for every NAL unit, the negative influence on compression performance is limited (see Table 1, “with IVs”). The examined bitstreams contain low resolution substreams (QCIF) and MGS enhancement layers for CIF. The compression overhead will be even lower for video content of higher spatial resolutions which will have bigger NAL units.

Table 1
Compression performance

Sequence	Relative overhead for SVC-specific encryption	Relative overhead with IVs
Foreman	0.21%	3.40%
Crew	0.14%	2.21%
Harbour	0.10%	1.65%
Football	0.08%	1.30%

7 Performance Evaluation

For a direct quantitative comparison of the SVC-specific encryption and SRTP-based in-network adaptation, the following metrics will be used:

- Transmission delays of coded pictures between server and clients, i.e., the delay between the start of processing (before encryption) of the last NAL unit of a coded picture at the server and its full delivery at the client (including decryption). This metric is to a large degree determined by the delays incurred by the adaptation MANE, because its processing is based on picture (AU) level.
- Load on the server and adaptation MANE, in terms of CPU usage.

These metrics are directly linked to the utility of the approaches in real streaming and adaptation systems. In order to achieve results of practical relevance, we have implemented streaming and adaptation prototypes using standard technologies as described below. The implementations are affected by operating systems and network behavior, e.g., context switches and socket processing efforts, as well as by complex scheduling effects on various levels, e.g., in the streaming server or in the multimedia and networking libraries employed.

7.1 Prototype Implementations

The goal of the evaluation is to compare the two video encryption schemes with respect to performance. Hence, only the RTP transport (data path) was implemented and evaluated. The signaling overhead (RTSP, RTCP) was regarded as identical for both encryption schemes and therefore omitted.

Our prototype implementations are based on standard open source streaming technologies, namely:

- libSRTP [6] for the AES [29] implementation and SRTP packet encryption.

- Live555 [22] modules for RTP transport and adaptation.

The streaming server is based on the Live555 RTP stack and packetizes H.264/SVC according to the recent RTP specification [46]. Live555 is also the basis for the adaptation MANE and a Live555 module is used to perform the actual adaptation. It is located between the Live555 RTP-source module (more or less an RTP client) and the RTP-sink module, which sends the adapted video stream to the end-user client. The end-user client is based on Live555's RTP-source module. All implemented servers, MANEs and clients have the same software basis.

7.2 Test Setup and Methodology

As all implementations share the same software basis, we can assume that the measurements indeed show the difference between the different encryption and adaptation schemes.

In addition to the MANEs described in Section 4, an *encryption-less system* was implemented to serve as a reference. It has the same adaptation capabilities as the SVC-specific system described in Section 4.1, but does not encrypt the H.264/SVC video stream. This MANE enables us to measure how much basic processing effort the RTP handling and adaptation processes will induce.

In summary, the following encryption and adaptation systems have been examined:

- Encryption-less system (*no encryption*).
- Selective SVC-specific encryption system with only base layer encrypted (*base layer*).
- Selective SVC-specific encryption system with only IDR frames encrypted (*IDR frames*).
- SVC-specific encryption system with all NAL units encrypted (*SVC encryption*).
- Secure RTP system (*SRTP*).

For delay measurements, we used a special method to retrieve accurate end-to-end delays. In general, network delay measurements are problematic, because of clock (de-)synchronization issues between the server and the clients. We solve this issue by running the streaming server and the client processes on the same computer (computer1 in Figure 8). The video is streamed from the server to the MANE (running on computer2, see Figure 8). The MANE performs adaptation and streams the adapted content to the client running on the same computer as the server. We can safely assume that the measured results will only be inferior to a solution with the server separate from the clients,

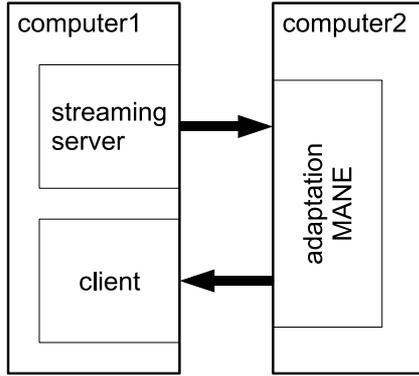


Figure 8. Evaluation setup

because the concurrency of the server and clients reduces the responsiveness of the operating system. The test setup consists of two DELL PowerEdge 1850 computers with two Intel Xeon 3.0 GHz EM64T processors with Hyperthreading disabled. In each computer, the main memory comprises 2 GB and the operating system is Ubuntu Linux 6.06.1 (dapper) with kernel 2.6.27.2 (x86 64). The computers are connected via an Intel(R) PRO/1000 Gb network card to a Gigabit Ethernet network switch.

For all measurements, 100 clients were receiving the same content from the streaming server via the adaptation MANE. There was no packet loss during transmission for all clients and measurements. The first 100 seconds were removed from the result data sets, because results in the startup phase may be overly optimistic. In the startup phase the number of clients is not constant, as the clients are started one after the other. The delay between the streaming server and the client is measured on a picture-by-picture basis. Before sending the NAL units of the picture and after fully sending/receiving/decrypting the last NAL unit of the picture a timestamp is recorded. This is done for each stream served by the streaming server and on each client. The test sequences were played in a loop, resulting in a total of 54000 pictures, i.e., 30 minutes at 30fps. The adaptation process at the MANEs was selected to let all NAL units pass. This is the worst case scenario, because all of the data has to be handled by the adaptation MANE and transmitted to the client. In the evaluations, this “pass-through” operation has been employed for all schemes in order to obtain comparable results.

7.3 Test Video Streams

In our performance evaluation of in-network adaptation we focused on four different well-known video sequences (*Foreman*, *Crew*, *Harbour*, *Football*). They were encoded using the settings defined in Section 5.1. The video streams used for the evaluations and their bit rates are described in Table 2. *Base layer de-*

Table 2

Video streams used for performance evaluation

Sequence	GOP size (IDR frame interval)	Bit rate in kbps	Base layer in kbps	IDR frames in kbps
Foreman	32	398	60	115
Crew	32	611	90	92
Harbour	32	822	90	195
Football	32	1045	158	133

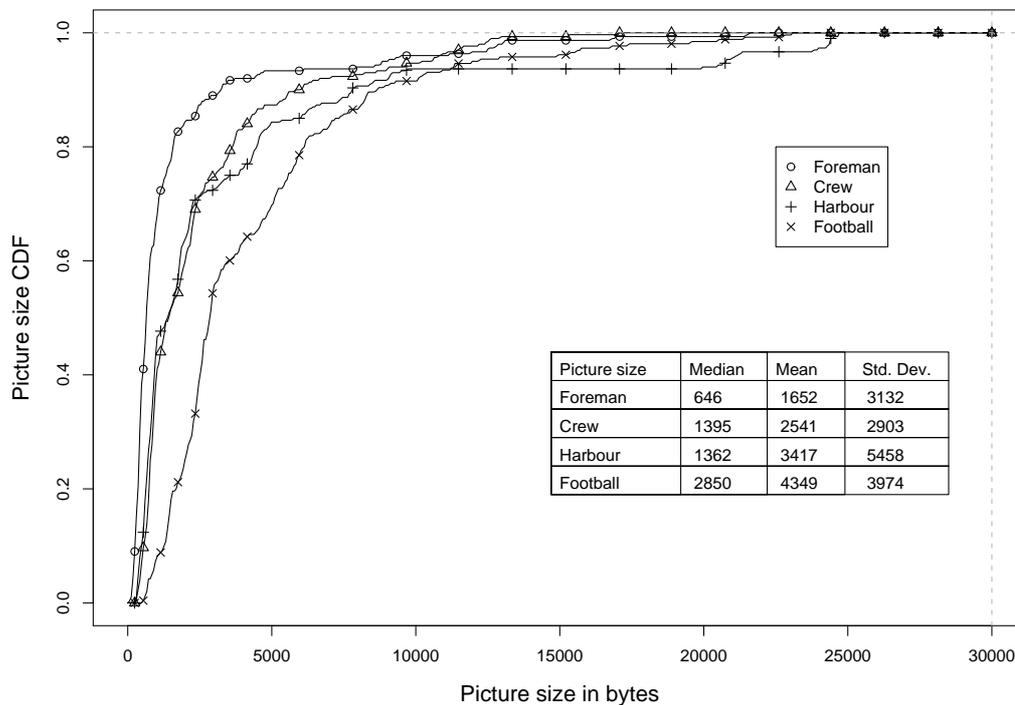


Figure 9. Coded picture size distribution

notes the bit rate of the H.264 compatible base layer, *IDR frames* the bit rate of the IDR frames with all scalability layers.

The transmission delay defined above depends on the coded picture size, which is the amount of processed data (encryption, transmission, decryption) per single delay measurement. The cumulative distribution function (CDF) of the coded picture sizes for each sequence (with all scalability layers) is presented in Figure 9. The distribution of the coded picture sizes is a result of the video characteristics (e.g., intra-picture complexity and inter-picture redundancy) and the predictive coding in H.264/SVC.

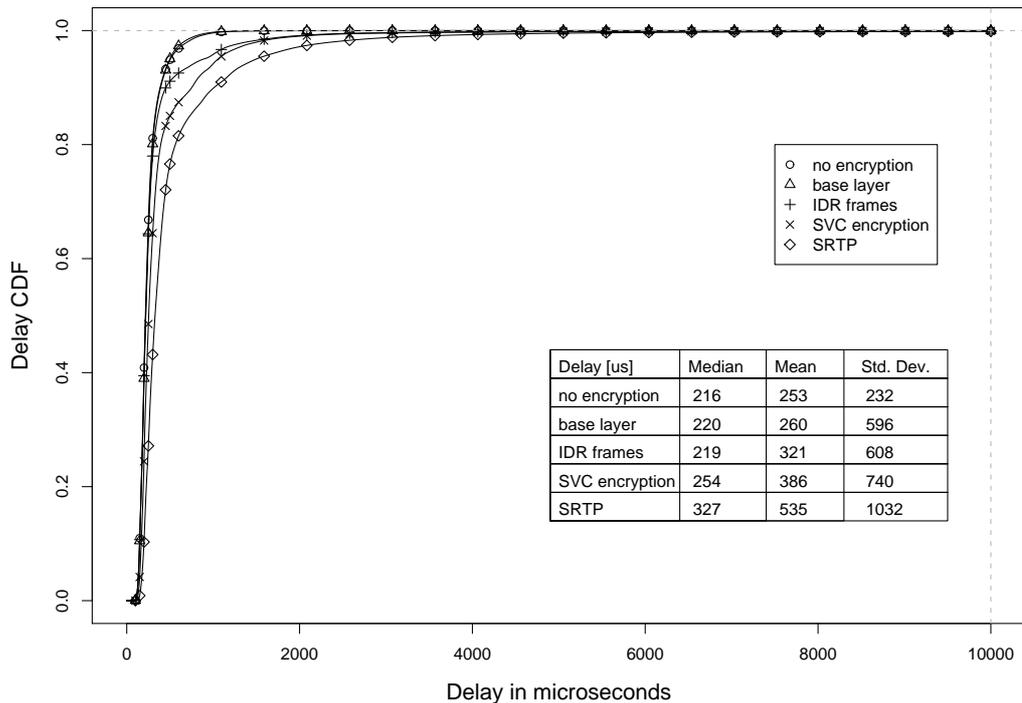


Figure 10. Cumulative delay distribution function for sequence *Foreman*

7.4 Evaluation Results

This quantitative evaluation shows the impact of the different encryption schemes on the transmission delay. In addition to the delay introduced by network transmission, the components responsible for most of the delay are en-/decryption, RTP de-/packetization and adaptation. Figures 10 and 11 show exemplarily that the delay increases with the degree of en-/decryption applied (by the investigated schemes), which is indicated by increasing mean and standard deviation values. The scalability (in terms of the number of served clients) of the different systems can be deduced from the load on the server and the adaptation MANE. This is shown in Figures 12 and 13, where the load for multiple streams is compared for the different encryption schemes. For the metrics (the delay and the load), an encryption-less system (*no encryption*) acts as an indicator of the basic effort needed for transmission and adaptation.

Figures 10 and 11 show that each adaptation system has a similarly shaped cumulative delay distribution function (CDF), which is basically a result of the picture size distribution. The different mean and median values and standard deviations are results of the different complexities of the encryption schemes. Not surprisingly, *no encryption* offers the lowest delay, followed by the two

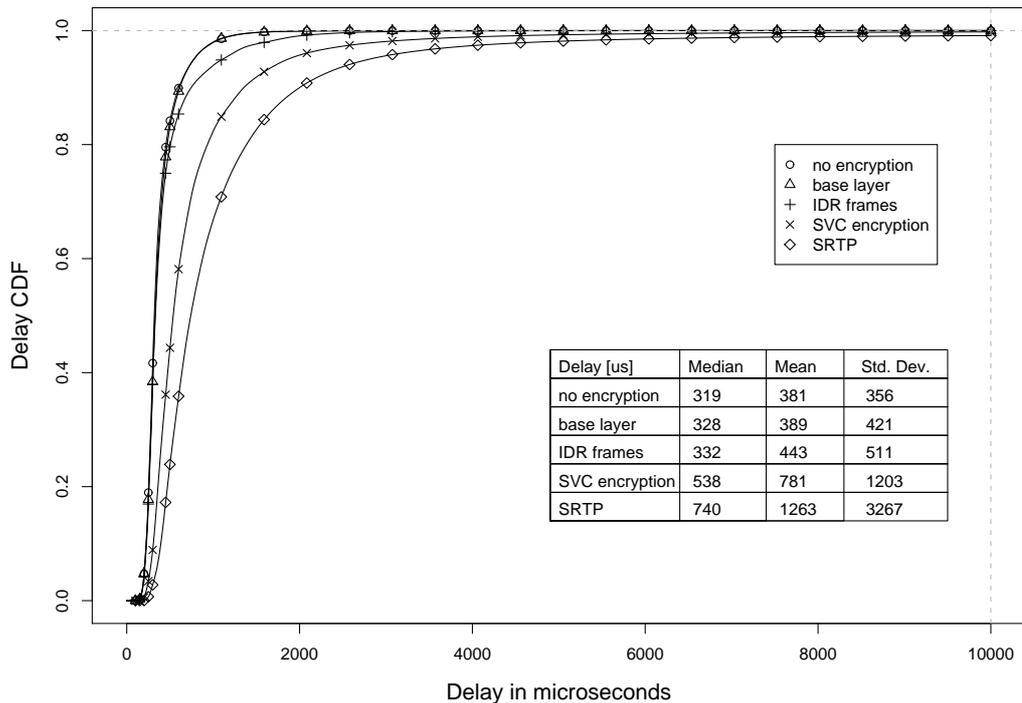


Figure 11. Cumulative delay distribution function for sequence *Football*

selective encryption schemes *base layer* and *IDR frames*. The full encryption schemes *SVC encryption* and *SRTP* are showing notably higher delays because of the higher processing effort. Compared to the other approaches, the delays of the *SRTP* approach are notably higher, mainly because of the required de-/encryption on the MANE. The delay jitter increases as well, as indicated by the standard deviation results.

The delay CDFs of the sequences *Crew* and *Harbour* were omitted because their shapes are similar to the delay CDFs of *Foreman* and *Football* and only differ in mean and standard deviation values. In direct comparison, the delay CDFs for the sequences *Foreman*, *Crew*, *Harbour* and *Football* reveal that the delay distribution is also directly dependent on the bit rate of the video: higher bit rates lead to increases of both delay and jitter.

When comparing *no encryption* and *base layer* encryption, it is notable that the delays differ only very little. Because the base layer portion of each frame is small, it does only slightly affect the delay (higher standard deviation). The *IDR frames* encryption system induces additional delay only to IDR frames, which are in general the largest ones. This results in larger mean and standard deviation values.

Figure 12 shows the CPU consumption (in the following denoted as load)

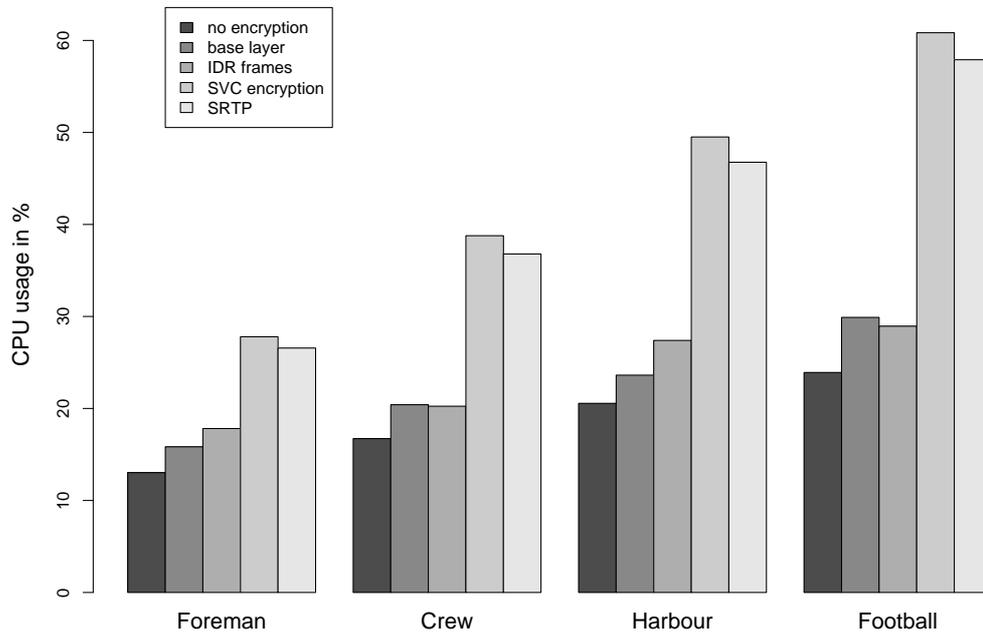


Figure 12. Average CPU usage on server

on the server, which obviously increases in general with the bit rate of the streamed content. In the case of *no encryption*, this is only due to the higher data throughput, but for *SVC encryption* and *SRTP* also the encryption effort adds to the CPU load. Because the bit rates of the encrypted parts in *base layer* and *IDR frames* are rather low, encryption does not lead to a significant load increase. When comparing the loads of the selective systems it can be seen that also in this case the CPU consumption is directly dependent on the bit rate of the encrypted stream (see Table 2). The underlying AES implementation is the same for all encryption schemes. Only the emulation prevention deployed in *SVC encryption* adds about five percent load relative to *SRTP* for our test sequences.

On the MANE, the SVC-specific encryption systems induce the same load (Figure 13) as the *no encryption* system, because the adaptation systems can operate on the unencrypted NAL unit headers. For the *SRTP* system, the decryption and re-encryption steps on the MANE additionally increase the CPU consumption. For the other systems, the load increases rather slowly with the bit rate. Depending on the test sequence (bit rate), the *SRTP* system induces a two to four times higher load on the MANE than the SVC-specific encryption systems.

In our IPTV and VoD application scenario, where the MANE may be a low-

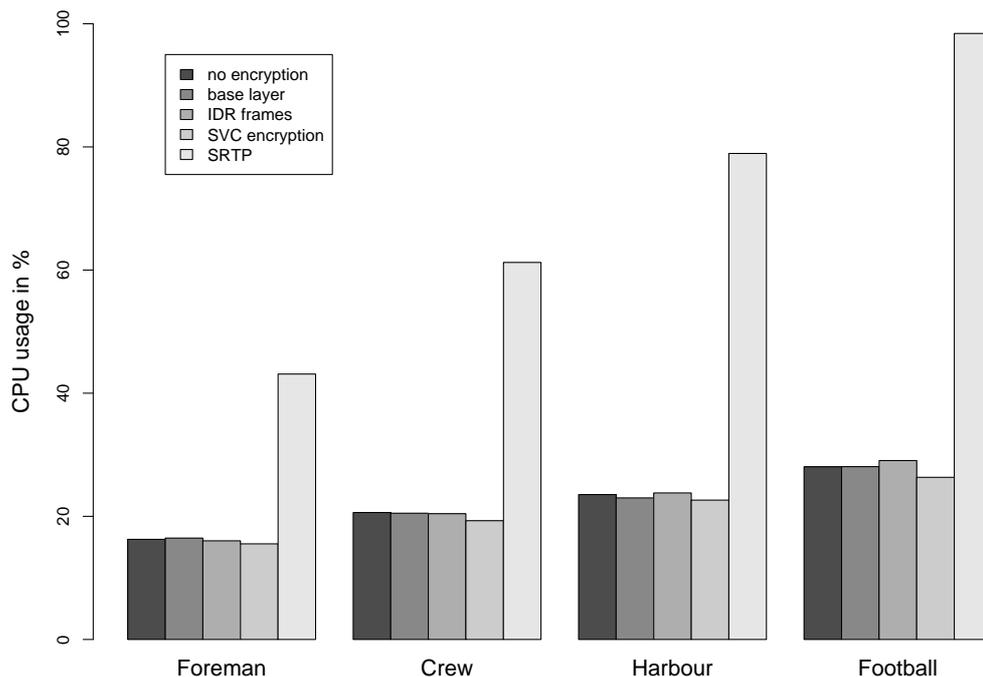


Figure 13. Average CPU usage on MANE

end, inexpensive device (e.g., wireless access point), computationally cheap adaptation mechanisms are required. The significantly higher load of the SRTP-based scheme is a decisive disadvantage for its deployment in an IPTV/VoD system, as compared to SVC-specific encryption and adaptation techniques.

8 Related Work

Related work originates in different fields.

The application of SRTP for secure multimedia transmission is discussed in a Cisco white paper on “Securing Internet Telephony Media with SRTP and SDP” [25], which gives a practical guide for applying SRTP [5] to IP telephony media. It presents an overview of IP telephony security as well as an in-depth view on transport level security with SRTP and its deployment.

An industry standard [2] for secure streaming of multimedia data has been published by the Internet Streaming Media Alliance (ISMA). It offers secure streaming solutions on the basis of both the MPEG-4 file format [14] and RTP. ISMA Encryption and Authentication (ISMACryp) [2] handles content encryption and message authentication/integrity services. Content delivery is based on the RTP protocol, signaling and control is realized by means of

the Session Description Protocol (SDP) [11] and the Real Time Streaming Protocol (RTSP) [37]. ISMACryp provides cryptographic metadata for each access unit and features, amongst others, selective encryption, key rotation and random-access. In [41], ARMS streaming (Adaptive Rich Media Secure) is introduced, which enables secure transmission of media data encoded in multiple independent streams. ARMS utilizes only stream switching, so no in-network adaptation actually takes place. The issue of secure adaptation on untrusted nodes is not dealt with in either ISMACryp or ARMS.

The latter topic, i.e., enabling adaptation on untrusted nodes within the network, has been subject of considerable research carried out at the HP Labs [43,42,44,4,28,27,3]. The term “Secure Scalable Streaming” (SSS) has been introduced in [43], where SSS has been investigated for wireless networks. The approach has been extended in [42]. The basic idea is to add control information (truncation points, prioritization information) to secured (encrypted) scalable packets in an unencrypted packet header. By truncating or dropping these packets, specific adaptations can be applied. The adaptation system is stateless, features low complexity, and enables adaptation without decryption. These efforts [44,4] also had great influence on the development of JPSEC, the Secure JPEG2000 standard [16], which enables secure scalable streaming for JPEG2000. JPSEC standardizes a secure meta file format exclusively for JPEG2000.

Further there are proposals which try to solve scalable streaming and security format-independently via application of a generic meta format, such as MPEG-21’s generic bitstream description language [28,27]. The application of gBSD has also been proposed specifically for H.264 [13]. These approaches, relying on the usage of gBSD, may lead to performance losses in terms of runtime and compression. In [32] it is shown that a gBSD metadata-driven adaptation solution performs significantly worse than an SVC-specific adaptation solution. An alternative metadata-driven approach can be rate distortion hints embedded in MPEG-4 [3]. Although these format-independent approaches have disadvantages in terms of runtime and compression performance, they offer conceptually clearer solutions which may outweigh their disadvantages.

On the other hand, there has been a considerable amount of format specific encryption proposals, including several for H.264/AVC: many of the previously proposed approaches implement encryption during compression, e.g., the scrambling of the intra prediction modes [12] or of motion vector data [21], the encryption of coefficient data and the perturbation of motion vectors [24], and the encryption of coefficient signs [30]. Most of the previous work on SVC encryption [19][48][3] is based on a draft standard that has significantly changed (e.g., in the meantime FGS has been removed). A compression integrated approach for SVC encryption is presented in [48]: sign encryption of “texture, motion vector, and FGS data” is proposed, and in [19] this idea

is extended to protect regions of interest. For all these approaches, the computationally demanding compression has to be conducted for encryption and decryption, which definitely presents a drawback. This disadvantage is argued to be compensated by the advantage of format compliance, i.e., that the encrypted video data is still a valid and decodeable H.264 bitstream. This is a functionality that our SVC-specific encryption scheme offers as well.

In [3] principles for SSS and SVC are discussed. It is pointed out that SSS packets include an unencrypted header which contains the scalability information. This is similar to our SVC-specific encryption scheme. Major attention is paid to the rate-distortion optimal truncation of SSS packets, a process that would require the support of FGS in SVC. The most common protocol for realtime multimedia transmission is RTP [36]. The SSS adaptation system is stateless [42]. This is not aligned with the RTP MANE concept, because within RTP it is not possible to statelessly identify payload types. Payload types are no longer statically defined but dynamically negotiated via RSTP. Thus, for the application of RTP, an RSTP signaling-aware, stateful MANE, as employed in this work, is a prerequisite. A second major difference is that our SVC-specific encryption scheme can offer format-compliant encryption, which offers the possibility to seamlessly integrate encryption and also meets the requirements of advanced application scenarios.

In [39] we have presented an approach that can be applied to both H.264/AVC and SVC. The original H.264 and SVC headers are preserved, conforming to the principles for SSS stated in [3]. In this work, we have adopted the main idea of the SVC-specific encryption scheme of [39], namely to format-compliantly signal the encrypted data. The approach has been improved with respect to the applied encryption mode. Furthermore the actual format-compliant encryption routines have been specified in detail and the construction of IVs is discussed.

For extensive overviews on multimedia encryption the interested reader is referred to [40] and [10].

9 Conclusions

In previous work [32], we showed how to enable RTP-compliant in-network H.264/SVC adaptation with a signaling-aware and stateful MANE. When using SRTP – the standard way of encrypting RTP-transported media – the MANE has to be within the security context. This leads to increased organizational and computational effort on the MANE and, moreover, to a computationally more expensive and less scalable adaptation system. The CPU load induced by the STRP-based system on MANE is two to four times higher than

that of the SVC-specific encryption and adaptation schemes. The end-to-end delays for SRTP are higher as well, because of the additional decryption and re-encryption steps to be done on the MANE.

Our SVC-specific encryption scheme offers almost the same encryption performance as SRTP, but offers far more flexibility, e.g., it supports selective encryption, pre-encryption of the content and SVC format-compliant encryption. The mechanism to enable SVC format compliance only introduces minimal performance overhead in the encryption process at the server and the decryption processes at the client. This shows that the additional syntax checks for SVC-specific encryption are inexpensive. However, in the adaptation process at the MANE the SVC-specific encryption scheme offers substantial performance gains. .

From a cryptographic point of view, both schemes leak information (the packet lengths) almost to the same extent. It is an application-dependent decision whether this information leakage is critical for its security. For many practical systems, such as IPTV, this information leakage does not impose a security threat. The security breach due to information leakage (network packet lengths) is alike for both schemes and poses a certain threat to the privacy of the communication (e.g., in the VoD application scenario).

With respect to runtime performance (delay and CPU usage), the proposed selective encryption schemes do make sense, as they offer almost the same performance as the no encryption case. However, the security of selective encryption schemes has to be researched; transparent/perceptual and sufficient encryption are within reach, while content confidentiality is assumed to be unachievable.

References

- [1] Martín Abadi and Phillip Rogaway. Reconciling two views of cryptography (the computational soundness of formal encryption). In *Proceedings of the International Conference IFIP on Theoretical Computer Science, TCS '00*, pages 3–22, London, UK, 2000. Springer-Verlag.
- [2] Internet Streaming Media Alliance. ISMA Encryption and Authentication Specification 2.0, Nov 2007.
- [3] J. Apostolopoulos. Architectural principles for secure streaming & secure adaptation in the developing scalable video coding (SVC) standard. In *Proceedings of the IEEE International Conference on Image Processing, ICIP '06*, pages 729–732, October 2006.
- [4] J. Apostolopoulos, S. Wee, F. Dufaux, T. Ebrahimi, Q. Sun, and Z. Zhang.

- The emerging JPEG2000 security (JPSEC) standard. In *Proceedings of International Symposium on Circuits and Systems, ISCAS'06*. IEEE, May 2006.
- [5] M. Baugher, D. McGrew, M. Naslund, E. Carrara, and K. Norrman. The Secure Real-time Transport Protocol (SRTP). RFC 3711 (Proposed Standard), March 2004.
 - [6] Cisco Systems, Inc. libSRTP: a library for secure rtp. <http://srtp.sourceforge.net/srtp.html>.
 - [7] Dominik Engel, Thomas Stütz, and Andreas Uhl. Format-compliant JPEG2000 encryption in JPSEC: Security, applicability and the impact of compression parameters. *EURASIP Journal on Information Security*, 2007(Article ID 94565):doi:10.1155/2007/94565, 20 pages, 2007.
 - [8] Dominik Engel, Thomas Stütz, and Andreas Uhl. A survey on JPEG2000 encryption. *Multimedia Systems*, January 2009.
 - [9] B. Furht and D. Kirovski, editors. *Multimedia Security Handbook*. CRC Press, Boca Raton, Florida, 2005.
 - [10] B. Furht, E. Muharemagic, and D. Socek. *Multimedia Encryption and Watermarking*, volume 28 of *Multimedia Systems and Applications*. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 2005.
 - [11] M. Handley, V. Jacobson, and C. Perkins. SDP: Session Description Protocol. RFC 4566, July 2006.
 - [12] Guang-Ming Hong, Chun Yuan, Yi Wang, and Yuzhuo Zhong. A quality-controllable encryption for H.264/AVC video coding. In *Proceedings of the Pacific Rim Conference on Multimedia, PCM '06*, volume 4261 of *Lecture Notes in Computer Science*, pages 510–517. Springer-Verlag, November 2006.
 - [13] Razib Iqbal, Shervin Shirmohammadi, Abdulmotaleb El Saddik, and Jiyong Zhao. Compressed-domain video processing for adaptation, encryption, and authentication. *IEEE Multimedia*, 15(2):38–50, April 2008.
 - [14] ISO/IEC 14496-14. Information technology – coding of audio-visual objects, Part 14: MPEG-4 file format, 2003.
 - [15] ISO/IEC 15444-12. Information technology – JPEG2000 image coding system, Part 12: ISO base media file format, April 2005.
 - [16] ISO/IEC 15444-8. Information technology – JPEG2000 image coding system, Part 8: Secure JPEG2000, April 2007.
 - [17] ITU-T H.264. Advanced video coding for generic audiovisual services, November 2007.
 - [18] S. Kent and K. Seo. Security Architecture for the Internet Protocol. RFC 4301 (Proposed Standard), December 2005.

- [19] Y. Kim, S. Yin, T. Bae, and Y. Ro. A selective video encryption for the region of interest in scalable video coding. In *Proceedings of the TENCON 2007 - IEEE Region 10 Conference*, pages 1–4, Taipei, Taiwan, October 2007.
- [20] I. Kofler, M. Prangl, R. Kuschnig, and H. Hellwagner. An H.264/SVC-based adaptation proxy on a WiFi router. In *Proceedings ACM NOSSDAV*, May 2008.
- [21] Sang Gu Kwon, Woong Il Choi, and Byeungwoo Jeon. Digital video scrambling using motion vector and slice relocation. In *Proceedings of Second International Conference of Image Analysis and Recognition, ICIAR'05*, volume 3656 of *Lecture Notes in Computer Science*, pages 207–214, Toronto, Canada, September 2005. Springer-Verlag.
- [22] Live Networks Inc. LIVE555 Streaming Media. <http://www.live555.com/liveMedia/>.
- [23] Benoit M. Macq and Jean-Jacques Quisquater. Cryptology for digital TV broadcasting. *Proceedings of the IEEE*, 83(6):944–957, June 1995.
- [24] Enrico Magli, Marco Grangetto, and Gabriella Olmo. Conditional access to H.264/AVC video with drift control. In *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME'06*. IEEE, July 2006.
- [25] Mark Baugher, David McGrew, Melinda Shore. Securing Internet Telephony Media with SRTP and SDP. Cisco White Paper. <http://www.cisco.com/web/about/security/intelligence/securing-voip.html>.
- [26] Steven McCanne, Van Jacobson, and Martin Vetterli. Receiver-driven layered multicast. In *SIGCOMM '96: Conference proceedings on Applications, technologies, architectures, and protocols for computer communications*, pages 117–130, New York, NY, USA, August 1996. ACM.
- [27] D. Mukherjee, A. Said, and S. Liu. A framework for fully format-independent adaptation of scalable bit streams. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1280–1290, October 2005.
- [28] D. Mukherjee, H. Wang, A. Said, and S. Liu. Format independent encryption of generalized scalable bit-streams enabling arbitrary secure adaptations. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05*, Philadelphia, March 2005.
- [29] National Institute of Standards and Technology. FIPS-197 - advanced encryption standard (AES), November 2001.
- [30] Thomas Nithin, Damien Lefol, David Bull, and David Redmil. A novel secure H.264 transcoder using selective encryption. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'07)*. IEEE, September 2007.
- [31] Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG. Joint Scalable Video Model. *Doc. JVT-X202*, July 2007.

- [32] R. Kuschnig, I. Kofler, M. Ransburg, H. Hellwagner. Design options and comparison of in-network H.264/SVC adaptation. *Journal of Visual Communication and Image Representation*, September 2008.
- [33] Eric Rescorla. *SSL and TLS — Designing and Building Secure Systems*. Addison-Wesley, second edition, March 2001.
- [34] A. Said. Measuring the strength of partial encryption schemes. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'05)*, volume 2, September 2005.
- [35] H. Schulzrinne and S. Casner. RTP Profile for Audio and Video Conferences with Minimal Control. RFC 3551 (Standard), July 2003.
- [36] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A Transport Protocol for Real-Time Applications. RFC 3550, July 2003.
- [37] H. Schulzrinne, A. Rao, and R. Lanphier. Real Time Streaming Protocol (RTSP). RFC 2326, April 1998.
- [38] Daniel Socek, Hari Kalva, Spyros Magliveras, Oge Marques, Dubravko Čulibrk, and Borko Furht. New approaches to encryption and steganography for digital videos. *Multimedia Systems*, 13(3):191–204, 2007.
- [39] Thomas Stütz and Andreas Uhl. Format-compliant encryption of H.264/AVC and SVC. In *Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'08)*, Berkeley, CA, USA, December 2008. IEEE Computer Society.
- [40] A. Uhl and A. Pommer. *Image and Video Encryption. From Digital Rights Management to Secured Personal Communication*, volume 15 of *Advances in Information Security*. Springer-Verlag, 2005.
- [41] Chitra Venkatramani, Peter Westerink, Olivier Verscheure, and Pascal Frossard. Securing Media for Adaptive Streaming. *ACM Multimedia*, pages 307–310, 2003.
- [42] S.J. Wee and J.G. Apostolopoulos. Secure scalable streaming enabling transcoding without decryption. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'01)*, Thessaloniki, Greece, October 2001.
- [43] S.J. Wee and J.G. Apostolopoulos. Secure scalable video streaming for wireless networks. In *Proceedings of the 2001 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2001)*, Salt Lake City, Utah, USA, April 2001. invited paper.
- [44] S.J. Wee and J.G. Apostolopoulos. Secure scalable streaming and secure transcoding with JPEG2000. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'03)*, volume I, pages 547–551, Barcelona, Spain, September 2003.
- [45] S. Wenger, M.M. Hannuksela, T. Stockhammer, M. Westerlund, and D. Singer. RTP Payload Format for H.264 Video. RFC 3984, February 2005.

- [46] S. Wenger, Y. Wang, T. Schierl, and A. Eleftheriadis. RTP Payload Format for SVC Video. Internet Draft draft-ietf-avt-rtp-svc-14, September 2008.
- [47] T. Wiegand, J. Ohm, G. Sullivan, and A. Luthra. Special Issue on Scalable Video Coding - Standardization and Beyond. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9), September 2007.
- [48] Y. G. Won, T. M. Bae, and Y. M. Ro. Scalable protection and access control in full scalable video coding. In *Proceedings on the 5th International Workshop on Digital Watermarking, IWDW '06*, volume 4283 of *Lecture Notes in Computer Science*, pages 407–421, Korea, November 2006. Springer.