This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

TCSVT 6212 Copyright (c) 2012 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

Overview of HEVC high-level syntax and reference picture management

Rickard Sjöberg, Ying Chen, *Senior Member, IEEE,* Akira Fujibayashi, Miska M. Hannuksela, *Member, IEEE,* Jonatan Samuelsson, Thiow Keng Tan, *Senior Member, IEEE,* Ye-Kui Wang and Stephan Wenger, *Senior Member, IEEE*

***Abstract*—The increasing proportion of video traffic in telecommunication networks puts an emphasis on efficient video compression technology. High Efficiency Video Coding (HEVC) is the forthcoming video coding standard that provides substantial bit rate reductions compared to its predecessors. In the HEVC standardization process, technologies such as picture partitioning, reference picture management and parameter sets are categorized as "high-level syntax". The design of the high-level syntax impacts the interface to systems, error resilience, and provides new functionalities. This paper presents an overview of the HEVC high-level syntax, including NAL unit headers, parameter sets, picture partitioning schemes, reference picture management, and SEI messages.**

***Index Terms*— High efficiency video coding (HEVC), JCT-VC, parameter set, reference picture list, reference picture set, video coding.**

I. Introduction

During the last decade there has been a steady increase in transmission speeds over fixed and mobile networks, as well as a large capacity increase in storage devices such as hard disks. One could therefore be led to believe that the need for video compression has decreased, whereas in fact the opposite is true: Faster networks and more efficient video compression have enabled previously infeasible applications, such as over-the-top video streaming, triggering an explosion in video traffic [1]. The demand for higher resolution content has further raised video traffic, to the point where, today, it occupies over 50% and 40% of transmitted data in fixed and mobile networks, respectively [2] [3].

High Efficiency Video Coding (HEVC) is the forthcoming video coding standard expected to be published in early 2013. It has been reported that HEVC provides a bit rate reduction of about 50% at the same subjective quality when compared to Advanced Video Coding (H.264/AVC) [4]. While improved compression efficiency is crucial for the success of the codec, the high-level syntax of HEVC also plays an important role, especially in how the features of the codec are exposed to systems. In this paper we describe the high-level design of HEVC, with a focus on those novel high-level elements that provide new functionalities and contribute to improved robustness against transmission errors. Our paper is based on the HEVC Draft International Standard (DIS) [5]. The DIS stage is produced as an interim step in the ISO/IEC approval process, and is generally believed to be mostly stable. Nevertheless, it is conceivable that the final ISO/IEC standard or ITU Recommendation differ in small details from the DIS.

The high-level architecture of video coding standards changed dramatically with H.264/AVC [6]. Previous standards, such as MPEG-2 Video [7], H.263 [8], and MPEG-4 Visual [9] were designed with a continuous video stream in mind. Although picture segmentation (through slices and similar tools) was available, it was not designed to be loss robust with respect to information above the slice header level, such as picture headers. Especially when transmitting video over lossy packet networks, header information had to be repeated by external, "bolt-on" technologies, for example by picture header repetition in H.263/RFC2429 [10] and Header Extension Codes (HEC) in MPEG-4 Visual. Although H.264/AVC can be transmitted as a stream, it was designed from the outset for packet-based transmission. The H.264/AVC Network Abstraction Layer (NAL) provides for self-contained packets, allowing the video layer to be identical for different network environments. The parameter set concept removes the need for header information duplication. Header parameters are signaled in parameter set NAL units, which are referenced by coded video NAL units.

HEVC inherits a number of high-level features from H.264/AVC, such as the NAL unit and parameter set concepts, the use of picture order count (POC), and supplemental enhancement information (SEI) messages for supplemental data signaling to give a few examples. Some high-level features that

Manuscript received April 13, 2012.

R. Sjöberg and J. Samuelsson are with Ericsson Research, Multimedia Technologies, Färögatan 6, SE-16480, Stockholm, Sweden (e-mail: rickard.sjoberg@ericsson.com; jonatan.samuelsson@ericsson.com).

Y. Chen and Y.-K. Wang are with Multimedia R&D and Standards group, QCT, Qualcomm Inc., 5775 Morehouse Dr, San Diego, CA 92121, USA (e-mail: cheny@qualcomm.com; yekuiw@qualcomm.com).

A. Fujibayashi is with NTT DOCOMO, Inc., Research Laboratories, 239-8536 Japan (e-mail: fujibayashi@nttdocomo.com).

M. M. Hannuksela is with Nokia Research Center, Visiokatu 3, 33720 Tampere, Finland (e-mail: miska.hannuksela@nokia.com).

T.K. Tan is with M-Sphere Consulting Pte. Ltd. 808379 Singapore. He is a consultant for NTT DOCOMO, Inc (e-mail: tktan@m-sph.com).

Stephan Wenger is CTO of VidyoCast, a division of Vidyo, Inc, 433 Hackensack Ave., Hackensack N.J. 07601, USA (e-mail: stewe@stewe.org).

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

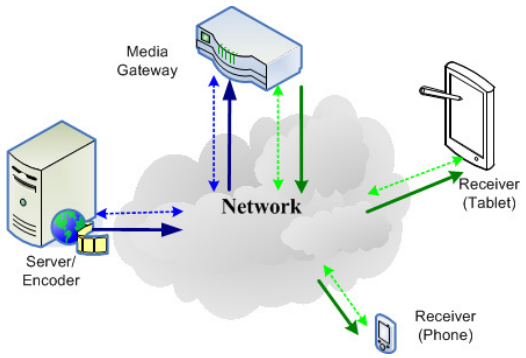

Fig. 1. Video transmission scenario.

```
+---------------+
|0|1|2|3|4|5|6|7|
+-+-+-+-+-+-+-+-+
|F|NRI| NALType |
+---------------+
```

(a) H.264/AVC NAL unit header

```
+---------------+---------------+
|0|1|2|3|4|5|6|7|0|1|2|3|4|5|6|7|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|F|  NALType  |   R6bits  | TIDP|
+---------------+---------------+
```

(b) HEVC NAL unit header

Fig. 2. NAL unit headers of H.264/AVC and HEVC.

exist in H.264/AVC were not included in HEVC, for example flexible macroblock order (FMO), redundant slices, arbitrary slice order (ASO), data partitioning, and SP/SI pictures. A number of new high-level features are introduced, such as the video parameter set (VPS), clean random access (CRA) picture, broken link access (BLA) picture, and temporal sub-layer access (TSA) pictures, the tiles and wavefront tools for parallel processing, the dependent slice tool for reduced delay, and the reference picture set concept for reference picture management.

This paper is organized as follows. Section II provides an overview of the HEVC NAL unit header and discusses future HEVC extensions. In Section III, we describe the HEVC parameter sets, including the new video parameter set. Random access and temporal switching are discussed in Section IV, and Section V addresses the HEVC picture partitioning schemes and parallel processing. Section VI provides an overview of HEVC reference picture management including the reference picture set (RPS) concept. Reference picture lists are described in Section VII and SEI messages are discussed in Section VIII. Finally, Section IX concludes the paper.

II. NAL UNIT HEADER

Similarly to H.264/AVC, HEVC uses a NAL unit based bitstream structure. A coded bitstream is partitioned into NAL units which, when conveyed over lossy packet networks, should be smaller than the maximum transfer unit (MTU) size. Each NAL unit consists of a NAL unit header followed by the NAL unit payload. There are two conceptual classes of NAL units. Video Coding Layer (VCL) NAL units containing coded sample data, e.g., coded slice NAL units, whereas non-VCL NAL units that contain metadata typically belonging to more than one coded picture, or where the association with a single coded picture would be meaningless, such as parameter set NAL units, or where the information is not needed by the decoding process, such as SEI NAL units.

The NAL unit header is designed to co-serve as part of the packet header in RTP based packet networks, and to be processed by Media-Aware Network Elements (MANEs). Further, to enable efficient processing by MANEs, NAL unit headers are of a fixed format (as described below) and do not include variable length codes. All this renders bits in the NAL unit header among the most "expensive" real estate in terms of coding efficiency, and the JCT-VC has been very careful in assigning only those codepoints to the NAL unit header that are required by MANEs for media aware processing.

A typical HEVC video transmission scenario involving MANEs is shown in Fig. 1, wherein a MANE, also known as media gateway, is present between the sender and the receiver(s). Receivers, which can be of different types, receive and decode the bitstream or a part thereof. The MANE is in the signaling context, and therefore aware of key properties of the video bitstream such as profile and level. Based on such information as well as information carried in the NAL unit header, the MANE may perform intelligent media-aware stream adaptation. One example is local repair and/or local redundancy coding—a MANE can, for example, duplicate critical NAL units such as parameter set NAL units on the transmission path to only a subset of receivers with bad connectivity. In order to do that, the MANE needs to know the type of the NAL unit. Another example of such adaptation is bitstream thinning where certain packets are removed from a bitstream based on network conditions or decoder/display capabilities. For example, a MANE can remove slice data belonging to a temporal enhancement layer when it senses congestion between itself and a given receiver.

H.264/AVC specifies a one-byte NAL unit header, which was extended by three bytes in its scalable (H.264/SVC) [11] and multiview (H.264/MVC) extensions [12]. The extensions are required to signal different scalable dimensions, e.g., temporal, spatial, quality or view dimensions.

In HEVC, a two-byte NAL unit header was introduced with the anticipation that this design is sufficient to support the HEVC scalable and 3D Video Coding (3DV) [13] extensions, as well as other future extensions, as briefly described below.

Fig. 2 shows both the H.264/AVC and the HEVC NAL unit headers. In both standards, the forbidden_zero (F) bit must be zero. It is included so to prevent start code emulations in MPEG-2 systems legacy environments. In H.264/AVC, the nal_ref_idc (NRI) was a two-bit codeword. The main motivation for two bits has been the support of different transport priority signaling to support data partitioning. HEVC does not include data partitioning, and the disposable nature of NAL units is indicated by the NAL unit type, rather than by dedicated bits. Therefore, the NRI field has become

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

unnecessary. One bit is used to increase the numeric range of the NAL unit type to 64 types. The other bit is reserved for future extensions.

The second part of the HEVC NAL unit header includes two syntax elements: reserved_zero_6bits (R6bits, 6 bits, one of which is part of the first byte as already described), and temporal_id_plus1 (TIDP, 3 bits). With TIDP, temporal scalability is supported (with the temporal identifier ranging from 0 to 6 inclusive). The reserved_zero_6bits are widely expected to carry some form of layer identification information in future extensions, as described below.

An HEVC bitstream might consist of several temporal sub-layers. Each NAL unit belongs to a specific sub-layer as indicated by the TemporalId (equal to temporal_id_plus1-1). All VCL NAL units of the same picture must belong to the same sub-layer, thus it can be said that the picture itself belongs to that sub-layer. HEVC prohibits any kind of dependency on data in a higher sub-layer in the decoding process of a lower sub-layer. It is required that a sub-bitstream, created from an HEVC bitstream by removing all NAL units with TemporalId higher than a specific value, by itself is a bitstream conforming to HEVC. It is the responsibility of the encoder to ensure that all conditions for bitstream conformance (e.g. buffer restrictions) are fulfilled for each sub-bitstream.

In the first version of HEVC, the reserved_zero_6bits shall be set to '000000' for all NAL units. HEVC version 1 conforming decoders ignore NAL units with reserved_zero_6bits other than '000000'.

It is widely anticipated that, in the scalable or 3DV extensions, the R6bits are renamed as layer_id, to describe all scalability dimensions but the temporal dimension. In 3DV, layer_id would identify view and depth, and in a scalable extension, it would be used to jointly indicate the spatial and quality scalability dimensions. Adaptations based on either one of temporal_id_plus1 and layer_id or both of them are possible by the HEVC NAL unit header design.

III. Parameter sets

HEVC inherits the parameter set concept of H.264/AVC [14] with a few modification and additions. The modifications and additions can be sub-categorized into three groups: (1) additions and modification made necessary by different coding tools of HEVC when compared to H.264/AVC, (2) additions and modifications resulting from operational experience with H.264/AVC, and (3) the newly introduced Video Parameter Set (VPS).

Parameter sets were introduced in H.264/AVC in response to the devastating effects of a loss of the sequence header and picture header, if a picture is partitioned into multiple segments (i.e. slices) and those segments are transported in their own transport unit (e.g. RTP packet)—which is desirable for MTU size matching. The loss of the first packet of a picture, which carries not only the first picture segment data, but also the picture header (and sometimes also the GOP and sequence header), might lead to a completely incorrectly reconstructed picture (and sometimes also the following pictures), even if all other packets were not lost. Some decoder implementations would not even attempt to decode the received packets of a picture, if the packet with the picture header was lost. To combat this vulnerability, transport layer based mechanisms were introduced. For example, the RTP payload format for H.263, specified in RFC 2429 [10], allowed for carrying a redundant copy of the picture header in as many packets as the encoder/packetizer chooses. During the design of H.264/AVC, it was recognized that the vulnerability of a picture header is an architectural issue of the video codec itself, rather than a transport problem, and therefore the parameter set concept was introduced as a fix for the issue.

Parameter sets can be either part of the video bitstream or can be received by a decoder through other means (including out-of-band transmission using a reliable channel, hard coding in encoder and decoder, and so on). A parameter set contains an identification, which is referenced, directly or indirectly, from the slice header as discussed in more detail later. The referencing process is known as “activation”. Depending on the parameter set type, the activation occurs per picture or per sequence. The concept of activation through referencing was introduced, among other reasons, because implicit activation by virtue of the position of the information in the bitstream (as common for other syntax elements of a video codec) is not available in case of out-of-band transmission.

The VPS was introduced to convey information that is applicable to multiple layers as well as sub-layers. H.264/AVC (in all its versions) did not contain a comparable parameter set, requiring a complex modeling of the layering structure for purposes such as capability exchange and session negotiation. In H.264/AVC’s scalable extension, the scalability information SEI message offers approximately the same content, but by its nature of being an SEI message, most of the same information has to be repeated in sequence parameter sets (SPSs), which in some applications scenarios also need to be transmitted out-of-band, and consequently cause increased initial delay, particularly when the retransmission gets involved to guarantee reliability in out-of-band transmission. In cases of broadcast and multicast with in-band transmission of parameter sets, repeating of the same information can be significant overhead as parameter sets need to be repeated at each random access point for tuning in and channel switching. The VPS was introduced to address these shortcomings as well as to enable a clean and extensible high-level design of multi-layer codecs.

Each layer of a given video sequence, regardless of whether they have the same or different SPSs, refer to the same VPS. The VPS conveys information including (1) common syntax elements shared by multiple layers or operation points, in order to avoid unnecessary duplications; (2) essential information of operation points needed for session negotiation, including e.g., profile and level; and (3) other operation point specific information, which doesn’t belong to one SPS, e.g., Hypothetical Reference Decoder (HRD) parameters for layers or sub-layers. The parsing of essential information of each

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

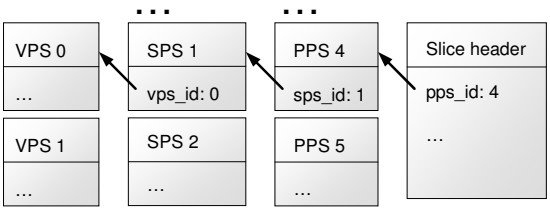

Fig. 3. Slice header referring to a PPS, indirectly an SPS and indirectly a VPS.

operation point doesn't require variable length coding, thus is considered as lightweight for most network elements. It is expected that the VPS extension, to be specified in HEVC extensions, may contain more syntax elements than those in the current VPS, for efficient parameter signaling, flexible and lightweight session negotiation as well as advanced bitstream adaptation, e.g., based on view identifier in 3DV extension.

According to the HEVC DIS [5], some information is duplicated between the VPS and the SPSs belonging to the layer. This duplication was introduced to allow a version 1 decoder to disregard the VPS NAL unit and still have available all information required to decode the bitstream.

In H.264/AVC as well as in HEVC, SPSs contain information which applies to all slices of a coded video sequence. In HEVC, a coded video sequence starts from an instantaneous decoding refresh (IDR) picture, or a BLA picture, or a CRA picture that is the first picture in the bitstream, include all subsequent pictures that are not an IDR or BLA picture. A bitstream consists of one or more coded video sequences. The content of the SPS can be roughly subdivided into six categories: (1) a self-reference (its own ID); (2) decoder operation point related (profile, level, picture size, number sub-layers, and so on); (3) enabling flags for certain tools within a profile, and associated coding tool parameters in case the tool is enabled; (4) information restricting the flexibility of structures and transform coefficient coding; (5) temporal scalability control (similar to H.264/SVC [11]); and (6) Visual Usability Information (VUI), which includes HRD information.

HEVC's picture parameter set (PPS) contains such information which could change from picture to picture. The PPS includes information roughly comparable what was part of the PPS in H.264/AVC, including (1) a self-reference; (2) initial picture control information such as initial quantization parameter (QP), a number of flags indicating the use of, or presence of, certain tools or control information in the slice header; and (3) tiling information.

The slice header contains information that can change from slice to slice, as well as such picture related information that is relatively small or relevant only for certain slice or picture types. The size of slice header may be noticably bigger than the PPS, particular when there are tile or wavefront entry point offsets in the slice header and RPS, prediction weights, or reference picture list modifications are explicitly signaled.

Activation of parameter sets is similar to H.264/AVC. As shown in in Fig. 3, the slice header contains a reference to PPS. The PPS, in turn, contains a reference to the SPS and the SPS contains a reference to the VPS. One common implementation strategy for parameter sets is to keep all parameter sets of a given type (PPS, SPS, and VPS) in tables, whose maximum size is indirectly specified by the numbering range of the parameter set IDs. Under this implementation strategy, the parameter set activation can be as simple as accessing the PPS tables based on information in the slice header, copying the information found into the relevant decoder data structures, and following the reference in the PPS to the relevant SPS, and following the reference in SPS to the relevant VPS. As these operations need to be performed only once per picture (worst case), the operation is lightweight. Similarly, the handling of the reception of a parameter set NAL unit, regardless of its type, is also straightforward. Parameter set NAL units do not contain parsing dependencies—which means they are self-contained and do not require context derived from other NAL units for parsing. Although this may cost a few more bits, it enables straightforward parsing and storage of parameter sets in their respective table entries.

Finally, each type of parameter set contains an extension mechanism, which allows extending the parameter set in future versions of HEVC without breaking backward compatibility and without creating a parsing dependency to the profile/level information carried in the VPS and SPS.

IV. Picture types

As illustrated in Table I, picture types can be classified into the following groups in HEVC: a) random access point (RAP) pictures, b) leading pictures, c) sub-layer access pictures, and d) pictures that do not fall into the three mentioned groups. The picture types and their sub-types as described in Table I are identified by the NAL unit type in HEVC. RAP picture types include IDR picture, BLA picture, and CRA picture, and can be further characterized based on the leading pictures associated with them as indicated in Table I. The CRA picture facilitates decoding beginning from any random access point in the middle of a coded video sequence, which is more efficient from the compression efficiency point of view than inserting an IDR picture (and, thereby, splitting the coded video sequence into two coded video sequences). RAP pictures and leading pictures are reviewed in further details in section IV.A.

Temporal sub-layer access pictures, TSA and STSA, indicate valid temporal sub-layer switching points and are reviewed in section IV.B.

A. *Random access point and leading pictures*

In many video applications, such as broadcasting and streaming, an important feature for users is to be able to switch between different channels and to jump to specific parts of the video with minimum delay. This feature is enabled by inserting RAP pictures in (regular) intervals into the video bitstream. The IDR picture can be used for random access. However, pictures following the IDR in decoding order cannot use pictures decoded prior to the IDR picture as reference. Consequently, bitstreams relying on IDR pictures for random access can have significantly lower coding efficiency (e.g., 6%, as reported in

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

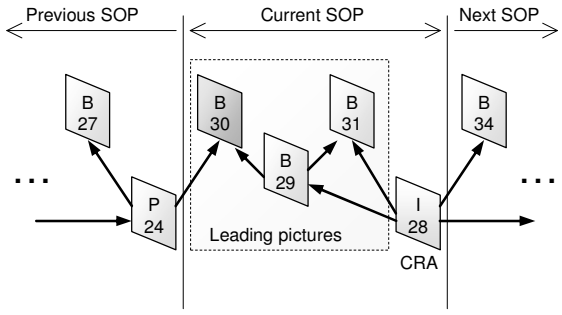

Fig. 4. CRA picture and leading pictures.

[15]). To improve coding efficiency, CRA pictures in HEVC allows pictures that follow the CRA picture in decoding order but precede it in output order to use pictures decoded before the CRA picture as reference and still allow similar clean random access functionality as an IDR picture. Clean random access is ensured by guaranteeing that pictures that follow a CRA picture in both decoding and output order are decodable if random access is performed at the CRA picture.

A typical prediction structure around a CRA picture is shown in Fig. 4, which refers to the concept of Structure of Pictures (SOP) defined as one or more coded pictures consecutive in decoding order, in which the first coded picture in decoding order is a reference picture at the lowest sub-layer and no coded picture except potentially the first coded picture in decoding order is a RAP picture. The relative decoding order of the pictures is illustrated by the numerals inside the pictures. Any picture in the previous SOP has a smaller decoding order than any picture in the current SOP and any picture in the next SOP has a larger decoding order than any picture in the current SOP. The CRA picture, I_{28}, belongs to a SOP which also contains the pictures B_{29}, B_{30}, and B_{31}. They follow the CRA picture in decoding order but precede the CRA picture in output order. These pictures are called leading pictures of the CRA picture. Since B_{29} and B_{31} do not refer to any picture preceding the CRA picture in decoding order, they can be correctly decoded when the decoding starts from the CRA picture and are therefore DLP pictures. Picture B_{30} is a TFD picture which can be correctly decoded if the decoding starts from a RAP picture before the current CRA picture. TFD pictures cannot be correctly decoded when random access from this CRA picture occurs; hence TFD pictures are typically discarded during random access decoding. If I_{28} had been an IDR picture, it would have cleared the DPB and hence B_{30} would not have been able to use P_{24} as a reference. Having I_{28} being a CRA picture makes this possible.

To prevent error propagation from reference pictures that may not be available depending on where the decoding starts, all pictures in the next SOP as shown in Fig. 4, that follow the CRA picture both in decoding order and output order, shall not use any picture that precedes the CRA picture either in decoding order or output order (which includes the leading pictures) as reference.

When a part of a bitstream starting from a CRA picture is included in another bitstream, the TFD pictures associated with the CRA picture cannot be decoded, because some of their reference pictures are not present in the combined bitstream. To make such splicing operation straightforward, the NAL unit type of the CRA picture can be changed to indicate that it is a BLA picture. The TFD pictures associated with a BLA picture are typically not correctly decodable hence should not be output/displayed.

TABLE I. PICTURE TYPE CATEGORIES, PICTURE TYPES AND PICTURE SUB-TYPES

a) Random access point pictures		
IDR	Instantaneous decoding refresh	without associated leading pictures
		may have associated leading pictures
BLA	Broken link access	without associated leading pictures
		may have associated DLP pictures but without associated TFD pictures
		may have associated DLP and TFD pictures
CRA	Clean random access	may have associated leading pictures

b) Leading pictures	
DLP	Decodable leading picture
TFD	Tagged for discard

c) Temporal sub-layer access pictures		
TSA	Temporal sub-layer access	not used for reference in the same sub-layer
		may be used for reference in the same sub-layer
STSA	Step-wise temporal sub-layer access	not used for reference in the same sub-layer
		may be used for reference in the same sub-layer

d) Picture that is not RAP, leading or temporal sub-layer access picture	
	not used for reference in the same sub-layer
	may be used for reference in the same sub-layer

Similar random access functionalities are supported in H.264/AVC with the recovery point SEI message. An H.264/AVC decoder implementation may or may not support the functionality. In HEVC, a bitstream starting with a CRA or BLA picture is considered as a conforming bitstream [16]. When a bitstream starts with a CRA picture, the TFD pictures of the CRA picture may refer to unavailable reference pictures and hence cannot be correctly decoded. However, HEVC specifies that the TFD pictures of the starting CRA picture are not output, hence the name “clean random access”. For establishment of bitstream conformance requirement (e.g., to mandate that the encoders follow the syntax as well as constraints to syntax element values for leading pictures), HEVC specifies a decoding process to generate unavailable reference pictures for decoding of the TFD pictures. However, conforming decoder implementations do not have to follow that decoding process, as long as it can generate identical output compared to when the decoding process is performed from the beginning of the coded video sequence. The same as described above applies to TFD pictures associated with BLA pictures regardless of whether the BLA picture starts a bitstream.

It is worth to note that, in HEVC a conforming bitstream may

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

contain no IDR pictures at all.

B. *Temporal sub-layer access pictures*

The term temporal sub-layer switching point refers to when a picture in a sub-layer has no dependency on any other picture in the same sub-layer that precedes the picture in decoding order.

A MANE may apply bitstream thinning to an HEVC bitstream that is encoded with multiple sub-layers. At any point in the bitstream a MANE can start removing NAL units of higher sub-layers with the knowledge that the pictures in the lower sub-layers are still decodable since the decoding process for the pictures in the lower sub-layers does not depend on the NAL units of the higher sub-layers. The action of starting to remove all NAL units with TemporalId higher than a certain value can be referred to as temporal down-switching. Temporal down-switching is always possible.

The action of starting to forward NAL units of a certain sub-layer that has not been forwarded up until that point can be referred to as temporal up-switching. Temporal up-switching is only possible if none of the pictures in the layer that is switched to depend on any picture in the same sub-layer prior to the point in the bitstream at which the switch was performed.

In H.264/SVC temporal sub-layer switching points can be indicated through temporal_id_nesting_flag in the SPS equal to 1, if all pictures in the sequence with temporal_id greater than 0 are temporal layer switching points, or through the temporal level switching point SEI message, which also contains information for how long period temporal layer M should have been decoded prior to the switch point in order to switch up to temporal layer M + 1 at the switch point.

In HEVC, just as in H.264/SVC, it is possible to indicate that all pictures with TemporalId greater than 0 are sub-layer switching points by setting sps_temporal_id_nesting_flag in the SPS equal to 1. In HEVC there are two picture types, the TSA and STSA picture types, that can be used to indicate temporal sub-layer switching points.

The TSA picture type is defined as a "switch-to" switching point meaning that when a picture is coded as a TSA picture it is possible to perform sub-layer switching to the layer in which the TSA picture is contained (as opposed to a "switch-from" definition for which an indication that it is possible to perform sub-layer switching would be contained in the sub-layer that it is possible to switch from). The TSA picture type imposes restrictions on the TSA picture itself and all pictures in the same sub-layer that follow the TSA picture in decoding order. None of these pictures shall use inter prediction from any picture in the same sub-layer that precedes the TSA picture in decoding order. The TSA definition further imposes restrictions on the pictures in higher sub-layers that follow the TSA picture in decoding order. None of these pictures shall reference a picture that precedes the TSA picture in decoding order if that picture belongs to the same or higher sub-layer as the TSA picture. It is specified that TSA pictures must have TemporalId greater than 0. The STSA is similar to the TSA picture but does not impose restrictions on the pictures in higher sub-layers that follow the STSA picture in decoding order and hence enable up-switching only onto the sub-layer where the STSA picture resides.

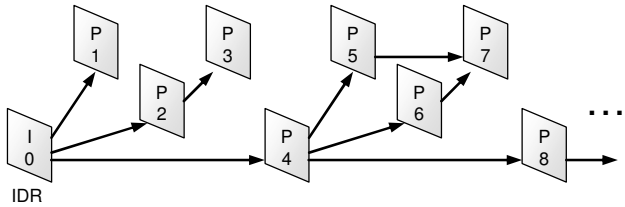

Fig. 5. Coding structure with three temporal layers.

In Fig. 5, horizontal axis represents the output order, the vertical axis represents the sub-layer, and the subscript numbers represents the decoding order (which is the same as the output order for the given example). The arrows represent inter-prediction. The pictures P_1, P_2, P_3, and P_5 are valid sub-layer switching points and may use the TSA picture type. P_6 and P_7 are not switching points since P_7 uses P_5 for prediction.

V. Picture Partitioning Schemes

HEVC includes four different picture partitioning schemes, namely regular slices, dependent slices, tiles, and Wavefront Parallel Processing (WPP), which may be applied for MTU size matching, parallel processing, and reduced end-to-end delay.

Regular slices are similar as in H.264/AVC. Each regular slice is encapsulated in its own NAL unit, and in-picture prediction (intra sample prediction, motion information prediction, coding mode prediction) and entropy coding dependency across slice boundaries are disabled. Thus a regular slice can be reconstructed independently from other regular slices within the same picture (though there may still be interdependencies due to loop filtering operations).

The regular slice is the only tool that can be used for parallelization that is also available, in virtually identical form, in H.264/AVC. Regular slices based parallelization does not require much inter-processor or inter-core communication (except for inter-processor or inter-core data sharing for motion compensation when decoding a predictively coded picture, which is typically much heavier than inter-processor or inter-core data sharing due to in-picture prediction). However, for the same reason, the use of regular slices can incur substantial coding overhead due to the bit cost of the slice header and due to the lack of prediction across the slice border. Further, regular slices (in contrast to the other tools mentioned below) also serve as the key mechanism for bitstream partitioning to match MTU size requirements, due to the in-picture independence of regular slices and that each regular slice is encapsulated in its own NAL unit. In many cases, the goal of parallelization and the goal of MTU size matching place contradicting demands to the slice layout in a picture. The realization of this situation led to the development of the parallelization tools mentioned below.

Dependent slices have short slice headers and allow partitioning of the bitstream at treeblock boundaries without breaking any in-picture prediction. Basically, dependent slices provide fragmentation of regular slices into multiple NAL units, to provide reduced end-to-end delay by allowing a part of a

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

regular slice to be sent out before the encoding of the entire regular slice is finished.

In WPP, the picture is partitioned into single rows of coding tree blocks (CTBs). Entropy decoding and prediction are allowed to use data from CTBs in other partitions. Parallel processing is possible through parallel decoding of CTB rows, where the start of the decoding of a CTB row is delayed by two CTBs, so to ensure that data related to a CTB above and to the right of the subject CTB is available before the subject CTB is being decoded. Using this staggered start (which appears like a wavefront when represented graphically), parallelization is possible with up to as many processors/cores as the picture contains CTB rows. Because in-picture prediction between neighboring treeblock rows within a picture is permitted, the required inter-processor/inter-core communication to enable in-picture prediction can be substantial. The WPP partitioning does not result in the production of additional NAL units compared to when it is not applied, thus WPP is not a tool for MTU size matching. However, if MTU size matching is required, regular slices can be used with WPP, with certain coding overhead.

Tiles define horizontal and vertical boundaries that partition a picture into tile columns and rows. The scan order of CTBs is changed to be local within a tile (in the order of a CTB raster scan of a tile), before decoding the top-left CTB of the next tile in the order of tile raster scan of a picture. Similar to regular slices, tiles break in-picture prediction dependencies as well as entropy decoding dependencies. However, they do not need to be included into individual NAL units (same as WPP in this regard); hence tiles cannot be used for MTU size matching. Each tile can be processed by one processor/core, and the inter-processor/inter-core communication required for in-picture prediction between processing units decoding neighboring tiles is limited to conveying the shared slice header in cases a slice is spanning more than one tile, and loop filtering related sharing of reconstructed samples and metadata. When more than one tile or WPP segment is included in a slice, the entry point byte offset for each tile or WPP segment other than the first one in the slice is signaled in the slice header.

For simplicity, restrictions on the application of the four different picture partitioning schemes have been specified in HEVC. A given coded video sequence cannot include both tiles and wavefronts. For each slice and tile, either or both of the following conditions must be fulfilled: 1) all coded treeblocks in a slice belong to the same tile; 2) all coded treeblocks in a tile belong to the same slice. Finally, a wavefront segment contains exactly one CTB row, and when WPP is in use, if a slice starts within a CTB row, it must end in the same CTB row.

VI. Reference picture sets

A. Introduction

The reference picture set (RPS) concept in HEVC defines how previously decoded pictures are managed in a decoded picture buffer (DPB) in order to be used for reference, i.e., sample data prediction and motion vector prediction. Pictures in the DPB can be marked as "used for short-term reference", "used for long-term reference" or "unused for reference". Once a picture has been marked "unused for reference" it can no longer be used for prediction, and when it is no longer needed for output it can be removed from the DPB. The RPS concept for reference picture management is fundamentally different from the reference picture management of previous video coding standards. Instead of signaling relative changes to the DPB, the status of the DPB is signaled in every slice. A goal in the HEVC development for reference picture management was to have a basic level of error robustness in all standard-conforming bitstreams and decoders.

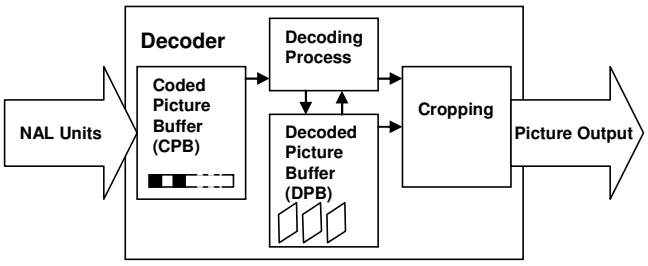

Fig. 6. HRD buffer model.

B. Background

H.263 Annex U and H.264/AVC [6] constituted a technological shift in the area of reference picture usage for motion compensated prediction of image sample data in video coding. The support for flexible reference picture selection was introduced in Annex N and U of H.263 and also adopted in H.264/AVC and allowed for up to 16 reference pictures to be used providing improved compression efficiency as well as improved possibilities for error (packet-loss) recovery using feed-back channels.

H.264/AVC as well as HEVC defines an HRD, which a decoder and describes the usage of a coded picture buffer (CPB) and a decoded picture buffer (DPB) as illustrated in Fig. 6. In both H.264/AVC and HEVC the decoding order of coded pictures is the same as the order in which the coded pictures occur in the bitstream. Both standards support an output order of decoded pictures that is different from the decoding order of the pictures. Each picture is associated with a picture order count (POC) value that represents the output order.

In H.264/AVC, there are two types of reference pictures, short-term and long-term. A reference picture may be marked as "unused for reference" when it becomes no longer needed for prediction reference. The conversion among these three statuses (short-term, long-term, and unused for reference) is controlled by the decoded reference picture marking process. There are two alternative decoded reference picture marking mechanisms, the implicit sliding window process and the explicit memory management control operation (MMCO) process. The sliding window process marks a short-term reference picture as "unused for reference" when the number of reference frames is equal to a given maximum number (max_num_ref_frames in SPS). The short-term reference pictures are stored in a first-in, first-out manner so that the most recently decoded short-term pictures are kept in the DPB.

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

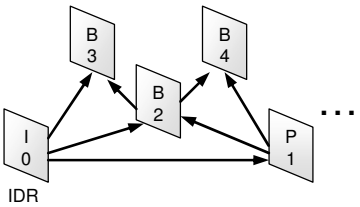

Fig. 7. Coding structure for reference picture set example

The explicit MMCO process may include multiple MMCO commands. An MMCO command may mark one or more short-term or long-term reference picture as "unused for reference", mark all the pictures as "unused for reference", or mark the current reference picture or an existing short-term reference picture as long-term, and assign a long-term picture index to that long-term picture.

In H.264/AVC the reference picture marking operations as well as the processes for output and removal of pictures from the DPB are performed after a picture has been decoded. Some aspects related to H.264/AVC reference picture marking mechanisms are discussed below.

1) Gaps in frame_num and "non-existing" pictures

In H.264/AVC each reference picture is associated with a number, FrameNum (derived from the frame_num syntax element in the slice header), which indicates the decoding order. Normally this number increases by one for each reference picture but gaps in FrameNum may be allowed (by setting the sequence level parameter gaps_in_frame_num_allowed_flag to one), such that an encoder or a MANE can deliver a bitstream in which FrameNum increases by more than one for a reference picture relative to the preceding reference picture in decoding order. This was allowed in order to support temporal scalability. A decoder that receives a sequence with gaps in FrameNum shall create "non-existing" pictures to fill the gap. The "non-existing" pictures are assigned with FrameNum values in the gap and are considered as reference pictures during decoded reference picture marking but will not be used for output (hence not displayed). The "non-existing" pictures ensure that the status of the DPB, with respect to the FrameNum of the pictures residing in it, is the same for a decoder that received the pictures as for a decoder that did not receive the pictures.

2) Loss of a reference picture when using sliding window

When a reference picture is lost in H.264/AVC, a decoder can try to conceal the picture (and possibly report the loss to the encoder if a feedback channel is available) given that the loss is detected. If gaps in FrameNum are disallowed, a discontinuity in FrameNum values indicates an unintentional loss of a reference picture. If gaps in FrameNum are allowed, a discontinuity in FrameNum values may be caused by either intentional removal of temporal layers or sub-sequences or an accidental picture loss, and decoders should infer a picture loss only if a "non-existing" picture is referred in the inter prediction process. The POC of a concealed picture may not be known which can cause the decoder to use incorrect reference pictures without detecting any errors when decoding B-pictures.

TABLE II. REFERENCE PICTURE SET EXAMPLE

Picture	RPS {reference picture, used by current picture}
I_0	-
P_1	{ I_0, 1}
B_2	{ I_0, 1}, { P_1, 1}
B_3	{ I_0, 1}, { P_1, 0}, { B_2, 1}
B_4	{ P_1, 1}, { B_2, 1}

3) Loss of a reference picture with MMCO

In H.264/AVC, when losing a reference picture that contains an MMCO command marking a short-term reference picture as "unused for reference", then the status of reference pictures in the DPB becomes incorrect and consequently, reference picture lists for a few pictures following the lost picture may become incorrect.

If a picture containing MMCO commands related to long-term reference pictures is lost there is a risk that the number of long-term reference pictures in the DPB is different from what it would have been if the picture was received, resulting in an "incorrect" sliding window process for all the following pictures. That is, the encoder and decoder will contain a different number of short-term reference pictures resulting in out-of-sync behavior of the sliding window process. What makes the situation even worse is that a decoder will not necessarily know that the sliding window process is out-of-sync.

C. Reference picture set concept

HEVC introduces a completely new approach for reference picture management, referred to as reference picture set (RPS) or buffer description [17].

The most fundamental difference with the RPS concept compared to MMCO/sliding window of H.264/AVC is that for each particular slice a complete set of the reference pictures that are used by the current picture or any subsequent picture must be provided. Thus, a complete set of all pictures that must be kept in the DPB for use by the current or future picture is signaled. This is different from the H.264/AVC scheme where only relative changes to the DPB are signaled. With the RPS concept, no information from earlier pictures in decoding order is needed to maintain the correct status of reference pictures in the DPB.

D. Reference picture set example

An example of a coding structure is shown in Fig. 7. The RPSs for the pictures in Fig. 7 are shown in Table II. The first picture in decoding order is an IDR picture, I_0, for which no reference picture set is signaled since it is the first picture in the coded video sequence and no picture that precedes the IDR picture in decoding order can be used for reference by the IDR picture or by any picture that follows the IDR picture in decoding order. The second picture in decoding order, P_1, uses I_0 for reference. It must therefore include I_0 in its RPS. Picture B_2 uses both I_0 and P_1 for reference so they are both included in the RPS of B_2.

B_3 uses I_0 and B_2 for reference so they are included in the RPS

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

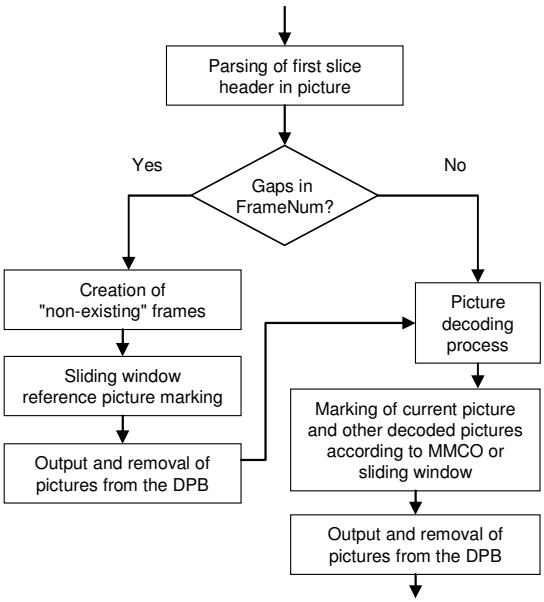

Fig. 8. H.264/AVC decoding order.

of B_3. But also P_1 must be included since this picture will be used for reference for future pictures. Finally, picture B_4 will use B_2 and P_1 for prediction. Note that the RPS of B_4 does not contain I_0. Since I_0 is not listed, it will be marked "unused for reference". This means that I_0 cannot be used for reference by B_4 or by any picture that follows B_4 in decoding order.

E. Order of picture decoding and DPB operations

The order of picture decoding and DPB operations in HEVC is changed compared to H.264/AVC in order to exploit the advantages of RPS and improve error resilience. In H.264/AVC picture marking and buffer operations (both output and removal of decoded pictures from the DPB) are applied after a current picture has been decoded, with the exception of when gaps in FrameNum are detected. In HEVC, the RPS is first decoded from a slice header of the current picture, then picture marking and buffer operations are applied before decoding the current picture.

The order of different processes for H.264/AVC and HEVC are shown in Fig. 8 and Fig. 9, respectively. As can be seen, the high-level decoding processes based on syntax elements in the slice header and above in HEVC have been significantly simplified compared to H.264/AVC thanks to the introduction of the RPS concept.

F. Signaling of reference picture sets

Each slice header in HEVC must include parameters for signaling of the RPS for the picture containing the slices. The only exception is that no RPS is signaled for IDR slices, instead the RPS is inferred to be empty. For I slices that do not belong to an IDR picture, an RPS may be provided, even if they belong to an I picture since there may be pictures following the I picture in decoding order which use inter-prediction from pictures that preceded the I picture in decoding order. The number of pictures in an RPS shall not exceed the DPB size limit as

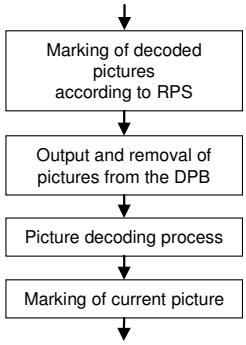

Fig. 9. HEVC decoding order.

specified by the sps_max_dec_pic_buffering syntax element in the SPS.

Each picture is associated with a POC value that represents the output order. The slice headers contain a fixed-length codeword, pic_order_cnt_lsb, representing the least significant bits of the full POC value, also known as the POC LSB. The length of the codeword is signaled in the SPS and can be between 4 and 16 bits. The RPS concept uses POC to identify reference pictures. Besides its own POC value, each slice header directly contains or inherits from the SPS a coded representation of the POC values of each picture in the RPS.

The RPS for each picture consists of five different lists of reference pictures, also referred to the five RPS subsets: RefPicSetStCurrBefore consists of all short-term reference pictures that are prior to the current picture in both decoding order and output order, and that may be used in inter prediction of the current picture. RefPicSetStCurrAfter consists of all short-term reference pictures that are prior to the current picture in decoding order, that succeed the current picture in output order, and that may be used in inter prediction of the current picture. RefPicSetStFoll consists of all short-term reference pictures that may be used in inter prediction of one or more of the pictures following the current picture in decoding order, and that are not used in inter prediction of the current picture. RefPicSetLtCurr consists of all long-term reference pictures that may be used in inter prediction of the current picture. RefPicSetLtFoll consists of all long-term reference pictures that may be used in inter prediction of one or more of the pictures following the current picture in decoding order, and that are not used in inter prediction of the current picture.

The RPS is signaled using up to three loops iterating over different types of reference pictures; short-term reference pictures with lower POC value than the current picture, short-term reference pictures with higher POC value than the current picture and long-term reference pictures. In addition, a flag (used_by_curr_pic_X_flag) is sent for each reference picture indicating whether the reference picture is used for reference by the current picture (included in one of the lists RefPicSetStCurrBefore, RefPicSetStCurrAfter or RefPicSetLtCurr) or not (included in one of the lists RefPicSetStFoll or RefPicSetLtFoll).

In the example in Fig. 10, the picture B_{14} contains exactly one picture in each of the five RPS subsets; P_8 is in

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

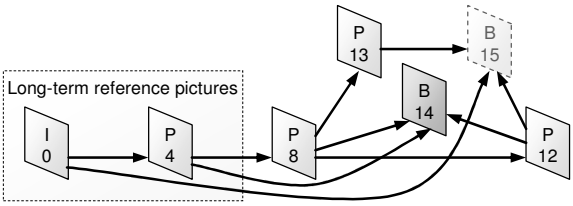

Fig. 10 Example of a picture, B_{14}, with entries in all subsets of the RPS.

RefPicSetStCurrBefore since it is before in output order and used by B_{14}, P_{12} is in RefPicSetStCurrAfter since it is after in output order and used by B_{14}, P_{13} is in RefPicSetStFoll since it is a short-term reference picture that is not used by B_{14} (but must be kept in the DPB since it is used by B_{15}), P_4 is in RefPicSetLtCurr since it is a long-term reference picture that is used by B_{14}. I_0 is in RefPicSetLtFoll since it is a long-term reference picture that is not used by the current picture (but must be kept in the DPB since it is used by B_{15}).

1) Signaling of short-term reference pictures in an RPS

The short-term part of an RPS may be included directly in the slice header. Alternatively, the slice header may contain only a syntax element which represents an index, referencing to a predefined list of RPSs sent in the active SPS. The short-term part of an RPS can be signaled using either of two different schemes; Inter RPS, as described in the next subsection, or Intra RPS, as described in this subsection. When Intra RPS is used, num_negative_pics and num_positive_pics are signaled representing the length of two different lists of reference pictures. These lists contain the reference pictures with negative POC difference and positive POC difference compared to the current picture, respectively. Each element in these lists is encoded with a variable length code representing the difference in POC value relative to the previous element in the list minus one. For the first picture in each list the signaling is relative to the POC value of the current picture minus one.

2) Signaling short-term pictures in an RPS using Inter RPS scheme

When encoding the recurring RPSs in the sequence parameter set, it is possible to encode the elements of one RPS with reference to another RPS already encoded in the sequence parameter set. This is referred to as Inter RPS. There are no error robustness problems associated with this method as all the RPSs of the sequences parameter set are in the same NAL unit.

The Inter RPS syntax exploits the fact that the RPS of the current picture can be predicted from the RPS of a previously decoded picture. This is because all the reference pictures of the current picture must either be reference pictures of the previous picture or the previously decoded picture itself. It is only necessary to indicate which of these pictures should be reference pictures and be used for the prediction of the current picture [18].

Therefore the syntax comprises of the following: an index pointing to the RPS to use as a predictor, a delta_POC to be added to the delta_POC of the predictor to obtain the delta POC of the current RPS, and a set of indicators to indicate which pictures are reference pictures and whether they are only used for the prediction of future pictures.

3) Signaling of long-term reference pictures in an RPS

Encoders that would like to exploit the use of long-term reference pictures must set the SPS syntax element long_term_ref_pics_present_flag to one. Long-term reference pictures can then be signaled in the slice header by fixed-length codewords, poc_lsb_lt, representing the least significant bits of the full POC value of each long-term picture. Each poc_lsb_lt is a copy of the pic_order_cnt_lsb codeword that was signaled for a particular long-term picture. It is also possible to signal a set of long-term pictures in the SPS as a list of POC LSB values. The POC LSB for a long-term picture can then be signaled in the slice header as an index to this list.

The delta_poc_msb_cycle_lt_minus1 syntax element can additionally be signaled to enable the calculation of the full POC distance of a long-term reference picture relative to the current picture. It is required that the codeword delta_poc_msb_cycle_lt_minus1 is signaled for each long-term reference picture that has the same POC LSB value as any other reference picture in the RPS.

G. Picture marking

Before picture decoding, there will typically be a number of pictures present in the DPB. Some of them may be available for prediction and thus marked as "used for reference". Others may be unavailable for prediction but waiting for output, thus marked as "unused for reference". When the slice header has been parsed, a picture marking process is carried out before the slice data is decoded. Pictures that are present in the DPB and marked as "used for reference" but are not included in the RPS are marked "unused for reference". Pictures that are not present in the DPB but are included in the reference picture set are ignored if the used_by_curr_pic_X_flag is equal to zero. However, if the used_by_curr_pic_X_flag instead is equal to one, this reference picture was intended to be used for prediction in the current picture but is missing. Then an unintentional picture loss is inferred and the decoder should take appropriate actions.

After decoding the current picture, it is marked "used for short-term reference".

H. POC, FrameNum and "non-existing" pictures

HEVC does not include FrameNum signaling and processes for generating and handling of "non-existing" pictures. In H.264/AVC FrameNum was used to identify pictures in the DPB when performing buffer operations (i.e. MMCO). FrameNum was also used to detect gaps in the decoding order of reference pictures (i.e. due to temporal scaling) in order to generate "non-existing" pictures. The "non-existing" pictures were introduced to H.264/AVC in order to keep correct status of the reference pictures in the DPB in the case of temporal scaling. In HEVC the RPS contains the status of all reference pictures in the DPB, thus generating and handling of "non-existing" pictures are not needed. Within the RPS concept, reference pictures are identified by POC. This has the advantage

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

that output order is known even for reference pictures that have been removed from the bitstream or unintentionally lost. Since POC is used for reference picture identification there is no need to signal FrameNum.

I. Error resilience aspects of RPS

Since all reference pictures that are used by the current picture must be included in the RPS there is no risk that the loss of one or more packets of data will lead to undetected usage of an incorrect reference picture. As soon as there is a reference picture with used_by_curr_pic_X_flag set to one in the RPS but no corresponding reference picture in the DPB the decoder will know that an unintentional picture loss has occurred. The decoder side can react to the detection of a lost picture as appropriate in the specific application, e.g. create a concealed picture or report the loss to the encoder side through a feedback channel. How to handle picture losses is not in the scope of the HEVC specification. In H.264/AVC the POC of a lost reference picture is typically not known. In HEVC, the RPS contains the POC of all reference pictures, which improves the possibility of creating a good concealment of lost reference pictures.

VII. Reference picture lists

In HEVC, the term inter prediction is used to denote prediction derived from data elements (e.g. sample values or motion vectors) of reference pictures other than the current decoded picture. Like in H.264/AVC, a picture can be predicted from multiple reference pictures. The reference pictures that are used for inter prediction are organized in one or more reference picture lists. The reference index identifies which of the reference pictures in the list should be used for creating the prediction signal.

A single reference picture list, List 0, is used for a P slice and two reference picture lists, List 0 and List 1 are used for B slices. Similar to H.264/AVC, the reference picture list construction in HEVC includes reference picture list initialization and reference picture list modification.

In H.264/AVC the initialization process for List 0 is different for P slice (for which decoding order is used) and B slices (for which output order is used). In HEVC, output order is used in both cases.

Reference picture list initialization creates default List 0 and List 1 (if the slice is a B slice) based on three RPS subsets: RefPicSetStCurrBefore, RefPicSetStCurrAfter, and RefPicSetLtCurr. Short-term pictures with earlier (later) output order are firstly inserted into the List 0 (List 1) in ascending order of POC distance to the current picture, then short-term pictures with later (earlier) output order are inserted into the List 0 (List 1) in ascending order of POC distance to the current picture, and finally the long-term pictures are inserted at the end. In terms of RPS, for List 0, the entries in RefPicSetStCurrBefore are inserted in the initial list, followed by the entries in RefPicSetStCurrAfter. Afterwards, the entries in RefPicSetLtCurr, if available, are appended.

In HEVC, the above process is repeated (reference pictures that have already been added to the reference picture list are added again) when the number of entries in a list is smaller than the target number of active reference pictures (signaled in the picture parameter set or slice header). When the number of entries is larger than the target number the list is truncated.

After a reference picture list has been initialized, it may be modified such that the reference pictures for the current picture may be arranged in any order, including the case where one particular reference picture may appear in more than one position in the list, based on the reference picture list modification commands. When the flag that indicates if the presence of list modifications is set to one, a fixed number (equal to the target number of entries in the reference picture list) of commands are signaled, and each command inserts one entry for a reference picture list. A reference picture is identified in the command by the index to the list of reference pictures for the current picture derived from the RPS signaling. This is different from reference picture list modification in H.264/AVC, wherein a picture is identified either by the picture number (derived from the frame_num syntax element) or the long-term reference picture index, and it is possible that fewer commands are needed e.g. for swapping the first two entries of an initial list or inserting one entry at the beginning of the initial list and shifting the others.

A reference picture list is not allowed to include any reference picture with TemporalId greater than the current picture.

VIII. Supplemental Enhancement Information (SEI)

The SEI mechanism enables encoders to include such metadata in the bitstream that is not required for correct decoding of the sample values of the output pictures but can be used for various other purposes, such as picture output timing, displaying, as well as loss detection and concealment. Encoders can include any number of SEI NAL units in an access unit, and each SEI NAL unit may contain one or more SEI messages. The HEVC standard includes the syntax and semantics for several SEI messages, but the handling of the SEI messages is not specified, because they do not affect the normative decoding process. One reason to have SEI messages in the HEVC standard is to enable supplemental data being interpreted identically in different systems using HEVC. Specifications and systems using HEVC may require encoders to generate certain SEI messages or may define specific handling of particular types of received SEI messages.

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

TABLE III. OVERVIEW OF SEI MESSAGES

SEI message	Purpose
Buffering period	Initial delays for hypothetical reference decoder (HRD) operation
Picture timing	Picture output time and picture/sub-picture removal time for HRD operation
Pan-scan rectangle	Displaying at a different picture aspect ratio (PAR) than the PAR of the output pictures
Filler payload	Adjusting the bitrate to meet specific constraints
User data registered User data unregistered	SEI messages specified by external entities
Recovery point	Additional information for clean random access. Gradual decoding refresh.
Scene information	Information about scene changes and transitions
Full-frame snapshot	Indication to label the associated decoded picture as a still-image snapshot of the video content
Progressive refinement segment	Indicates that certain consecutive pictures represent a progressive refinement of the quality of a picture rather than a moving scene
Film grain characteristics	Enables decoders to synthesize film grain
Deblocking filter display preference	Recommends whether or not displayed pictures should undergo the in-loop deblocking filter process
Post-filter hint	Provides suggested post-filter coefficients or correlation information for post-filter design
Tone mapping information	Remapping to another color space than that used or assumed in encoding
Frame packing arrangement	Packing of stereoscopic video into an HEVC bitstream
Display orientation	Specifies flipping and/or rotation that should be applied to the output pictures when they are displayed
Field indication	Provides information related to interlaced video content and/or field coding, e.g. indicates whether the picture is a progressive frame, a field, or a frame containing two interleaved fields
Decoded picture hash	Checksum of the decoded picture, which may be used for error detection
Sub-picture timing	Sub-picture removal time for HRD operation
Active parameter sets	Provides information on active VPS, SPS, etc.
Structure of Pictures description	Describes the temporal and inter prediction structure of the bitstream

Table III lists the SEI messages specified in HEVC and briefly describes their purposes. The SEI mechanism and several SEI messages in HEVC were inherited from H.264/AVC and hence not described in detail in this paper. The field indication, decoded picture hash, the structure of pictures (SOP) description SEI messages were introduced in the HEVC standard. The field indication SEI message enables the use of interlaced video content with HEVC. It is similar to the respective message in the H.263 standard [8]. The decoded picture hash SEI message contains a checksum derived from the decoded samples of the associated picture, hence enabling error detection. The sub-picture timing SEI message provides removal times for sub-pictures in hypothetical reference decoder operation for very low-delay applications, such as remote screen sharing. The active parameter set SEI message provides the identifier of the active VPS, such that a MANE does not need to parse slice header, PPS, and SPS to find the active VPS, which includes information that is essential for stream adaptation and other intelligent media-aware operations.

The SOP description SEI message describes the structure of the bitstream through RPSs and is described in further details below. The motivation to design the SOP description SEI message arose from several use cases where the knowledge of the temporal and inter prediction structure is helpful [19]. For example, a gateway can use the SOP information in bit rate adaptation to determine a set of inter-related pictures that can be dropped without affecting the decoding of the forwarded bitstream. Such bitstream trimming can have a finer granularity than the sub-bitstream extraction based on TemporalId and can therefore be more suitable for subtle temporary bit rate adaptation. The SOP description SEI message resides in the first access unit of a SOP. It provides the following information for each picture in the SOP: an indication whether the picture is a reference or a non-reference picture, the TemporalId value of the picture, the short-term RPS index used by the picture, and the picture order count relative to the first picture of the SOP. These pieces of information represent the temporal structure and the inter prediction hierarchy of the SOP comprehensively.

IX. CONCLUSION

We presented various aspects of HEVC's high-level syntax, which form the system and transport interfaces of the HEVC codec to applications. While based on H.264/AVC's concepts, several changes were introduced in HEVC. Compared to H.264/AVC, a number of source coding tools commonly associated with high-level designs have been eliminated, such as flexible macroblock order, data partitioning, and redundant slices. New concepts, including video parameter set, parallel coding tools, additional random access point, and temporal sub-layer switching picture types, were shown to be beneficial to the design and have been included,. Some mechanisms available in H.264/AVC were kept in spirit, while the details of their implementation were changed, occasionally severely. Those include the reference picture set architecture, and the NAL unit header syntax. We believe that with those modifications, one cornerstone has been set for the success of HEVC in future video coding applications.

REFERENCES

[1] Sandvine, "Sandvine Global Internet Phenomena Report: Fall 2011, " 2011.

[2] Cisco, "Visual Networking Index: Forecast and Methodology, 2010-2015", June 2011.

[3] Allot MobileTrends: Global Mobile Broadband Report, H2, 2011. http://www.allot.com/index.aspx?id=4019&fileID=4117

[4] T.K. Tan, A. Fujibayashi, Y. Suzuki and J. Takiue, "[AHG 8] Objective and subjective evaluation of HM5.0," Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-H0116, 8th Meeting: San Jose, USA, 1-10 February, 2012.

[5] B. Bross, W.J. Han, J. Ohm, G. Sullivan, T. Wiegand, "High efficiency video coding (HEVC) text specification draft 8" Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

JTC1/SC29/WG11, JCTVC-J1003_d2.doc, 10th Meeting: Stockholm, Sweden, 11-20 July, 2012.

[6] Advanced Video Coding for Generic Audiovisual Services, ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2003.

[7] "Generic Coding of Moving Pictures and Associated Audio Information - Part 2: Video," ITU-T and ISO/IEC JTC 1, ITU-T Recommendation H.262 and ISO/IEC 13 818-2 (MPEG-2), 1994.

[8] "Video Coding for Low Bit Rate Communication," ITU-T, ITU-T Recommendation H.263 version 2, 2005.

[9] "Coding of audio-visual objects—Part 2: Visual," ISO/IEC 14492-2 (MPEG-4 Visual), ISO/IEC JTC 1, Version 3: May 2004.

[10] C. Bormann, L. Cline, G. Deisher, T. Gardos, C. Maciocco, D. Newell, J. Ott, G. Sullivan, S. Wenger, and C. Zhu, "RTP payload format for the 1998 version of ITU-T Rec. H.263 video (H.263+)", RFC2429, May 1999.

[11] Y.-K. Wang, M. M. Hannuksela, S. Pateux, A. Eleftheriadis, and S. Wenger, "System and transport interface of SVC," *IEEE Trans. on Circuits and Systems for Video Technology,* vol.17, no.9, pp.1149-1163, Sep. 2007.

[12] Y. Chen, Y. -K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, "The Emerging MVC Standard for 3D Video Services," *EURASIP Journal on Advances in Signal Processing*, Volume 2009, Article ID 786015.

[13] "Call for Proposals on 3D Video Coding Technology," ISO/IEC JTC1/SC29/WG11/N12036, Geneva, Switzerland, Mar. 2011.

[14] S. Wenger, "H.264/AVC over IP," *IEEE Trans. on. Circuits and Systems for Video Technology*, vol. 13, no. 7, July 2003.

[15] A. Fujibayashi and T.K. Tan, "Random access support for HEVC," Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-D234, 4th Meeting: Daegu, Korea, 20-28 January, 2011.

[16] Y. -K. Wang, Y. Chen, M. Karczewicz, and J. L. Chen, "On bitstreams starting with CRA pictures," Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-H0496, 8th Meeting: San Jose, CA, USA, 1–10 February, 2012.

[17] R. Sjöberg, J. Samuelsson, "Absolute signaling of reference pictures," Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-F493, 6th Meeting: Torino, IT, 14-22 July, 2011.

[18] T.K. Tan and CS Boon, "AHG21: Inter reference picture set prediction syntax and semantics", Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-G198, 7th Meeting: Geneva, CH, 21-30 November, 2011.

[19] J. Boyce, D. Hong, and A. Eleftheriadis, "High layer syntax to improve support for temporal scalability," Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCTVC-D200, 4th Meeting: Daegu, Korea, 20-28 January, 2011.

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

Rickard Sjöberg received his Master of Science degree in computer science from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1997.

He has been with Ericsson in Sweden and the UK (one year) since 1996, working with still image and video coding research, real-time video codec implementations of H.263, MPEG-4 part2 and H.264/AVC as well as subjective video encoding optimizations. Since 1997, he has been an active contributor with more than a hundred proposals to the video coding standards of ITU-T, MPEG, JVT and JCT-VC. Currently he is a Senior Specialist in Video Coding at Multimedia Technologies, Ericsson Research, Stockholm, Sweden, working as the technical leader of the Ericsson 2-D video coding research.

Ying Chen (M'05 - SM'11) received his B.S. degree in Applied Mathematics and M.S. degree in Electrical Engineering & Computer Science, both from Peking University, in 2001 and 2004 respectively. He received his PhD in Computing and Electrical Engineering from Tampere University of Technology (TUT), Finland, in 2010.

He is currently a Senior Staff Engineer at Qualcomm Incorporated, San Diego, CA, USA. Dr. Chen joined Qualcomm in Mar. 2009. His earlier working experiences include Researcher in TUT and Nokia Research Center, Finland from 2006 to Feb. 2009 and Research Engineer in Thomson Corporate Research, Beijing, from 2004 to 2006.

Dr. Chen has been actively contributing to MPEG, JVT and JCT-VC, on Scalable Video Coding, Multiview Video Coding, and 3D Video (3DV) Coding extensions of H.264/AVC, as well as high-level syntax and 3DV extension of HEVC. Dr. Chen has also been involved in standardization activities of MPEG systems, including MPEG File Format, MPEG-2 Systems and DASH (Dynamic Adaptive Streaming over HTTP). Dr. Chen has co-authored a couple of hundreds of standard contribution documents to JVT, JCT-VC and MPEG and around 30 academic papers in the fields of image processing, video coding, and video transmission.

Akira Fujibayashi received the B.S. and M.S. degrees in Electronics, Information and Communication Engineering from Waseda University, Japan in 2002 and 2004, respectively.

He has been with NTT DOCOMO, Inc. since 2004. He is an active participant at the ITU-T/ISO/IEC Joint Collaborative Team for Video Coding (JCT-VC) standardization activities and has published some contributions in JCT-VC.

He is interested in the area of image and video coding, perceptual coding, video enlargement technologies.

Miska M. Hannuksela (M'03) received his Master of Science degree in engineering and Doctor of Science degree in technology from Tampere University of Technology, Finland, in 1997 and 2010, respectively.

He has been with Nokia since 1996 in different roles including research manager/leader positions in the areas of video and image compression, end-to-end multimedia systems, as well as sensor signal processing and context extraction. Currently he works as Distinguished Scientist in Multimedia Technologies in Nokia Research Center, Tampere, Finland. He has published about 100 journal and conference papers and hundreds of standardization contributions in JCT-VC, JVT, MPEG, 3GPP, and DVB. He has granted patents from more than 60 patent families. His research interests include video compression, multimedia communication systems and formats, user experience and human perception of multimedia, as well as sensor signal processing.

Dr. Hannuksela received an award of the best doctoral thesis of Tampere University of Technology in 2009 and Scientific Achievement Award nominated by the Centre of Excellence of Signal Processing, Tampere University of Technology, in 2010. He has been an Associate Editor in IEEE Transactions on Circuits and Systems of Video Technology since 2010.

Jonatan Samuelsson received his Master of Science degree in computer science from Royal Institute of Technology (KTH), Stockholm, Sweden, in 2007.

He has been with Ericsson at Multimedia Technologies in Stockholm, Sweden, since 2007. His work has included different aspects of video compression with specific emphasis on software implementations of real time video codecs. Areas of focus have been rate-controllers, adaptive QP algorithms, packetization and transport of video, video buffering and error feedback. He currently holds a position as Senior Researcher, is active in standardization and has published a dozen contributions in JCT-VC.

Thiow Keng Tan (S'89–M'94–SM'03) received the Bachelor of Science and Bachelor of Electrical and Electronics Engineering degrees from Monash University, Australia in 1987 and 1989, respectively. He later received the Ph.D. degree in Electrical Engineering in 1994 from the same university.

He currently consults for NTT DOCOMO, Inc., Japan. He is an active participant at the video subgroup of the ISO/IEC JCT1/SC29/WG11 Moving Picture Experts Group (MPEG), the ITU-T SG16 Video Coding Experts Group (VCEG) as well as the ITU-T/ISO/IEC Joint Video Team (JVT) and the ITU-T/ISO/IEC Joint Collaborative Team for Video Coding (JCT-VC) standardization activities. He has also served on the editorial board of the IEEE Transaction on Image Processing.

Dr. Tan was awarded the Dougles Lampard Electrical Engineering Medal for his Ph.D. thesis and 1st prize IEEE Region 10 Student Paper Award for his final year undergraduate project. He was also awarded three ISO certificates for outstanding contributions to the development of the MPEG-4 standard. He is the inventor in at least 50 granted US patents. His research interest is in the area of image and video coding, analysis and processing.

Ye-Kui Wang received his BS degree in industrial automation in 1995 from Beijing Institute of Technology, Beijing, China, and his PhD degree in electrical engineering in 2001 from the Graduate School in Beijing, University of Science and Technology of China, Beijing, China. He is currently a Senior Staff Engineer at Qualcomm, San Diego, CA, USA. His earlier working experiences and titles include Multimedia Standards Manager at Huawei Technologies, Bridgewater, NJ from Dec. 2008 to Aug. 2011, various positions, including Principal Member of Research Staff, at Nokia Corporation, Tampere, Finland from Feb. 2003 to Dec. 2008, and Senior Researcher at Tampere International Center for Signal Processing, Tampere University of Technology, Tampere, Finland from Jun. 2001 to Jan. 2003. His research interests include video coding, multimedia transport and systems.

Dr. Wang has been an active contributor to various multimedia standards, including video codecs, file formats, RTP payload formats and streaming systems, developed in ITU-T VCEG, ISO/IEC MPEG, JVT, JCT-VC, 3GPP SA4, IETF and AVS. He has been an editor for several standard specifications, including ITU-T H.271, Scalable Video Coding (SVC) File Format, Multiview Video Coding (MVC), IETF RFC 6184 and IETF RFC 6190. He has co-authored over 300 technical standardization contributions, about 50 academic papers, and over 100 granted or pending patents or patent applications in his areas of interest.

Stephan Wenger (M'99) received his Diploma and PhD in computer science from Technische Universitat Berlin, Germany, in 1989 and 1995, respectively.

After working as assistant professor at TU Berlin, various consulting roles, and five years in Nokia's Research Center and IPR/legal groups, since 2007 he is Chief Technology Officer of VidyoCast, a division of Vidyo, in Hackensack, NJ, USA. He has published dozens of Journal and Conference papers, Internet RFCs, and hundreds of standardization contributions to many standards setting organizations, as well as more than 30 granted or pending patents and patent applications. He is interested in the interface between media coding and transport over a variety of networks.

Copyright (c) 2011 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.