Transport and Signaling of SVC in IP Networks

Stephan Wenger, *Member, IEEE*, Ye-Kui Wang, *Member, IEEE*, and Thomas Schierl, *Member, IEEE*

(Invited Paper)

***Abstract*—The transport of scalable media, and in particular of scalable video conforming to the forthcoming Scalable Video Coding (SVC) technology, presents challenges not only in the video compression technology, but also in transport and signaling. This paper discusses the current status of standardization of the support for scalable media, and SVC in particular, over IP based networks. Both the transport of SVC over the Real-time Transport Protocol (RTP), and the signaling support—namely the additional mechanisms in the Session Description Protocol (SDP)—are covered. As it turns out, the support of SVC over RTP is not quite as straightforward as that of nonscalable video bit streams. Specifically, the signaling architecture requires an almost complete overhaul, and new protocol mechanisms need to be introduced into the packetization.**

***Index Terms*—H.264/AVC, Real-time Transport Protocol (RTP) payload, Session Description Protocol (SDP) signaling, Scalable Vido Coding (SVC).**

I. Introduction

SCALABILITY in video coding and transmission has been a research topic for at least a decade. Years ago, the key driver has been the desire for complexity scalability, i.e., to allow decoding of a complex bit stream even on computationally limited (and cheap) devices. One incarnation of this train of thought has been MPEG-1's B-picture concept. B-pictures are roughly twice as complex to decode as P-pictures, and, therefore, the committee felt that it should be possible to discard these "expensive" pictures without a penalty in reproduced quality, except the loss in frame rate. This lead to the often-lamented, and technically not justifiable, coupling of the bidirectional-predicted nature of B-pictures and their disposability—a defect that was cured only in 2003 with the advent of H.264/AVC.

As time went by, the focus of requirement discussions for scalable architectures changed towards bandwidth availability on the transmission path to the receiver, and on screen size. Nevertheless, until 2003, scalability was not on the top of the topic list of researchers in the field, although several attempts have been made around 1992 (MPEG-2 scalability), and 1997 (H.263+ and MPEG-4 scalable profiles). However, from 2003 onwards, the Joint Video Team (JVT, consisting of members of ITU-T VCEG group and MPEG's video group) took on scalability in earnest once more, and this project is now coming to a successful conclusion in the form of Annex G of H.264/AVC [1]. This annex specifies H.264/AVC's scalable extension, and is known as Scalable Video Coding (SVC) [2].

In the most basic form of SVC, a video signal is represented by one base layer and one or more enhancement layers. An enhancement layer may increase the temporal resolution (i.e., the frame rate), the spatial resolution, or the quality of the video content, compared to what is available when decoding only a layer the enhancement layer is based on. Enhancement layers can be "stacked" on top of each other. In SVC, it is even possible to make an enhancement layer directly dependent on more than one "lower layer," and quite complex graphs of layer dependencies can be implemented, subject to the constraint, though, that in one access unit a layer picture can directly depend only on one lower layer.

Each layer, together with all its dependent lower layers, forms one representation of the video signal at a certain spatial resolution, temporal resolution and quality level. In this paper a scalable layer representation is being referred to as one given layer together with all lower layers it directly or indirectly depends on. One scalable bit stream contains layers that form at least two, but sometimes many more scalable layer representations. Each scalable layer representation can be extracted from the scalable bit stream by low complexity bit stream extraction operations without transcoding.

SVC's recently finalized specification (Phase 1) offers several forms of scalability. In alignment with the terminology used within the JVT community, Coarse-grained scalability (CGS) is referred to as the traditional quality [signal-to-noise ratio (SNR)] scalability. Here, spatial resolution and frame rate stay constant, but the number of bits spent per pixel increases, resulting in better quality. The form of scalability that allows for changes in the spatial resolution (pixel count) is referred to as spatial scalability, and the scalability with frame rate change is referred to as temporal scalability. JVT also entertained the concept of fine granularity scalability (FGS) over a long period of time; however in SVC Phase 1, FGS is not supported. Instead, the so-called medium-grained (granularity) scalability (MGS) is supported. MGS is similar to CGS in that only quality enhancement is involved and data unit truncation is not possible, with the difference being that MGS data units can be freely dropped without affecting the conforming of the resulting bit stream, while in case of CGS, all data units of the complete layer are either processed or not.

Even the most powerful video compression technology is rather useless without an application. Broadly put, applications

Manuscript received October 6, 2006; revised July 13, 2007. This paper was recommended by Guest Editor T. Wiegand.

S. Wenger is with Nokia Corporation, Palo Alto, CA 94304, USA (e-mail: stephan.wenger@nokia.com).

Y.-K. Wang is with Nokia Research Center, 33721 Tampere, Finland (e-mail: ye-kui.wang@nokia.com).

T. Schierl is with Fraunhofer—Heinrich Hertz Institute (HHI), D-10587 Berlin, Germany (e-mail: thomas.schierl@hhi.fraunhofer.de).

Digital Object Identifier 10.1109/TCSVT.2007.905523

using video compression can be categorized in store-forward applications—with the DVD as the prime example—and communication applications. Of the latter, the already very significant, but soon predominant infrastructure is based on protocols known as the Internet Protocol (IP, STD 0005, RFC 791 [3]), User Datagram Protocol (UDP, STD 0006, RFC 768 [4]), and Real-time Transport Protocol (RTP, RFC 3550 [5]). RTP covers the media format independent real-time transport for point-to-point and multicast scenarios, and relies on so-called RTP payload formats for media adaptation.

The most recent incarnation of SVC is a backward compatible enhancement of H.264/AVC [1], and therefore it is not a surprise that the draft SVC RTP payload format [6] is based on the corresponding specification for H.264/AVC—namely RFC 3984 [7]. However, in addition to straightforward enhancements of the semantics of previously existing RFC3984 codepoints, the current draft contains numerous conceptually new enhancements, whose discussion occupies a large part of the present paper.

While for the packetization a relatively straightforward enhancement of RFC 3984, as discussed below, is sufficient to create a state-of-the-art payload specification, this is not the case for the signaling. Specifically, signaling of layers in the RTP and SDP context has not really been revisited since its invention, which was in the context of the MBONE project and in the late 1990s. It was believed then that a straightforward mapping of layers to RTP sessions residing on ascending IP addresses in the IP multicast address space would be sufficient. Today, however, IP multicast is not quite as widely deployed as the MBONE pioneers envisioned—in practice, it's not used for media transmission at all (that may change in due course in certain "private" IP networks, e.g., those of operators in the 3GPP world who want to offer Multimedia Broadcast/Multicast Service (MBMS) [8]). Furthermore, the common use of network address translation (NAT) and firewalls presents new challenges. Specifically, it is in practice not possible—or at least overly costly in terms of manual configuration—to open many pinholes in a firewall (or translate many transport addresses) just to convey a single media stream which happens to be in a layered architecture. Without an improved signaling architecture, this situation—bluntly put—prevents the deployment of SVC over IP networks—which are envisioned as the key network architecture for SVC. Therefore, the SVC community took the initiative to redesign the signaling model for layered codecs altogether, to make the signaling compatible with today's IP world.

As for the placement of the aforementioned specification text in the IETF document space, it is obvious that a new RFC would be required, covering the RTP payload specification for SVC and the SVC-related signaling aspects. Confronted with the outdated signaling mechanisms as specified in [9], however, it was further decided to make an attempt to divide the signaling into two specifications—one that covers all aspects of signaling for a generic, hypothetical multilayered codec (extended in scope towards concepts such as multiple-description coding with no clean hierarchy of layers, following the lead of [10]), and the other to build on that generic specification and cover only the SVC specific details. As a result, this paper covers the status of the development of two Internet Drafts as per mid 2007, namely [6] and [11].

SVC retains H.264/AVC's Network Abstraction Layer (NAL) concept and key properties. NAL units form the basic structure of an SVC bit stream. The parameter set concept is still used to convey most important information that pertains to more than one NAL unit. Due to space constraints, readers are referred to [12] for a detailed introduction of SVC system and transport interface, including the NAL unit structure.

The rest of this paper is organized as follows. Section II covers the network and design constraints, followed by Section III containing a detailed description of the draft SVC payload specification. Section IV discusses the SDP extensions introduced to enable signaling layered codecs in general, and SVC in particular. Section V concludes the paper with a summary.

II. Network and Design Constraints

A. System Models/Topologies Envisioned

Before the detailed requirements discussion, the network and distribution models considered relevant shall be introduced. These follow, with a few detours of thought, from the most basic design criteria that needs to be obeyed for most (if not all) IETF standardization work—namely that the newly designed protocols must be compatible with design choices made earlier in "parent" protocols. Key examples of such parent protocols that need to be considered are RTP [5], some of its companion documents such as the audio-visual profile RFC 3551 [13] and secure RTP (SRTP) RFC 3711 [14].

Furthermore, it should be noted that the choice of network and distribution models is also influenced by practical design constraints of the current Internet. First and foremost, the lack of support of IP multicast in large parts of today's IP networks (including the Internet, but by no means limited to it) makes it impossible to rely exclusively on IP-multicast-based distribution models. Second, the practical requirement of compatibility with NATs and firewalls makes it necessary to consider network and distribution models that were undesirable if there were a world without NATs and firewalls. In particular, RFC 3984 already introduced the concept of a Media Aware Network Element (MANE)—a system that meaningfully manipulates the RTP stream in a lightweight fashion, based on information available only in the signaling and in the RTP header, RTP payload header, and perhaps NAL unit header. It is envisioned that this concept is of considerable value once SVC is in use. And third, it is also refrained from discussing a few corner-cases that may be implementable while obeying the first two constraints, but would—in our opinion unnecessarily—bloat this paper.

The latter applies particularly to the use of layered codecs in applications not using a server-client model, where very similar design choices can be made, although the terminology might be different. For example, in a multipoint video conferencing scenario with layered coding support, the equivalent of a server would be the sending endpoint, the equivalent of a client would be the receiving endpoint, and the equivalent of a MANE would be functionality residing in the multipoint control unit (MCU).

With these remarks, in the following the basic network distribution models are presented that are considered relevant. Note that in real applications, the topologies may be combined.

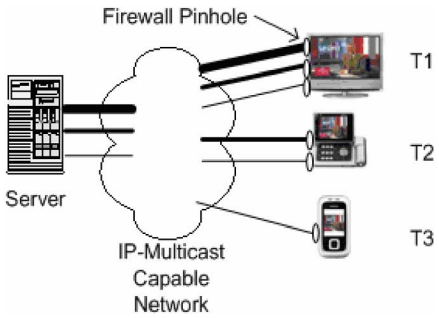

Fig. 1. Receiver driven layered multicast.

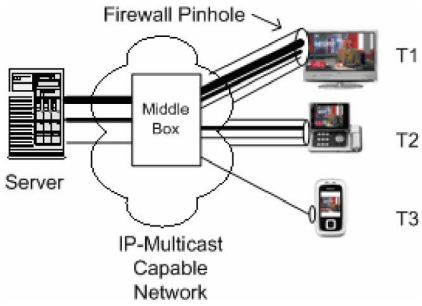

Fig. 2. MANEs in network to aggregate and/or trim RTP sessions.

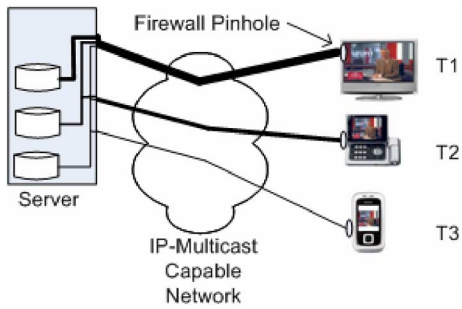

Fig. 3. Unicast—layers all in one session.

1) Multicast/broadcast of video data to receivers with heterogeneous connectivity, where layers are transported in separate RTP sessions on separate transport addresses.
2) Multicast/broadcast on the server side, with a MANE to aggregate and/or trim sessions. The NAL units of the aggregated and/or trimmed sessions are conveyed jointly on a single transport address, and in a single RTP session.
3) Starting from a layered representation in a file, the server generates and sends one RTP session containing possibly more than one layer.

Fig. 1 depicts use case 1. A server carries one base layer and two enhancement layers, forming a hierarchy. Terminals T1, T2, and T3 are connected to the server through the Internet, over links that allow for certain maximum bit rates. The capacity of the links and the bit rate demands of the streams are illustrated by the line width of the connections—the wider a line, the higher the bit rate. The end-to-end capacity is a function of both the connectivity of the endpoint and the congestion of each link. Therefore, the picture should be viewed as a snapshot of a configuration at a given time—the connectivity of each terminal may change frequently with the changes of the congestion situation. Note also that for the sake of simplicity, uncongested links from the server to the backbone are assumed.

According to the receiver driven layered multicast concept, first introduced by McCanne in [15], each layer is transported in its own IP multicast group identified by its own IP multicast address, and terminals subscribe to layers utilizing IP multicast mechanisms, namely IGMP [16]. This implies that one terminal may have to subscribe to many IP multicast groups for the best possible quality. While, considering Internet technologies in their purest form, this is not a problem and actually desirable, practical constraints—namely the existence of NATs and firewalls—make such an approach only feasible in certain academic and research environments.

This line of thought leads to a scenario as depicted in Fig. 2. As the server, in most cases, will be a professionally maintained device, it is reasonable to assume that its administrators have control over the firewall and can open as many pinholes as required. Therefore, the server sends to multiple IP multicast groups, each carrying a single layer. Close to the edge of the network, a middlebox, also known as MANE, is used to aggregate the content of potentially more than one multicast group into a single stream carrying one or more layers. When performing aggregation, the MANE may omit the unwanted layers of some multicast groups. Only for the single stream constructed by the MANE, a pinhole to a single transport address has to be opened in a firewall. MANEs operate as mixers, and, therefore, the outgoing RTP session is fully under their control and terminated by the MANE and the respective endpoint. Physically, middleboxes of this kind are likely to be co-located with wireless access gateways and similar entities.

The advantage of such a topology is reduced server and core network load, and reduced server complexity, as the server does not need to generate and simulcast multiple full representations.

In order to fulfill its role, the MANE has to be aware of the details of RTP, its payload format, and the signaling. It needs to be located inside the security context of the sessions. In short, these MANEs are not simple (network layer) routers that receive configuration information through IGMP, but need to be signaling aware application layer devices of high sophistication.

MANEs terminates RTP sessions. This implies per RFC 3550, a very loose relationship between the incoming and outgoing RTP sessions. In particular, there is no direct relationship between the incoming and outgoing RTP sequence numbers, RTP timestamps, payload types used, and so on.

MANEs are conceptually simple devices (though not necessarily trivial to implement) and can offer powerful features, primarily because they necessarily can “see” the payload (including the RTP payload headers), utilize the wealth of layering information available therein, and manipulate it.

A third scenario is presented in Fig. 3. Here, the terminals are connected directly to the server, utilizing only a single transport address (IP address and port number) for video. For each terminal, the server composes a bit stream tailored for the terminal’s needs, by aggregating NAL units of appropriate layers. The aggregation process is possible through the generation of a single RTP session carrying multiple layers.

B. Design Criteria for the Payload Format

The following design criteria have to be kept in mind when discussing the draft RTP payload specification for SVC, available in [6]. Note that these criteria are partly the result of commercial thinking; only very few are entirely technology-driven. They are presented in a rough order of importance.

1) No mechanisms should be included that would break commonly employed/envisioned RTP base technologies in use over today's Internet. Prime example: the payload format must not include any mechanisms that cannot be used in conjunction with encryption (SRTP) and feedback (AVPF RFC4585 [17]).
   Please note the language "over Internet" in contrast to "over IP." The distinction is both technical and procedural. The Internet, currently and in the vast majority of cases, is considered a best-effort IP network and offering end-to-end connectivity. Congestion control and security are major problems and have to be addressed by every specification under IETF control. Intelligence in the network is not considered a desirable feature; whenever possible, intelligence should be implemented in the endpoints. Private IP networks may set other priorities, and the best design choices for those may differ from those for the Internet.
2) In JVT it has been decided that the base layer of an SVC bit stream must be conforming to one of the old profiles of H.264/AVC, in order to allow legacy H.264/AVC decoders to use SVC bit streams tailored for them. For the same reason, on the wire and (to the extent possible) in the signaling, the RTP payload specification for SVC must "look" like RFC 3984.
3) Even for enhancement layers, it appears sensible to enable synergy effects by reusing as many mechanisms of RFC 3984 as possible, so to ease implementation burden, reuse stable specification text (preferably by citation), allow for code reuse, and so on.
4) Both layered multicast and the transport of multiple layers in a single RTP session must be possible, though it is not required to handle all layer combinations theoretically possible in SVC. In other words, imposing restrictions related to exotic configurations of layers are considered, when justified by simpler specification or implementation.
5) Optimizations for layered stream manipulations are desirable. Examples include lightweight means to associate individual RTP packets to layers, a table of content for RFC 3984's aggregation packets, support to reorder NAL units into decoding order even when the layers are conveyed across sessions.

C. Design Considerations for Signaling

The signaling support for SVC has to take at least the following design considerations—again in the rough order of importance—into account.

1) On many occasions, the IETF has displayed a preference for generic solutions over specific ones. The signaling support for SVC is no exception. The authors of the RTP payload specification have been asked to extract all mechanisms useful for "generic" signaling of layered and/or multiple description codecs into its own specification. This new specification should augment the rudimentary support for signaling of layered codecs as present since the early days of SDP [9]. The RTP payload specification should build on top of this generic layered coding support signaling specification.
2) As much legacy technology and behavior as possible should be carried forward from RFC 3984 and other commonly used RFCs. Ideally, a legacy receiver, supporting only old profiles of H.264/AVC and RFC 3984, should be able to connect to an SVC capable sender without excessive signaling overhead—and without the need to convey the base layer in more than one RTP session (even when both legacy and SVC capable receivers need to subscribe to the multicast group carrying the base layer).

D. Problems to Solve

So far, the following problems have been identified, which entail specifying protocol extensions relative to RFC 3984 (for both transport and signaling) that require additional bits on the wire, including a significant change in semantics for protocol elements already defined in RFC 3984, and nonstraightforward changes and extensions in the signaling. These problems are:

1) cross-layer synchronization when layers are sent in multiple RTP sessions;
2) need for a "table of content" of an aggregation packet;
3) SVC stream adaptation, i.e., enabling the pruning of a scalable bit stream carried in a given RTP session by removing layers and/or composing an RTP session containing layers previously carried in their own sessions;
4) signaling aspects; in particular an SDP [9] representation of layer dependencies and attributes, including the challenges related to the desire to carry more than one—but not all—layers of a scalable bit stream in one RTP session.

III. RTP Payload for SVC

The RTP payload for SVC follows, wherever possible, the guidance of RFC 3984. It retains the basic transport structures of RFC 3984, including the Aggregation and Fragmentation mechanisms. However, in certain cases, while the syntactical structures remain intact, the semantics of some fields change subtly. The most obvious of these cases is the redefinition of the decoding order number (DON) in the interleaved packetization mode, which now spans across more than one RTP session when layers are transported in more than one RTP session. See Section III-A for more details.

Backward compatibility is a design goal very high up on the priority list of the designers of the SVC payload. In particular, there is a value of mandating the use of RFC 3984, whenever a base layer is sent by its own. JVT has consciously made the decision that the base layer be H.264/AVC compatible (i.e., conforms to one of the established pre-SVC profiles of H.264/AVC) [18], and it would be outright imprudent to contradict this decision by enforcing the use of a different payload specification. Enhancement layer data is encapsulated according to the SVC payload specification.

Since there is such a high level of similarity between RFC 3984 and the forthcoming SVC payload specification, we first

briefly review the features of RFC 3984. Thereafter, the three most difficult problems encountered so far, and the current state of solutions, are discussed.

A. Review of RFC 3984

RFC 3984 supports encapsulating a single NAL unit, more than one NAL unit, or a fragment of a NAL unit into one RTP packet. A single NAL unit as specified in H.264/AVC can be included in the RTP packet "as is," and the NAL unit header co-serves as the payload header. Four types of aggregation NAL units are specified. The two single-time aggregation packet types, STAP-A and STAP-B allow encapsulating more than one NAL unit into one RTP packet that stem from the same picture (identified by identical RTP timestamp). The two multiple-time aggregation packet types, MTAP16 and MTAP24, respectively, can be used to aggregate NAL units from different pictures into one RTP packet. RFC 3984 also supports two types of fragmentation units, FU-A and FU-B, which enable fragmentation of one NAL unit into multiple RTP packets.

Figs. 4 and 5 depict an example of STAP-B and MTAP16, respectively. In both cases, two NAL units are aggregated into a single RTP packet. As can be seen, STAP-B includes a field DON, while MTAP16 includes fields DONB and DOND. These fields are used to indicate or derive the DON, which can be used to reorder NAL units into their decoding order. DON related fields are available in STAP-B, MTAP16, MTAP24, and FU-B. In STAP-B, the DON field indicates the DON value of the first NAL unit carried in the RTP packet. For the following NAL units contained in the same RTP packet, the DON increases by one for each NAL unit. The fields DONB, for the RTP packet, and DOND, for each NAL unit, included in MTAP16 can be used to derive the DON value of each NAL unit contained in the RTP packet. The introduction of the DON concepts allows transmitting NAL units out of their decoding orders, in the interleaved packetization mode.

RFC 3984 specifies three packetization modes, single NAL unit mode, noninterleaved mode and interleaved mode. STAP-B, MTAP16, MTAP24, and FU-B are allowed in the interleaved mode.

The benefits of having aggregation packets and out-of-decoding-order transmission of NAL units are as follows. Firstly, aggregation of multiple coded pictures into the same RTP packet can reduce packet header overhead. A bit rate saving of 5 to 10 percent is typical when the video bit rate is not larger than 64 kbps [19]. Secondly, when temporal scalability is supported, sending lower temporal layers earlier than other data can avoid rebuffering in mobile streaming, since after a handover, instead of rebuffering data, the player can play a lower frame rate [20]. Thirdly, improved error resilience can be achieved by sending more important data earlier such that there is more time for the retransmission [19]. As discussed in Section III-A, this would also allow for simple cross-layer synchronization of NAL units in different SVC layers transmitted in different RTP sessions.

B. Cross-Layer Synchronization

SVC, as all previous video compression standards, requires that syntactical entities of the bit stream be presented to the decoder in a certain order, the decoding order. In case of H.264/AVC and SVC, the decoding order is expressed in constraints for the sequencing of the NAL units. Some H.264/AVC and SVC profiles allow a certain amount of NAL unit reordering without breaking compliance, but others do not. In any case, it is necessary to include mechanisms in the transport layer that allow for efficient NAL unit reordering.

RTP supports packet reordering by the means of the RTP sequence number, and time synchronization between different RTP sessions by the means of the RTP timestamp and the RTCP sender reports.

The NAL unit decoding order, however, is not necessarily identical to the transmission order or the packet order. For example, when the interleaved packetization mode of RFC 3984 is used, it is sometimes impossible to infer the correct NAL unit ordering from the aforementioned information. When transporting layers in different RTP sessions, the situation gets even more complicated. Early versions of the SVC payload draft have attempted to specify an algorithm for this reordering [21], but the specification and implementation complexities are considered excessively high.

An alternative to this approach is an explicit signaling of the order of the NAL units in the packet stream. This requires that the interleaved mode is used, and the explicitly included or derived DON values indicate the NAL unit decoding order across all the layers.

The tradeoff between the two design choices is comparatively simple: an inference algorithm as specified in [21] requires a considerable amount of implementation complexity. The use of the interleaved mode has certain known commercial implications, and also adds a few bits on the wire. After careful consideration, the AVT WG in the IETF decided to use the Interleaved Mode to overcome this specification problem. This solution has been stable now for several IETF meeting cycles.

Fig. 6 illustrates an example of the cross-layer synchronization. The base layer (layer 0 in the figure) is a QCIF@15 Hz bit stream, and the quality is enhanced by a SNR scalable layer (layer 1) with the same frame rate. Based on the SNR scalable layer, a spatial enhancement layer (layer 2) of CIF@30 Hz is encoded, and finally an MGS layer (layer 3) is encoded based on the spatial enhancement layer. Three access units are depicted in the figure. Layer 0 is transported in RTP session S_1, using STAP-B. Layer 1 is transported in RTP session S_2, with the NAL unit in access unit 1 being fragmented to two FUs (FU-B with the first packet and FU-A with following packets). Layers 2 and 3 are transported in the same RTP session S_3, using both STAP-B and MTAP. As can be seen, the DON values indicate the decoding order of all the NAL units across all the layers. This way, the receivers can easily recover the NAL unit decoding order from the signaled or derived DON values. For a receiver that receives only the base layer, there will be gaps in the DON values of some NAL unit continuous in decoding order, which is compliant with RFC3984. However, and fortunately, the derived decoding order would still be correct.

C. Payload Content Scalability Information NAL Unit

One perceivable objective of MANEs is to control the bit rate of the forwarded bit stream according to the prevailing downlink network conditions. It is desirable to control the forwarded data

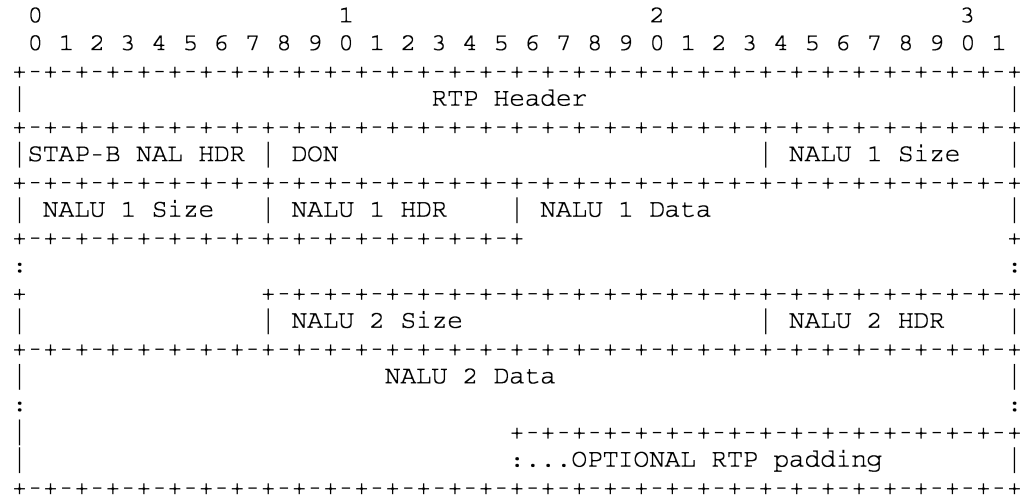

Fig. 4. RTP packet including an STAP-B carrying two NAL units.

rate without extensive processing of the incoming data, e.g., by simply dropping packets. Due to the requirement to handle large amounts of data, MANEs have to identify removable packets as quickly as possible. Furthermore, this identification is also helpful for the playback of a bit stream with a certain desired quality and complexity below what the bit stream offers, as receivers and players should be able to identify those data that they are incapable or unwilling to decode.

The interleaved packetization mode of RFC 3984 allows for the encapsulation of practically any NAL units of any access units into the same RTP payload of a given aggregation packet. In particular, it is not required to encapsulate entire coded pictures in one RTP payload, but rather the NAL units of a coded picture can be split into multiple RTP packets. The SVC payload format inherits this encapsulation capability. While the liberty of packet aggregation is welcome for many applications, it causes a number of complications in a MANE operation. First, given an aggregation packet, it is not known to which pictures the NAL units belong to, until the header of each NAL unit contained in the aggregation packet has been parsed. Therefore, when the interleaved packetization mode is applied for SVC, the layers in which the contained NAL units belong are not known, before parsing the header of each NAL unit in the packet. Consequently, a MANE has to parse each NAL unit header before deciding whether any, all, or some NAL units of the packet are to be forwarded. Second, for some NAL units, such as supplemental enhancement information (SEI) and parameter-set NAL units, it is not possible to identify the access unit they belong to before video coding layer (VCL) NAL units of the same access unit are received. Therefore, a MANE may need to maintain a buffer and some state information to resolve the mapping of non-VCL NAL units to their associated pictures.

A new NAL unit type has been specified to enable easy identification of scalability dependencies within the bit stream, thereby enabling fast and efficient bit stream manipulation. It is known as payload content scalability information (PACSI), whose structure is the same as the 4-byte SVC NAL unit header. The PACSI NAL unit, if present, must be the first NAL unit in an aggregation packet, and it must not be present in other types of packets. The PACSI NAL unit indicates scalability characteristics that are common for all the remaining NAL units in the payload, thus making it easier for MANEs to decide whether to forward or discard the packet. Senders may create PACSI NAL units and receivers can ignore them. The NAL unit type for the PACSI NAL unit is selected among those values that are unspecified in the H.264/AVC specification and in RFC 3984. Thus, SVC bit streams having H.264/AVC base layer and including PACSI NAL units can be processed with RFC 3984 receivers and H.264/AVC decoders.

IV. SDP Signaling for SVC

Modern RTP payload formats include sections on the signaling of the payload format. More specifically, they are as follows.

- MIME type registration reserves a unique name for the payload format in the MIME name space as administrated by IANA. The MIME type usually includes also a definition of the names, syntax and semantics of optional parameters of the type—which can be used during session setup for capability exchange:
- mapping of the MIME parameters on SDP;
- consideration when using the SDP in an offer-answer model;
- considerations when using the SDP in a declarative model, i.e., for the use in RTSP.

A. Session Description Protocol

Many systems employing RTP for media transport also rely on IETF-specified protocols for the session setup. The most commonly used protocols in this field are the Session Initiation Protocol (SIP) [22] and the Real Time Streaming Protocol (RTSP) [23]. Both rely on Session Description Protocol (SDP) for the representation of the description of the multimedia session and the individual media the session consists of. Although the term SDP suggests some form of protocol activity, the specification is perhaps best described as a definition for a lightweight language to define session and media. The emphasis of the designers was to create an easily parse-able, back-to-basics syntax,

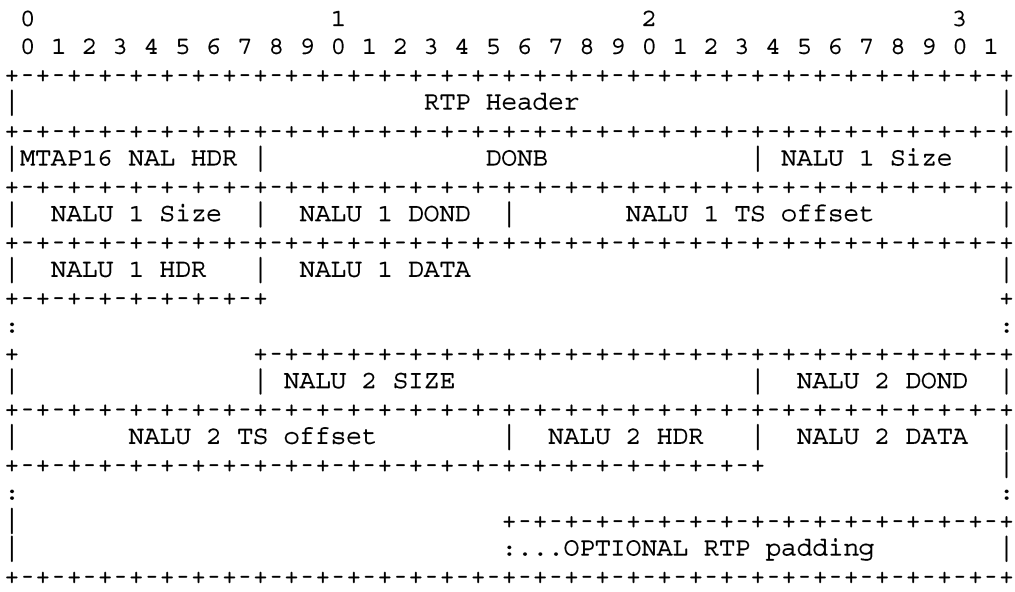

Fig. 5. RTP packet including a MTAP16 carrying two NAL units.

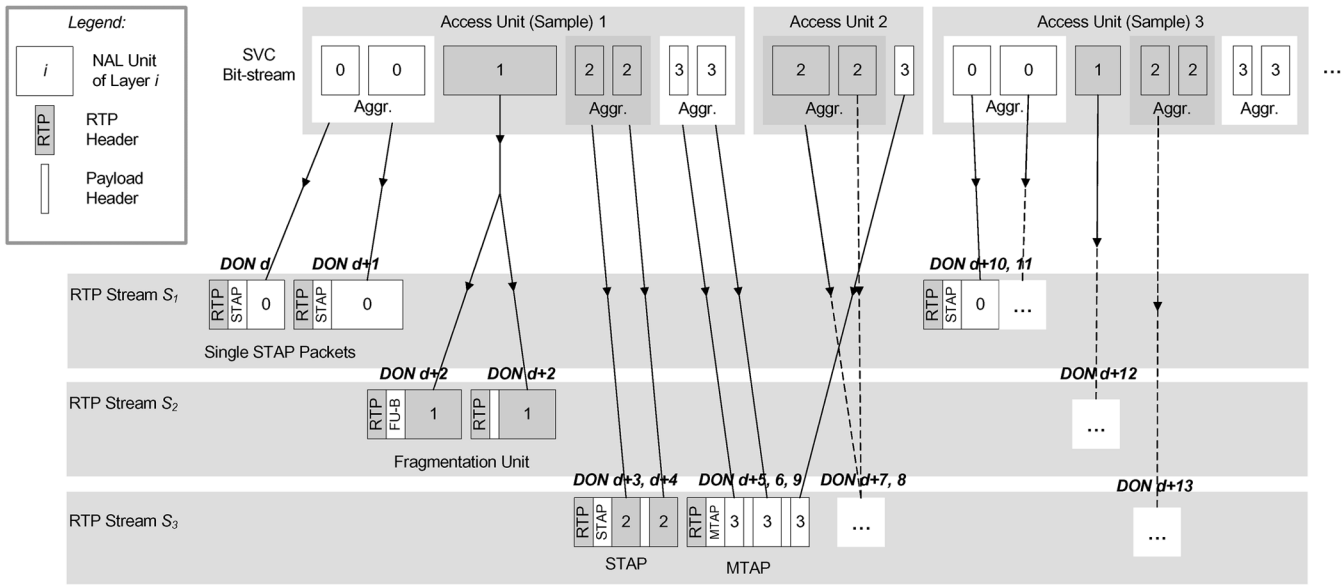

Fig. 6. Cross-layer synchronization with packetization in interleaved mode.

and the limitations of this design choice presents a major challenge for the description of media as powerful and flexible as SVC.

SDP implements two distinct levels of description. The session level describes the session itself, i.e., the session's name ("s="), originator with contact details ("e="), and a session-wide encryption key ("k="), when (a very simple form of) encryption is in use. It is also possible to specify here the transport address range ("c=")—however, most implementations rely on the corresponding field in the media level ("m="). The media level description is present for each media stream.

On both session and media levels, an attribute ("a=") can be used to further define the session/media. The SDP specification defines a few attributes itself. One of those in common use is "rtpmap," which maps a media session to a payload type (e.g., the dynamic type "99" for H264 in the example below), and RTP timestamp basis (e.g., "90000" ticks per second in the example below). The use of attributes defined on the media level is SDP's prime mechanism to convey media-specific information in the "fmtp" attribute during session setup.

Below is an example of an RTP session description for an SVC bit stream transport over two RTP session, one for the AVC base layer and one for the SVC enhancements. Please also refer to [9, Sec. 5], for an in-detail discussion of general SDP attributes.

```
v=0
o=jdoe 2890844526 2890842807 IN IP4
10.47.16.5
s=SDP SVC Secssion Seminar
i=A Seminar on the session description
```

```
    protocol using SVC
u=http://www.example.com/seminars/svc.pdf
e=j.doe@example.com (Jane Doe)
c=IN IP4 224.2.17.12/127
t=2873397496 2873404696
a=recvonly

a=group:DDP 1 2

m=audio 49170 RTP/AVP 0

m=video 51372 RTP/AVP 99
a=rtpmap:99 H264/90000
a=fmtp:99 profile-level-id=4d400a;
packetization-mode=2; init-buf-time=0;
sprop-parameter-sets=Z01ACprLFicg,
    aP4DGoA=;
a=mid:1

m=video 52012 RTP/AVP 100
a=rtpmap:100 SVC/90000
a=fmtp:100 profile-level-id=53000a;
    packetization-mode=2; init-buf-time=0;
    sprop-parameter-sets=Z01ACprLFicg,
    Z1MACksA1NZYsTk=,aP4DGoA=,aEvgZqA=,
    aGvgZiA=;
a=mid:2
a=depend:lay 1
```

An SDP session description may contain various media descriptions (e.g., for video as well as for audio), each identifying one RTP session. In some cases different media coding types are available per RTP session, identified by different RTP payload types. An RTP session may further consist of RTP streams (e.g., in a multipoint-to-multipoint scenario) from different sources, where each stream is identified by a separate and unique SSRC within the RTP session.

The "historic" support mechanisms for layering defined in today's SDP have their roots in the concepts of receiver driven layered multicast. These mechanisms allow for specifying a "start" IP (multicast) address or a "start" port number for a given media. For each layer, the IP (multicast) address or port number is increased by one. This implies that each layer is necessarily conveyed in its own multicast group, and poses restrictions on port numbers (in the nonmulticast sense) that are—in today's NAT-centric world—not implementable.

B. Generic Layering Support for SDP

At present, the generic SDP signaling mechanism for layered media appears to be best implemented on the (SDP) media session level. The current draft further relies on the transport of parts of a layered bit stream (be it a single layer, or a combination of layers) in their own respective RTP sessions. On the signaling side, this results in the need for potentially many media descriptions, some of which are only useful in combination with others. Therefore, there is a need to "group" these related media sessions, and furthermore describe their relationship to each other.

The proposed signaling mechanism is based on an extension to the SDP grouping of RFC3388 [24]. The reason for that is summarized as follows: An SDP session description may contain various media descriptions each identifying one media stream. If more than one media description exists indicating the same media type (e.g., video), a receiver or network element possibly cannot identify an existing relationship between those descriptions. This is certainly the case if the receiver or network element is not aware of the media specific information, which may be carried within in media specific attributes of the session. Relationships like dependencies of media streams may exist for different reasons as for transporting bit stream partitions of a layered coding process or of a multiple-description coding (MDC) process in different transport streams. SDP does not allow for signaling such relations.

Therefore, the SDP grouping has been extended by adding a new grouping type (decoding dependency "DDP"), indicating dependency between members (RTP sessions) of an SDP group. Further attributes indicate the type of dependency ("depend"). A layered dependency—as exists between layers of a hierarchical coding process—is identified by the attribute "lay." A multi descriptive decoding dependency—as exists between streams of a multiple-description coding process—is identified by "mdc." But the latter dependency is, today, neither related to an existing nor an upcoming coding standard.

C. Support of Legacy Devices

It is desirable that legacy devices, i.e., devices that implement nonscalable H.264/AVC and RFC 3984, but not SVC, can connect to SVC sources. From a media processing viewpoint, this is trivial—if the nonscalable base layer is conveyed in its own session and according to RFC 3984, then it can be decoded and reproduced without problems. Even if a session contains SVC NAL units, these NAL units are ignored by the legacy decoder as per H.264/AVC.

On the signaling side, it is unfortunately not possible to backward-compatibly enhance the signaling of RFC 3984. However, since two different media descriptions can legally "point" to the same RTP media stream, it appears possible to solve the legacy-problem. The RTP session containing the base layer is described by two media descriptions, one announcing the stream as nonscalable H.264/AVC or RFC 3984 and the other announcing it as SVC. It appears that such a mechanism would work with both RTSP and with the offer/answer model [25] commonly employed in SIP.

In point-to-multipoint scenarios, the more straightforward solution is to separate AVC and SVC NAL units into two different RTP sessions, one according to RFC 3984 and the other according to the new SVC payload.

D. An Example

The following SDP example illustrates the signaling of dependency relationships between media streams:

```
v=0
o=svcsrv 289083124 289083124 IN IP4
  host.example.com
s=LAYERED VIDEO SIGNALING Seminar
t=0 0
c=IN IP4 224.2.17.12/127
a=group:DDP 1 2 3 4

m=video 40000 RTP/AVP 94
b=AS:96
a=framerate:15
a=rtpmap:94 H264/90000
a=mid:1
a=depend:lay

m=video 40002 RTP/AVP 95

b=AS:96
a=framerate:15
a=rtpmap:95 SVC/90000
a=mid:2
a=depend:lay 1

m=video 40004 RTP/AVP 96
b=AS:64
a=framerate:30
a=rtpmap:96 SVC/90000
a=mid:3
a=depend:lay 1 2

m=video 40004 RTP/SAVP 100
c=IN IP4 224.2.17.13/127
b=AS:512
k=uri:conditional-access.example.com
a=framerate:15
a=rtpmap:100 SVC/90000
a=mid:4
a=depend:lay 1 2
```

In the example above, the separated transport in different media streams/RTP sessions is shown, as defined in [11]. A H.264/AVC base layer is transported in its own RTP session indicated by a media identifier of "mid:1." Further three SVC enhancement layers are separated into three different RTP sessions. For each SVC layer a separate RTP session is used and the relation is indicated by the "depend"-attribute. The RTP session marked with "mid:2" may contain a SVC enhancement layer of the same frame rate as the H.264/AVC base layer contained in the session with "mid:1," this SVC layer may be an CGS enhancement to the base layer. The SVC layer in RTP session marked with "mid:3" is of a higher frame rate than the SVC layer in session with "mid:2" and may be a further temporal enhancement. The last SVC layer signaled in the media session is an enhancement to SVC bit stream resulting from RTP sessions with "mid:2" and "mid:1," thus for this stream only a dependency to RTP session with "mid:1" and "mid:2" is signaled. This layer may be a spatial enhancement to a higher resolution. Further this RTP session is transported following the SRTP "RTP/SAVP" profile for encrypted sessions. This may be a high quality enhancement, which is under conditional access control.

V. Summary

The current state of standardization of the support for transport and signaling of SVC over IP has been presented. Specifically, the RTP payload format specification enhances RFC 3984 by two concepts. DON-based reordering of NAL units conveyed in separate RTP sessions enables the use of IP multicast transport towards MANEs close to the edge of the network, which combine data from different RTP sessions in a lightweight and efficient way into a single, unicast RTP stream—which is more friendly to NATs and firewalls. The PACSI NAL unit acts as a table of content of an aggregation packet.

The signaling support has been split into two documents. Responding to the outdated support for layered codecs in general—which has not been updated since 1998—a generic signaling enhancement for SDP is in the process of being specified. It attempts to cover not only layered coding schemes, but also related technologies such as multiple-description coding. Each layer (or a know group of layers handled as a unit) is being conveyed in its own RTP session, therefore requires its own media description. Following the lead of RFC 3388, a grouping scheme is introduced that describes the relationship between the various RTP sessions.

Beyond the generic signaling support, the need for SVC specific codepoints in the payload specification is obvious, but the current drafts do not yet address this problem.

References

[1] *Advanced Video Coding for Generic Audiovisual Services, v3*, ITU-T Rec. H.264|ISO/IEC IS 14496-10 AVC, 2005.

[2] T. Wiegand, G. Sullivan, J. Reichel, H. Schwarz, and M. Wien, *Joint Draft 11 of SVC Amendment* Joint Video Team (JVT), Doc. JVT-X201, Jun.–July 2007.

[3] J. Postel, "Internet protocol DARPA internet program protocol specification," IETF STD 0005, RFC 791, Sep. 1981.

[4] J. Postel, "User datagram protocol," IETF STD 0006, RFC 768, Aug. 1980.

[5] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," IETF STD 0064, RFC 3550, Jul. 2003.

[6] S. Wenger, Y.-K. Wang, and T. Schierl, "RTP payload format for SVC video," IETF Internet Draft Draft-Ietf-Avt-Rtp-Svc-02.txt, Jul. 2007.

[7] S. Wenger, M. M. Hannuksela, T. Stockhammer, M. Westerlund, and D. Singer, "RTP payload format for H.264 video," IETF RFC 3984, Feb. 2005.

[8] *Multimedia Broadcast/Multicast Service (MBMS); Protocols and Codecs, Version 6.1.0*, 3GPP TS 26.346, Jun. 2005.
[9] M. Handley, V. Jacobson, and C. Perkins, "SDP: Session description protocol," IETF RFC 4566, Jul. 2006.
[10] A. Vitali, A. Borneo, M. Fumagalli, and R. Rinaldo, "Video over IP using standard-compatible multiple-description coding: An IETF Proposal," *J. Zhejiang Univ. SCIENCE*, vol. 7, no. 5, pp. 668–676, Apr. 2006.
[11] T. Schierl and S. Wenger, "Signaling media decoding dependency in session description protocol (SDP)," IETF Internet Draft Draft-Schierl-Mmusic-Layered-Codec-04, Jun. 2007.
[12] Y.-K. Wang, M. M. Hannuksela, S. Pateux, A. Eleftheriadis, and S. Wenger, "System and transport interface of SVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1149–1163, Sep. 2007.
[13] H. Schulzrinne and S. Casner, "RTP profile for audio and video conferences with minimal control," IETF STD 00065, IETF RFC 3551, Jul. 2003.
[14] M. Baugher, D. McGrew, M. Naslund, E. Carrara, and K. Norrman, "The secure real-time transport protocol (SRTP)," IETF RFC 3711, Mar. 2004.
[15] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-driven layered multicast," in *Proc. ACM SIGCOMM*, Aug. 1996, pp. 117–130.
[16] B. Cain, S. Deering, I. Kouvelas, B. Fenner, and A. Thyagarajan, "Internet group management protocol, version 3," IETF RFC 3376, Oct. 2002.
[17] J. Ott, S. Wenger, N. Sato, C. Burmeister, and J. Rey, "Extended RTP profile for real-time transport control protocol (RTCP)-based feedback (RTP/AVPF)," IETF RFC 4585, Jul. 2006.
[18] G. Sullivan and T. Wiegand, *Joint Video Team , SVC Requirements Specified by VCEG* ITU-T SG16 Q16, Doc. JVT-N027, Jan. 2005.
[19] D. Tian, V. K. Malamal Vadakital, M. M. Hannuksela, S. Wenger, and M. Gabbouj, "Improved H.264/AVC video broadcast/multicast," in *Proc. VCIP*, Jul. 2005, pp. 71–82.
[20] T. Schierl, T. Wiegand, and M. Kampmann, "3GPP Compliant adaptive wireless video streaming using AVC," in *Proc. IEEE ICIP*, Sep. 2005, pp. 696–699.
[21] S. Wenger, Y.-K. Wang, and T. Schierl, "RTP payload format for SVC video," IETF Internet Draft draft-wenger-avt-rtp-svc-01.txt, Mar. 2006.
[22] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session initiation protocol," IETF RFC 3261, Jun. 2002.
[23] H. Schulzrinne, A. Rao, and R. Lanphier, "Real time streaming protocol (RTSP)," in *IETF RFC 2326*, Apr. 1998.
[24] G. Camarillo, J. Holler, and H. Schulzrinne, "Grouping of media lines in the session description protocol (SDP)," IETF RFC 3388, Dec. 2002.
[25] J. Rosenberg and H. Schulzrinne, "An offer/answer model with the session description protocol (SDP)," IETF RFC 3264, Jun. 2002.

Stephan Wenger (M'00) received the computer science diploma and the Ph.D. degree from the Technical University of Berlin, Germany, in 1989 and 1995.

Since early 2007, he has been with Nokia's IPR Group, Palo Alto, CA,. Before joining the IPR group, he worked as a Principal Scientist in the Nokia Research Center, and an Adjunct Professor at the Tampere University of Technology, Tampere, Finland. His main research interests lie in the field of media coding and transmission over digital links. Before joining Nokia, he had roles as a project manager at TELES AG, Germany, and technical advisor and consultant for several companies including Polycom, Siemens, Microsoft, and Intel. He has also helped start companies in the field of multimedia coding, and served on the Board of Directors of UB Video Inc. from incorporation until successful acquisition. In addition, he has also been active as a Lecturer and later Assistant Professor at Technical University of Berlin, Germany, from 1989 to 2003. He has authored or co-authored over 100 journals and conference publications, standardization contributions, Internet RFCs, and book chapters. He is very active in the standardization process for new Multimedia technologies, especially in the IETF and the ITU-T. He currently holds two international patents with several pending.

Ye-Kui Wang (M'02) received the B.S. degree in industrial automation from Beijing Institute of Technology, Beijing, China, and the Ph.D. degree in electrical engineering from the Graduate School at Beijing, University of Science and Technology of China, Beijing, China, in 1995 and 2001, respectively.

He is currently a Member of Research Staff in Nokia Research Center, Tampere, Finland. From February 2003 to April 2004, he was a Senior Design Engineer in Nokia Mobile Phone. Before joining Nokia, he worked as a Senior Researcher from June 2001 to January 2003 in Tampere International Center for Signal Processing, Tampere University of Technology. His research interests include video coding and transport, particularly doing those in an error resilient and scalable manner. He has been an active contributor to different standardization organizations, including ITU-T VCEG, ISO/IEC MPEG, JVT, 3GPP SA4, IETF, and AVS. He has acted as an Editor for several (draft) standard specifications, including ITU-T Rec. H.271, and the MPEG file format and the IETF RTP payload format for the scalable video coding (SVC) standard. He has also been a Chair of the special session of Scalable Video Transport in the 15th International Packet Video Workshop (2006). He has co-authored about 200 technical standardization contributions and about 30 academic papers. In addition, he has co-invented over 60 issued and pending patents in the fields of multimedia coding, transport, and application systems.

Thomas Schierl (M'06) received the Dipl.-Ing. degree in computer engineering from the Berlin University of Technology, Berlin, Germany in December 2003.

He has been with Fraunhofer Institute for Telecommunications, Heinrich-Hertz Institute (HHI), Berlin, Germany, since 2003. As Project Manager, he is responsible for various scientific and industry research projects. He conducted different research works on reliable real-time transmission of H.264/MPEG-4 AVC and scalable video coding (SVC) in mobile point-to-point, point-to-multipoint, and broadcast environments like used by 3 GPP and DVB-H. He submitted various inputs on real-time streaming to standardization committees like 3 GPP, JVT/MPEG; ISMA, and IETF. He is one of the authors of the IETF RTP SVC payload format and author of the related SDP draft on signaling layered codecs. Furthermore, he is one of the authors of the SVC Amendment for MPEG-2 TS. In 2007, he was visiting the group of Prof. B. Girod at Stanford University, Stanford, CA, for different research activities. His current research work mainly focuses on developing new real-time streaming techniques for video delivery in mobile ad hoc networks (MANETs). Further research interests are in reliable transmission of real-time media in mobile networks and joint source channel coding, as well as the deployment of SVC in mobile networks.