ETSI TR 126 948 V17.0.0 (2022-05)

TECHNICAL REPORT

Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; 5G; Study on video enhancements in 3GPP multimedia services (3GPP TR 26.948 version 17.0.0 Release 17)

Reference

RTR/TSGS-0426948vh00

Keywords

5G,GSM,LTE,UMTS

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

If you find a security vulnerability in the present document, please report it through our Coordinated Vulnerability Disclosure Program:
https://www.etsi.org/standards/coordinated-vulnerability-disclosure

Notice of disclaimer & limitation of liability

The information provided in the present deliverable is directed solely to professionals who have the appropriate degree of experience to understand and interpret its content in accordance with generally accepted engineering or other professional standard and applicable regulations.
No recommendation as to products and services or vendors is made or should be implied.
No representation or warranty is made that this deliverable is technically accurate or sufficient or conforms to any law and/or governmental rule and/or regulation and further, no representation or warranty is made of merchantability or fitness for any particular purpose or against infringement of intellectual property rights.
In no event shall ETSI be held liable for loss of profits or any other incidental or consequential damages.

Any software contained in this deliverable is provided "AS IS" with no warranties, express or implied, including but not limited to, the warranties of merchantability, fitness for a particular purpose and non-infringement of intellectual property rights and ETSI shall not be held liable in any event for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption, loss of information, or any other pecuniary loss) arising out of or related to the use of or inability to use the software.

Copyright Notification

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Legal Notice

This Technical Report (TR) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities. These shall be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between 3GPP and ETSI identities can be found under http://webapp.etsi.org/key/queryform.asp.

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Contents

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

1 Scope

The present document reports the study on video enhancements in 3GPP multimedia services. It firstly provides an overview of the video codecs and their configurations specified for existing 3GPP multimedia services, namely 3GP-DASH (TS 26.247 [1]), PSS (TS 26.234 [2]), MBMS (TS 26.346 [3]), MTSI (TS 26.114 [4], including multi-stream multiparty video conferencing), MMS (TS 26.140 [5]), and IMS Messaging and Presence (TS 26.141 [6]). Then use cases on video enhancements for existing 3GPP multimedia services are discussed, including a discussion on potential codec solutions for each of the use cases. To enable drawing conclusions, simulation conditions and simulation results for comparisons of different codecs and their configurations are provided. Performance is evaluated in typical 3GPP service environments taking into account bandwidth, quality and complexity. Based on the performance results, conclusions are made in terms of recommendations for support of enhanced video capabilities for 3GPP multimedia services.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TS 26.247: "Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)".

[2] 3GPP TS 26.234: "Transparent end-to-end Packet-switched Streaming Service (PSS); Protocols and codecs".

[3] 3GPP TS 26.346: "Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs".

[4] 3GPP TS 26.114: "IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction".

[5] 3GPP TS 26.140: "Multimedia Messaging Service (MMS); Media formats and codecs".

[6] 3GPP TS 26.141: "IP Multimedia System (IMS) Messaging and Presence; Media formats and codecs".

[7] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[8] ITU-T Recommendation H.263 (01/2005): "Video coding for low bit rate communication".

[9] ITU-T Recommendation H.264 (V9) (02/2014): "Advanced video coding for generic audiovisual services".

[10] ITU-T Recommendation H.265 (V3) (04/2015): "High efficiency video coding".

[11] J. Boyce, Y. Yan, J. Chen, and A. K. Ramasubramonian, "Overview of SHVC: Scalable Extensions of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. Circuits Syst. Video Technol.*, August 2015, to be published.

[12] R. Sjöberg, Y. Chen, A. Fujibayashi, M. M. Hannuksela, J. Samuelsson, T. K. Tan, Y.-K. Wang, and S. Wenger, "Overview of HEVC High-Level Syntax and Reference Picture Management," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1858–1870, Dec. 2012.

[13] 3GPP TR 26.906: "Evaluation of HEVC for 3GPP Services".

[14] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, August 2015, to be published.

[15] ITU-R Recommendation BT.709-6 (06/2015): "Parameter values for the HDTV standards for production and international programme exchange".

[16] ITU-R Recommendation BT.2020-1 (06/2014): "Parameter values for ultra-high definition television systems for production and international programme exchange".

[17] 3GPP TR 26.923: "Study on Media Handling Aspects of IMS-based Telepresence".

[18] 3GPP TS 26.948: "Video enhancements for 3GPP Multimedia Services".

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in TR 21.905 [7] apply.

3.2 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [7] and the following apply.

3D-HEVC	Three-Dimension High Efficiency Video Coding
3GPP	3rd Generation Partnership Project
API	Application Program Interface
AU	Access Unit
AVC	Advanced Video Coding
BD	Bjontegaard Delta
BL	Base Layer
BLA	Broken Link Access
BM-SC	Broadcast-Multicast - Service Centre
BP	Bitstream Partition
CABAC	Context Adaptive Binary Arithmetic Coding
CPB	Coded Picture Buffer
CRA	Clean Random Access
CTU	Coding Tree Unit
DASH	Dynamic Adaptive Streaming over HTTP
DPB	Decoded Picture Buffer
EL	Enhancement Layer
GOP	Group of Pictures
HDTV	High-Definition TeleVision
HEVC	High Efficiency Video Coding
HLS	High-Level Syntax
HRD	Hypothetical Reference Decoder
IDR	Instantaneous Decoding Refresh
IMS	IP Multimedia Subsystem
INBLD	Independent Non-Base Layer Decoding
ILP	Inter-Layer Prediction
IRAP	Intra Random Access Point
MANE	Media Aware Network Element
MBMS	Multimedia Broadcast/Multicast Service
MBMS-GW	MBMS Gateway
MMVC	Multi-stream Multiparty Video Conferencing
MMS	Multimedia Messaging Service
MRFP	Multimedia Resource Function Processor
MSB	Most Significant Bits
MTSI	Multimedia Telephony Service for IMS

MTU	Maximum Transmission Unit
MV	Motion Vector
MV-HEVC	MultiView High Efficiency Video Coding
NAL	Network Abstraction Layer
OLS	Output Layer Set
POC	Picture Order Count
PPS	Picture Parameter Set
PSNR	Peak Signal Noise Ratio
PSS	Packet-switched Streaming Service
PTL	Profile, Tier, and Level
QP	Quantization Parameter
RAP	Random Access Point
RPL	Reference Picture List
RPS	Reference Picture Set
SAO	Sample Adaptive Offset
SEI	Supplemental Enhancement Information
SPS	Sequence Parameter Set
SVC	Scalable Video Coding
SHVC	Scalable High efficiency Video Coding
TMVP	Temporal Motion Vector Prediction
UE	User Equipment
UHDTV	Ultra High-Definition TeleVision
VPS	Video Parameter Set
VUI	Video Usability Information
WPP	Wavefront Parallel Processing

4 Overview of video codecs specified for existing 3GPP multimedia services

The video support in 3GPP multimedia services in Release-12 is provided in Table 1.

Table 1: Video support in 3GPP multimedia services in Release-12

	H.263 [8]	H.264/AVC [9]	HEVC/H.265 [10]
DASH and PSS	Profile 0 Level 45	Constrained Baseline Profile, Level 1.3 Progressive High Profile Level 3.1 Frame-packed stereoscopic 3D video (H.264 Constrained Baseline Profile Level 1.3 or Progressive High Profile Level 3.1) Multiview stereoscopic 3D video (H.264 Stereo High Profile Level 3.1), but not for RTP based transmission	Main Profile, Main Tier, Level 3.1
MBMS		Constrained Baseline Profile, Level 1.3 Progressive High Profile Level 3.1 Frame-packed stereoscopic 3D video (H.264 Constrained Baseline Profile Level 1.3 or Progressive High Profile Level 3.1)	Main Profile, Main Tier, Level 3.1
MTSI and IMS Messaging and Presence		H.264 Constrained Baseline Profile, Level 1.2 Constrained Baseline Profile, Level 3.1	Main Profile, Main Tier, Level 3.1
MMS	Profile 0 Level 45	Constrained Baseline Profile, Level 1.3 Progressive High Profile Level 3.1 Frame-packed stereoscopic 3D video (H.264 Constrained Baseline Profile Level 1.3 or Progressive High Profile Level 3.1)	Main Profile, Main Tier, Level 3.1

5 Overview of SHVC

5.0 General

Scalable High efficiency Video Coding (SHVC) refers to the scalable extension of H.265/HEVC, specified in Annex H of the H.265/HEVC specification [10]. This clause provides an overview of SHVC, including the basic SHVC architecture, SHVC systems and transport interfaces, a comparison of SHVC and Scalable Video Coding (SVC), the scalable extension of H.264/AVC [9], and SHVC decoder and encoder complexity analyses.

5.1 Basic SHVC architecture

Inter-layer prediction is employed in a scalable system to improve the coding efficiency of the enhancement layers. In addition to the spatial and temporal motion-compensated predictions that are available in a single-layer codec, inter-layer prediction (ILP) in SHVC uses the reconstructed video signal from a reference layer to predict the current enhancement layer. Inter layer prediction in SHVC is built upon the so called "reference index" framework. With this framework, the collocated reconstructed picture from the reference layer is treated as a long-term reference picture, and is assigned a reference index (or reference indices) in the reference picture list(s) along with other temporal reference pictures in the current layer. Then, ILP is achieved at the block-level (Prediction Unit-level) by setting the value of the ref_idx syntax element to correspond to the inter-layer reference picture(s) in the reference picture list(s).

Figure 1 shows the SHVC codec architecture from the decoder's perspective. SHVC supports more layers, but for ease of explanation, Figure 1 only describes a two-layer scalable system consisting of the base layer (BL) and one enhancement layer (EL).

As will be discussed later in clause 8.3, one fundamental benefit of the "reference index" based SHVC architecture is that it allows the EL codec to maintain the same block level logics as a single-layer HEVC decoder. The EL codec differs from a single-layer HEVC decoder only at the high level syntax level, i.e. at or above the slice header level. Hence, the EL decoder is labeled as HEVC* in Figure 1 to reflect this. To achieve efficient inter-layer prediction, inter-layer processing is applied to the reconstructed BL pictures retrieved from the BL Decoded Picture Buffer (BL DPB); afterwards, the processed pictures are put into the EL Decoded Picture Buffer (EL DPB) and used as inter-layer reference pictures for predictive coding of the EL pictures. SHVC applies different forms of inter-layer processing depending on the types of scalability between the two layers. For example, for spatial scalability, resampling of texture and/or motion information from the reference layer is applied. By adjusting the sample bit depth during resampling, SHVC also supports bit depth scalability. For color gamut scalability, a color mapping process is applied. Further detailed discussion of inter layer processing modules supported by SHVC can be found in [11].

As shown in Figure 1, the base layer bitstream can be sent either as part of the SHVC bitstream "in-band", or obtained via "external means" in an "out-of-band" manner. In the former case when the base layer is embedded within the SHVC bitstream, the input bitstream is de-multiplexed into two separate layers. The base layer (BL) bitstream is sent to the base layer decoder and the enhancement layer (EL) bitstream is sent to the EL decoder. The BL decoder is an HEVC decoder; in the Scalable Main and Scalable Main 10 profiles currently defined in SHVC, the BL decoder conforms to either the HEVC Main or Main 10 profile. Additionally, SHVC also allows the base layer bitstream to be provided via external means, for example, through other system-level multiplexing methods. This latter function can be used to support the use case when the base layer bitstream is coded using a non-HEVC single-layer codec, for example, using H.264/AVC, MPEG-2, or even non-standardized codecs. Accordingly, this is also referred to as hybrid codec scalability. For hybrid codec scalability, the BL decoding operations are outside of the scope of the SHVC decoder. After decoding, the reconstructed BL pictures are provided to the SHVC decoder, along with some information associated with the BL pictures. The remaining SHVC decoding operations are the same as the former case with the embedded BL bitstream. The SHVC decoder applies inter-layer processing to the reconstructed BL pictures to obtain the inter-layer reference pictures for predictive coding of the EL video pictures. It is worth noting that, although BL bitstreams provided via external means are generally expected to be non-HEVC coded, an HEVC-coded BL bitstream can be provided via external means as well.

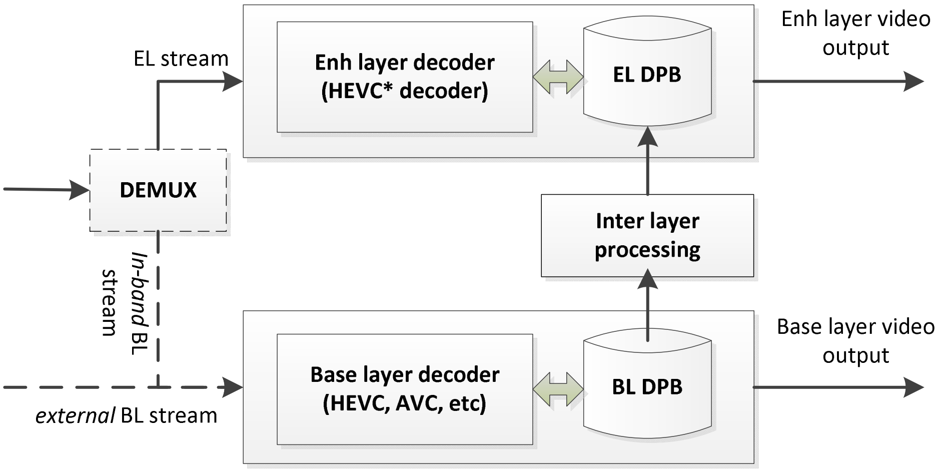

Figure 1: SHVC decoder architecture

5.2 Systems and transport interfaces of SHVC

5.2.1 Introduction

The systems and transport interfaces of a video codec, also referred to as high-level syntax (HLS), are an integral part of a video codec. An important part is the network abstraction layer (NAL), providing a (generic) interface of a video codec to (various) networks/systems. HLS topics include (but are not limited to) bitstream structure and coded data units structures; parameter sets signalling; support of random access and stream adaptation; error resilience; coded and decoded picture buffer management and buffering model (a.k.a. hypothetical reference decoder or HRD); scalability; byte stream format; profile and level signalling; signalling of supplemental enhancement information (SEI) and video usability information (VUI); extensibility and backward compatibility.

HEVC (single-layer coding) HLS was designed with significant consideration of extensibility mechanisms. These are also referred to as hooks, which basically allow future extensions that would be backward compatible to earlier versions of the standard. Important HLS hooks in HEVC include: a) Inclusion of layer identifier (ID) in the NAL unit header, whereby the same NAL unit header syntax applies to both HEVC single-layer coding and its multi-layer extensions; b) Introduction of the video parameter set (VPS), which was introduced mainly for use with multi-layer extensions, as VPS contains cross-layer information; c) Introduction of the layer set concept and the associated signalling of multi-layer HRD parameters; d) Addition of extensibility for all types of parameter sets and slice header, which allows the same syntax structures to be used for both the base layer and enhancement layers without defining new NAL unit types and to be further extended in the future when needed.

A common HLS framework has been jointly developed for SHVC and MV-HEVC (which is largely applicable to 3D-HEVC as well). This clause focuses on the new HLS features developed for the three multi-layer HEVC extensions compared to HEVC single layer coding HLS, for which an overview can be found in [12] and TR 26.906 [13]. More details of SHVC Systems and transport interfaces be found in [11] and [14].

5.2.2 Parameter Set and Slice Segment Header Extensions

The VPS has been extended by adding the VPS extension structure to the end, which mainly includes information on: a) Scalability type and division of NAL unit header layer ID to scalability IDs; b) Layer dependency, dependency type, and independent layers; c) Layer sets and output layer sets; d) Sub-layers and inter-layer dependency of sub-layers; e) Profile, tier, and level (PTL); f) Representation format (resolution, bit depth, color format, etc.); g) Decoded picture buffer (DPB) size; h) cross-layer video usability information (VUI), which includes information on cross-layer picture type alignment, cross-layer intra random access point (IRAP) picture alignment, bit rate and picture rate of layer sets, video signal format (color primaries, transfer characteristics, etc.), usage of tiles and wavefronts and other enabled parallel processing capabilities, and additional HRD parameters.

It should be noted that the VPS applies to all layers, while in the AU decoding order dimension it applies from the first AU where it is activated up to the AU when it is deactivated. Different layers (including the base layer and a non-base layer) may either share the same SPS or use different SPSs. Pictures of different layers or AUs can also share the same picture parameter set (PPS) or use different PPSs. To enable sharing between sequence parameter set (SPS) and PPS, all SPSs share the same value space of their SPS IDs, regardless of the layer ID values in their NAL unit headers; the same is true for PPSs.

Among other smaller extensions, the slice segment header has been extended in a backward compatible manner by adding the following information: a) The discardable flag that indicates whether the picture is used for at least one of inter prediction and inter-layer prediction or neither (when neither applies the picture can be discarded without affecting the decoding of any other pictures, in the same layer or other layers); b) A flag that indicates whether an IDR picture is a bitstream splicing point (if yes, then pictures from earlier AUs would be unavailable as references for pictures of any layer starting from the current AU); c) Information on lower-layer pictures used by the current picture for inter-layer prediction; and d) POC resetting and POC most significant bits (MSB) information. The latter two sets of information are used as the basis for derivation of the inter-layer reference picture set (RPS) and for guaranteeing cross-layer POC alignment, both of which are discussed later.

5.2.3 Layer and Scalability Identification

Each layer is associated with a unique layer ID, for which the value will be increasing across pictures of different layers in decoding order within an AU. In addition, a layer is associated with scalability IDs specifying its content, which are derived from the VPS extension and denoted as view order index and auxiliary ID.

All layers of a view have the same view order index. The view order index is required to be increasing in decoding order of views. Furthermore, a view ID value is signalled for each view order index, which can be chosen without constraints, but should indicate the view's camera position (e.g. in a linear setup).

The auxiliary ID signals whether a layer is an auxiliary picture layer carrying depth, alpha or other user defined auxiliary data. By design choice, auxiliary picture layers have no normative impact on the decoding of non-auxiliary picture layers (denoted as primary picture layers).

5.2.4 Layer sets

The concept of layer sets was already introduced in HEVC version 1. A layer set is a set of independent decodable layers that contains the base layer. Layer sets are signalled in the base part of the VPS. During the development of the common multi-layer HLS, two related concepts, namely output layer sets (OLSs) and additional layer sets, were further introduced. An OLS is a layer set for which the target output layers are specified (non-target-output layers are for example those layers that are used only for inter-layer prediction but not for output/display). For example, an OLS can have two layers for output (e.g. stereoscopic viewing) but contain three layers. An HEVC single-layer decoder would only process one target output layer, i.e. the base layer, regardless of how many layers the layer set contains. This is the reason why the concept of OLS layer set was not needed in HEVC version 1.

An additional layer set is a set of independent decodable layers that does not contain the base layer. For example, if a bitstream contains two simulcast (i.e. independently coded) layers, then the non-base layer itself can be included in an additional layer set. This concept can also be used for signalling the PTL for auxiliary picture layers, which are usually coded independently from the primary picture layers. For example, a depth or alpha (i.e. transparency) auxiliary picture layer can be included in an additional layer set and indicated to conform to the Monochrome (8 bit) profile, regardless of which single-layer profile the base (primary picture) layer conforms to. Without such a design, many more profiles would need to be defined to handle all the combinations of auxiliary picture layers with single-layer profiles. To realize the benefits of this design, an independent non-base layer rewriting process was specified, which "transcodes" independent non-base layers to a bitstream that conforms to a single-layer profile.

By design choice, an additional layer set is allowed to contain more than one layer, e.g. three layers with layer ID values equal to 3, 4, and 5, where the layer with layer ID equal to 3 is an independent non-base layer. Along with this, a bitstream extraction process for additional layer sets was specified. While the extracted sub-bitstream does not contain a base layer, it is still a conforming bitstream, i.e. the multi-layer extensions of HEVC allow for a conforming multi-layer bitstream to not contain the base layer, and compliant decoding of the bitstream may not involve the base layer at all.

5.2.5 Profile, Tier, and Level (PTL)

Compared to earlier multi-layer video coding standards, a fundamentally different approach was taken for MV-HEVC and SHVC for the specification and signalling of interoperability points (i.e. PTL in the context of HEVC and its extensions). Rather than specifying PTL for an operation point that contains a set of layers, in MV-HEVC and SHVC, PTL is specified and signalled in a layer specific manner. Consequently, a decoder that is able to decode two-layer bitstreams with 720p@30fps at the base layer and 1080p@60fps at the enhancement layer should express its capability as a list of two PTLs equivalent to {Main profile Main tier Level 3.1, Scalable Main profile Main tier Level 4.1}. A key advantage of this design is that it facilitates easy decoding of multiple layers by reusing single-layer decoders. If PTL was specified for the two layers together, then the decoder would need to be able to decode the two-layer bitstreams with both the base and enhancement layers of 1080p@60fps. In other words, over provisioning of resources would be required.

Another related innovation is the definition of the independent non-base layer decoding (INBLD) capability, which is associated with the decoding capability of one or more of the single-layer profiles. The INBLD capability, when supported, indicates the capability of a decoder to decode an independent non-base layer that is indicated in the active VPSs and SPSs to conform to a single-layer profile and is the layer with the smallest nuh_layer_id value in an additional layer set. Compared to conventional single-layer decoders, such single-layer decoders can also parse some multi-layer syntaxes such as VPS extension and handle NAL units with layer ID greater than zero. It is recommended that, when expressing the capabilities of a decoder for one or more single-layer profiles, whether the INBLD capability is supported for those profiles should also be expressed. As can be seen from its connection to additional layer sets, the INBLD capability is also part of the entire solution for auxiliary picture layers introduced to HEVC version 2.

5.2.6 RPS and Reference Picture List Construction

In addition to the five RPS lists (RefPicSetStCurrBefore, RefPicSetStCurrAfter, RefPicSetStFoll, RefPicSetLtCurr, and RefPicSetLtFoll) defined in HEVC version 1, two more RPS lists, RefPicSetInterLayer0 and RefPicSetInterLayer1 (denoted as RpsIL0 and RpsIL1, respectively), were introduced to contain inter-layer reference pictures. Given a current picture, those inter-layer reference pictures are included into two sets depending on whether they have view ID values greater or smaller than the current picture. If the base view has greater view ID than the current picture, then those with greater view IDs are included into RpsIL0 and those with smaller view IDs into RpsIL1, and vice versa. The derivation of RpsIL0 and RpsIL1 is based on VPS extension signalling (of layer dependency and inter-layer dependency of sub-layers) as well as slice header signalling (of lower-layer pictures used by the current picture for inter-layer prediction).

When constructing the initial reference picture list 0 (i.e. RefPicListTemp0), pictures in RpsIL0 are inserted immediately after pictures in RefPicSetStCurrBefore, and pictures in RpsIL1 are inserted last, after pictures in RefPicSetLtCurr. When constructing the initial reference picture list 1 (i.e. RefPicListTemp1), pictures in RpsIL1 are inserted immediately after pictures in RefPicSetStCurrAfter, and pictures in RpsIL0 are inserted last, after pictures in RefPicSetLtCurr. Otherwise the reference picture list construction process stays the same as for HEVC single-layer coding.

5.2.7 Random Access, Layer Switching, and Bitstream Splicing

Compared to AVC, HEVC provides more flexible and convenient random access and splicing operations, by allowing conforming bitstreams to start with a clean random access (CRA) or broken link access (BLA) picture. In addition, MV-HEVC and SHVC support non-cross-layer aligned IRAP pictures of any type (IDR, CRA, or BLA), and a conforming bitstream can start with any type of IRAP access unit, including an IRAP AU where the base layer picture is an IRAP picture while (some of) the enhancement layer pictures are non-IRAP pictures. This allows easy splicing of multi-layer bitstreams at any type of IRAP AU and random accessing from such AU. Non-cross-layer aligned IRAP pictures also allow for flexible layer switching.

To support non-cross-layer aligned IRAP pictures, the multi-layer POC design needs to ensure cross-layer POC alignment within any AU. Cross-layer POC alignment is needed to ensure that the in-layer RPS derivation and the output order of pictures of target output layers are correct.

The multi-layer HEVC design allows extremely flexible layering structures. Basically, a picture of any layer may be absent at any AU. For example, the highest layer ID value can vary from AU to AU, which was disallowed in SVC and MVC. Such flexibilities imposed a great challenge on the multi-layer POC design. In addition, although a bitstream after layer or sub-layer switching is not required to be conforming, the design should still enable a conforming decoding behaviour to work with layer and sub-layer switching, including cascaded switching behaviour. This is achieved by a POC resetting approach.

The basic idea of POC resetting is to reset the POC value when decoding a non-IRAP picture (as determined by the POC derivation process in HEVC version 1), such that the final POC values of pictures of all layers of the AU are identical. In addition, to ensure that POC values of pictures in earlier AUs are also cross-layer aligned and that POC delta values of pictures within each layer remain proportional to the associated presentation time delta values, POC values of pictures in earlier AUs are reduced by a specified amount.

To work with all possible layering structures as well as some picture loss situations, the POC resetting period is specified based on the POC resetting period ID that is optionally signalled in the slice header. Each non-IRAP picture that belongs to an AU that contains at least one IRAP picture will be the start of a POC resetting period in the layer containing the non-IRAP picture. In that AU, each picture would be the start of a POC resetting period in the layer containing the picture. POC resetting and decreasing of POC values of same-layer pictures are applied only for the first picture within each POC resetting period, such that these operations would not be performed more than necessary; otherwise POC values would be messed up.

5.2.8 Hybrid Codec Scalability and Multiview/3D Support

The HEVC multi-layer extensions support the base layer being coded by other codecs, e.g. AVC. A simple approach was taken for this functionality by specifying the base layer being provided by an external means, i.e. not specified by the standard. Basically, except for information on the representation format and whether the base layer is a target output layer as signalled in the VPS extension, no other information about the base layer is included in the bitstream (as input to the enhancement-layer decoder). Decoder implementations can implement an application program interface (API) to accept the sample values of the decoded base layer picture for each AU plus some other minimum amount of information required for decoding the enhancement layer pictures, including whether it is an IRAP picture and, if yes, the IRAP picture type (IDR, CRA, or BLA). The base layer pictures and this latter information may be provided by the base layer decoder through the API, however this is not part of the enhancement-layer decoder specified by HEVC version 2. The output of base layer pictures is the responsibility of the base layer decoder, and output synchronization between a base layer picture and an enhancement layer picture in the same AU, when needed, is externally controlled (e.g. by using presentation timestamps). The association of a base layer decoded picture to an AU is also the responsibility of external means.

5.2.9 Hypothetical Reference Decoder (HRD)

The main new developments of HRD of the common HLS compared to HEVC version 1 include the following three aspects relevant for MV- and 3D-HEVC. Firstly, the bitstream conformance tests specified in HEVC version 1 are classified into two sets and a third set is additionally specified. The first set of tests is for testing the conformance of the entire bitstream and its temporal subsets. The second set of bitstream conformance tests is for testing the conformance of the layer sets specified by the active VPS and their temporal subsets. For the first and second sets of tests, only the base layer pictures are decoded and other pictures are ignored by the decoder. The third set of tests is for testing the conformance of the OLSs specified in the VPS extension and their temporal subsets.

The second aspect is the introduction of bitstream partition (BP) specific coded picture buffer (CPB) operations, wherein each BP contains one or more layers, and CPB parameters for each BP can be signalled and applied. These parameters can be utilized by transport systems that transmit different sets of layers in different physical or logical channels; one extreme example is one channel for each layer. The layer specific CPB parameters are also a basis for defining the semantics of layer specific PTL. The third aspect is the layer specific DPB management operations, where each layer exclusively uses its own sub-DPB. To ensure the design works with (cascaded) layer switching behavior, sharing of a particular memory unit across layers is disallowed.

5.2.10 SEI Messages

SEI messages in HEVC version 1 have been adapted to be applicable in the multi-layer contexts, in a backward compatible fashion, some of them with significant semantics changes. In addition, some new SEI messages are specified that apply to all multi-layer HEVC extensions.

The layers not present SEI message can indicate which layers are dropped. The inter-layer constrained tile sets SEI message, which can indicate cross-layer region of interest coding based on tiles. The BP nesting SEI message and the BP initial arrival time SEI message can be used to signal BP buffering parameters for CPB operations. The sub-bitstream property SEI message provides the bit rate information for a sub-bitstream created by discarding those pictures in the layers that do not belong to the output layers of the OLSs specified by the active VPS and that do not affect the decoding of the output layers. The alpha channel information SEI message provides information about alpha channel sample values and post-processing applied to the decoded alpha plane auxiliary pictures, and one or more associated primary pictures. Other new SEI messages that apply to all multi-layer HEVC extensions include the temporal motion vector prediction constraints SEI message and the frame-field information SEI message.

5.3 A comparison of SHVC and SVC

The scalable extension to H.264/AVC, commonly referred to as SVC, is the most recent scalable video coding standard preceding SHVC. In this clause a brief comparison between SHVC and SVC is provided.

The first difference between SHVC and SVC is the indication (i.e. signalling) of inter layer prediction. The previous SVC standard uses a base_mode_flag signalled at the macroblock level to enable inter-layer prediction, and makes further block-level operation changes depending on the value of base_mode_flag. In comparison, SHVC uses the "reference index" based framework, as discussed above. Because the ref_idx syntax element already exists in the single-layer HEVC standard at the Prediction Unit-level, referencing an inter-layer reference picture can be carried out in a transparent manner at the block-level. In other words, all block-level logic, including parsing and interpretation of the syntax elements, decoding and reconstruction, loop filtering, and other related processes, of the EL codec can be kept unchanged from those of a single-layer HEVC codec. Any necessary changes to the EL decoder, denoted as HEVC* in Figure 1, are only at slice header-level and above, that is, at the high level syntax level. By keeping the detailed block-level operations compatible with a single-layer codec, SHVC can be implemented by reusing/repurposing most parts of an existing HEVC implementation; thus, the implementation cost of SHVC can be reduced significantly.

Secondly, to achieve inter-layer prediction, the only BL information that the EL needs to access is the reconstructed pictures from the BL DPB, which includes the reconstructed texture samples, and typically also the BL motion information. As the BL DPB needs to be provided as an open interface in a single-layer codec implementation, such scalable codec architecture requires no change at all to the BL codec, and allows the BL codec to essentially operate as a black box. In contrast, in addition to using the BL reconstructed texture to predict the EL, SVC also applies cross-layer syntax prediction and cross-layer residual prediction, the implementation of which requires the BL codec to be redesigned to provide much more information than what it would generally need to provide as a single-layer codec. As such, implementation of the SHVC codec is much simpler than that of SVC. Further, operating the BL codec as a "black box" also allows SHVC to support hybrid codec scalability discussed earlier, thus providing expanded backward compatibility support.

Thirdly, SVC applies a single-loop decoding constraint, whereby when decoding a bitstream containing multiple layers, full decoding of reference layer(s) may not be required (i.e. limited decoding may be sufficient) in order to fully decode the current enhancement layer picture. This limited reference layer decoding can include decoding of multiple layers of intra-coded macroblocks, but not decoding of multiple layers of inter-coded macroblocks. As a consequence, constrained intra prediction will be used for any layer which will be used as a reference layer, meaning that spatial intra prediction in the reference layer can only predict from intra-coded spatial neighbours and not from inter-coded spatial neighbours. This constraint can negatively impact the coding efficiency. A disadvantage of single loop decoding is that arbitrary down-switching at any picture is not supported. For example, consider a two-layer spatial scalable bitstream containing a lower resolution BL and a higher resolution EL. When the decoder receives both layers and operates in single loop decoding mode, it outputs only the high resolution EL, and does not fully decode the lower resolution BL. If a Media Aware Network Element (MANE) removes the enhancement layer of the bitstream in the middle of a coded sequence, this single loop decoder is unable to switch to decoding just the lower resolution BL, because the previous temporal reference pictures of the BL are not available. In comparison, the SHVC architecture is based on multi-loop decoding. It allows all samples from the reference layers within the specified reference regions to be used in inter-layer prediction. By not imposing the constraint, SHVC does not have the problems inherent to single loop decoding.

Lastly, a full list of scalability features supported by SHVC is summarized in Table 2 in comparison with SVC. The type of scalability feature(s) between two layers dictates the type of inter-layer processing applied (cf. Figure 1). Both scalable standards support the conventional set of scalability features including temporal scalability (lower frame rate to higher frame rate), spatial scalability (lower spatial resolution to higher spatial resolution) and SNR scalability (lower quality to higher quality). Temporal scalability is already fully supported by the single-layer HEVC standard; SHVC simply inherits this feature. For spatial scalability, the resampling process in SHVC provides more enhanced functionality compared to SVC, to be discussed next. As shown in Table 2, SHVC also supports three new scalability features not supported by SVC: 1) the hybrid codec scalability discussed above, where the base layer can be coded using non-HEVC codec; 2) bit depth scalability, where the base layer is of lower bit-depth (e.g. 8-bit) and the EL is of higher bit-depth (e.g. 10-bit); 3) color gamut scalability, where the BL has narrower color gamut (e.g. ITU-R recommendation BT.709 [15]) and the EL has wider color gamut (e.g. ITU-R recommendation BT.2020 [16]). Individual scalability features in Table 2 can be combined. In particular, the combination of spatial, bit depth, and color gamut scalability can be used to fully enable the migration from HDTV to UHDTV.

Table 2: Comparison of scalability features supported by SVC and SHVC

Scalability features	Scalable standard		Examples
	SVC	SHVC	
Temporal	X	X (in HEVC)	30fps to 60fps
Spatial	X	X	1080p to 4Kx2K
SNR (quality)	X	X	33 dB to 36 dB
Hybrid codec		X	AVC-coded BL
Bit depth		X	8-bit to 10-bit
Color gamut		X	BT.709 to BT.2020

In terms of spatial scalability, SHVC provides more flexibility, mainly in two aspects: 1) because of SHVC's relatively simple architecture design, arbitrary scaling ratio is included in SHVC's Scalable Main and Scalable Main 10 profiles. In comparison, arbitrary spatial ratio, also referred to as enhanced spatial scalability in SVC, is supported only in the Scalable High profile of SVC, and Scalable Baseline profile only allows 2x and 1.5x ratios; 2) SHVC supports flexible resampling phase adjustment in resampling. This allows the resampling filter phases to be selected to match those of the down-sampling filters used by the encoder. By default, the resampling process of SHVC assumes that the top left sample location of the pictures of the two spatial layers have zero phase shift. However, when generating the lower resolution layer, SHVC gives the encoder the freedom to choose non-default down-sampling filters; for example, the encoder can choose the down-sampling filters used by SVC with 0.5-sample phase shift [6]. By adjusting the filter phases during resampling at the decoder to match those used by the encoder during down-sampling, high coding efficiency can be maintained for all encoder designs. Additionally, the resampling process in SHVC allows the output sample bit depth to be different from (and larger than) the input sample bit depth. This provides a natural support for bit depth scalability.

Note that when Nx is used to refer to the spatial ratio for spatial scalable coding, the spatial resolution ratio is N is both the horizontal and the vertical dimensions, for example, for a two-layer spatial scalable coding, if the base layer is of 640x360 and the enhancement layer is of 1280x720, it is a 2x spatial scalable coding.

To summarize, Table 3 lists the comparison between SVC and SHVC made in this clause.

Table 3: Summary of comparison between SVC and SHVC

	SVC	SHVC
Inter-layer prediction signalling	Add flag in macroblock	Reuse ref_idx in Prediction Unit
Decoding type	Single-loop	Multi-loop
Spatial scalability ratio	Limited to 2x and 1.5x in Scalable Baseline	Arbitrary ratio
Spatial resampling phase	Fixed phase position	Arbitrary phase adjustment
Backward compatibility	AVC-coded BL only	HEVC or non-HEVC coded BL
BL decoder design	Needs new API	No change to the BL decoder
EL decoder design	Cannot directly reuse AVC decoder	Can repurpose existing HEVC decoder

5.4 SHVC decoder and encoder complexity analyses

5.4.1 Introduction

This clause provides a complexity comparison of SHVC and HEVC simulcast, including the encoder/decoder architecture and coding tools.

The following are concluded from the analysis:

- SHVC codec architecture is high-level syntax changes only which is easy to implement by re-using the existing HEVC codec designs.
- HEVC simulcast decoder and SHVC base layer decoder complexity are identical when the output layer is base layer.
- SHVC decoder complexity is around 1.25x as that of HEVC single layer decoder for 2x spatial scalability when the output layer is the enhancement layer.

5.4.2 HEVC simulcast encoder and decoder

HEVC is a block-based hybrid coding architecture, combining inter and intra prediction and transform coding with high-efficiency entropy coding. HEVC employs a quad-tree coding block partitioning structure based on a coding tree unit (CTU) instead of a macroblock, which enables a flexible use of large and small coding, prediction, and transform blocks. HEVC also allows for improved intra prediction and coding, adaptive motion parameter prediction and coding, a new loop filter and an enhanced version of context-adaptive binary arithmetic coding (CABAC) entropy coding. New high level structures, such as tiles or wavefront parallel processing (WPP), have also been designed to aid parallel processing. Figure 2 shows a general block diagram of HEVC encoders.

Figure 3 illustrates a block diagram of HEVC decoders that corresponds to the encoder diagram in Figure 2. The video bitstream is parsed and entropy decoded first, the coding mode and associated prediction information are passed to either intra prediction or motion compensated prediction to form the prediction block. The residual transform coefficients are then inverse quantized and inverse transformed to reconstruct the residual block. The prediction block and the reconstructed residual block are then added together to form the reconstructed block. The reconstructed block may further go through in-loop filtering before being stored into the reference picture buffer and used to predict future video blocks.

The total amount of memory required for HEVC decoding can be expected to be similar to that for H.264/AVC decoding, and most of the memory is required for the decoded picture buffer that holds multiple pictures. HEVC may require more cache memory due to larger block sizes that it supports. The complexity of some key modules such as transforms, intra prediction and motion compensation is likely higher in HEVC than in H.264/AVC, and the complexity was reduced in other modules such as entropy coding and deblocking. The implementation cost of an HEVC decoder is not expected to be much higher than that of an H.264/AVC decoder even with additional in-loop filter such as sample adaptive offset (SAO). From an encoder prospective, due to the flexibility of quad-tree structures and many more intra prediction modes, an encoder fully exploiting the capabilities of HEVC can be expected to be several times more complex than an H.264/AVC encoder. It was reported that real-time software decoding and display of a variety of 1080p sequences was feasible on a smartphone featuring an ARMv7 processor clocked at 1.3GHz with 30fps at 2Mbps back in 2012.

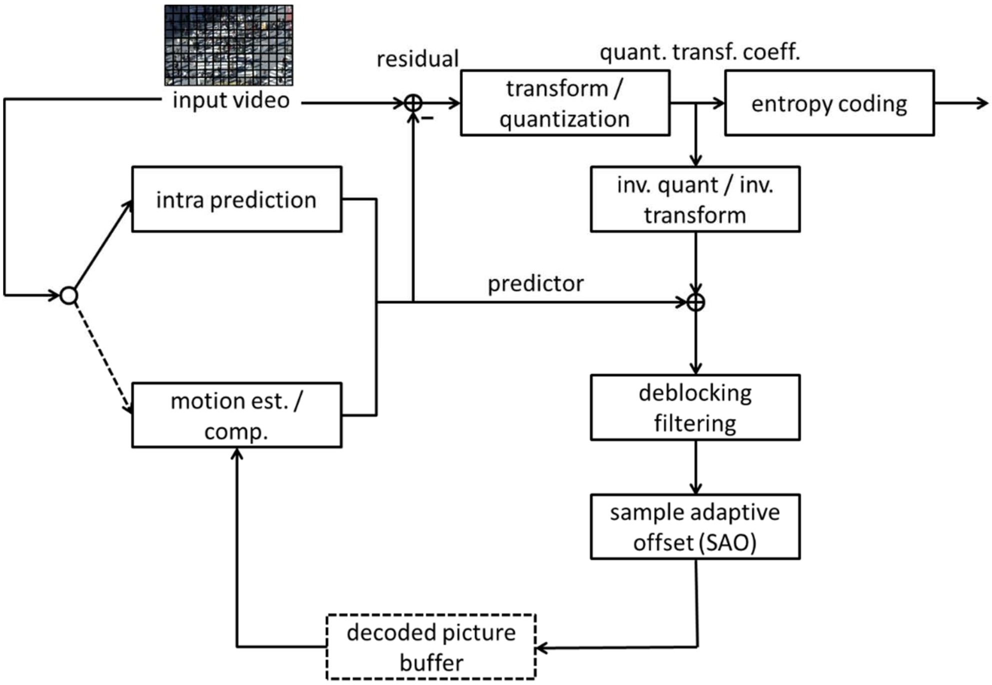

Figure 2: Generic diagram of HEVC encoders

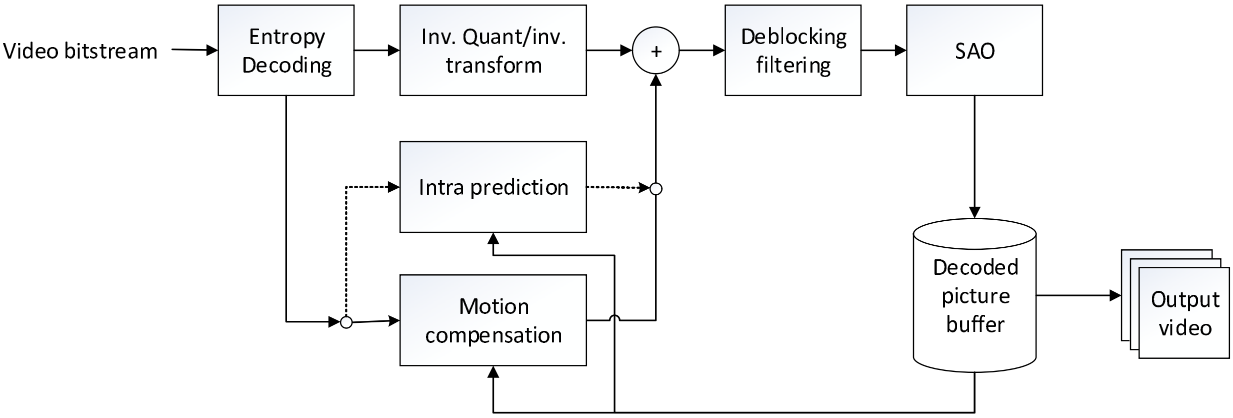

Figure 3: Generic diagram of HEVC decoders

For HEVC simulcast, each single layer video sequence is encoded by an HEVC encoder and decoded by an HEVC decode without any dependency. Figure 4 illustrates an HEVC simulcast example with two spatial resolutions, the high resolution video content is encoded by a single HEVC encoder and the generated video bitstream is streamed to the devices with high capabilities or sufficient network bandwidth. The down-sampled low resolution video content is encoded by another HEVC encoder, and the generated video bitstream is streamed to the devices with low capabilities or low network bandwidth. The decoded picture buffer (DPB) stores multiple reconstructed pictures as reference for the motion estimation and compensation.

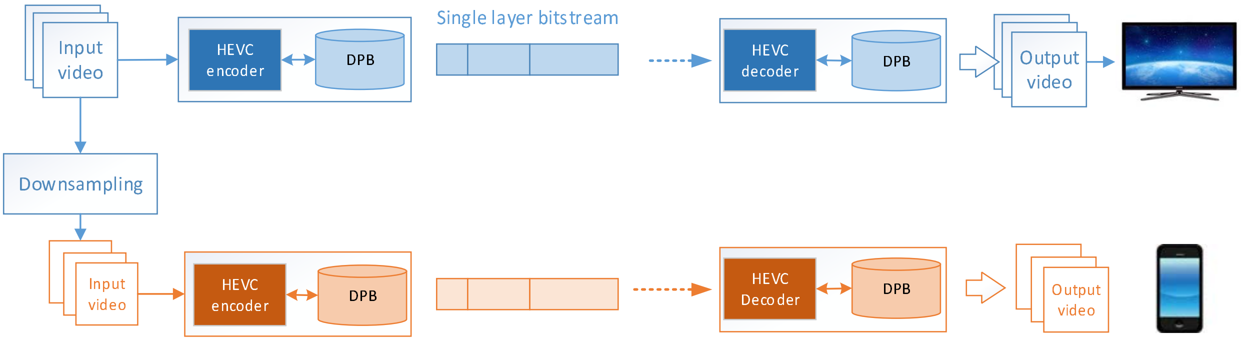

Figure 4: HEVC simulcast example

5.4.3 SHVC encoder and decoder

SHVC is the scalability extension to HEVC that enables spatial scalability, SNR scalability, bit depth scalability and color gamut scalability. The SHVC design uses a multi-loop coding framework, such that in order to decode an enhancement layer, its reference layers have to first be fully decoded to make them available as prediction references. The coding tools of SHVC are limited to changes at the slice level and above for ease of implementation, especially the possibility to re-use existing HEVC implementations.

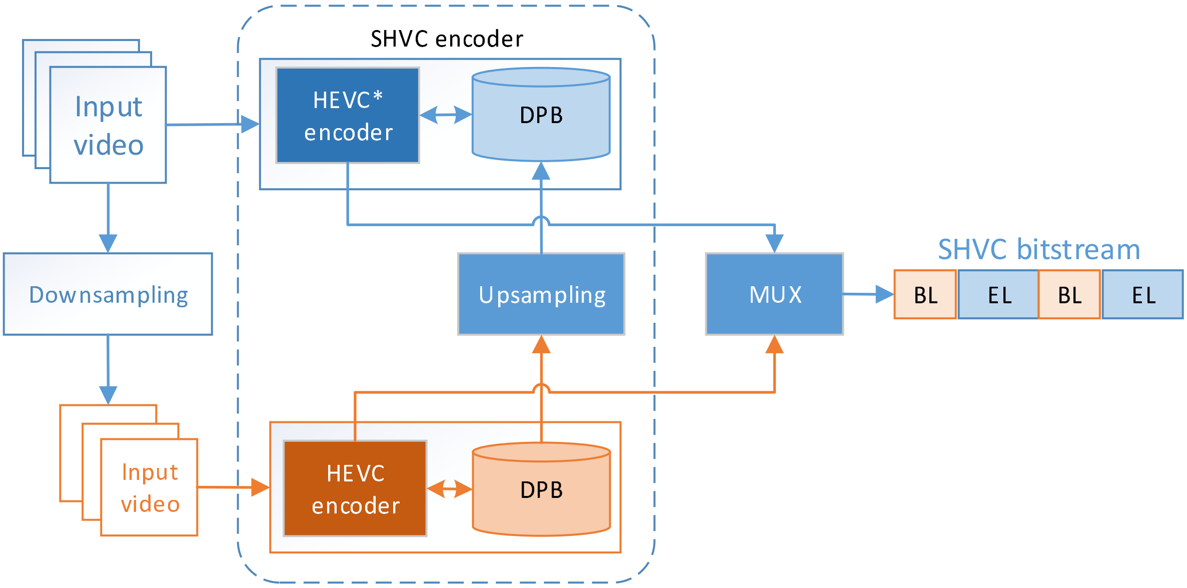

Figure 5: 2 layer spatial scalability SHVC encoder diagram

Figure 5 shows a two-layer spatial scalability SHVC encoder architecture consisting of a base layer (BL) encoder and an enhancement layer (EL) encoder. The base layer encoder is identical to a single HEVC encoder. The up-sampling module is used to map reconstructed sample values from the base layer to the higher-resolution sampling grid of the enhancement layer. This allows the use of the reconstructed base layer sample values for enhancement layer prediction.

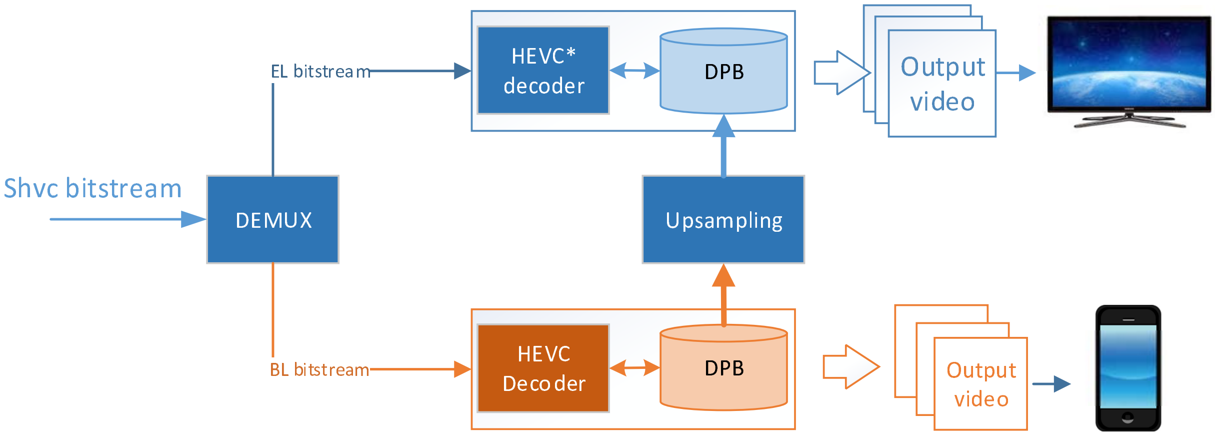

Figure 6: 2 layer spatial scalability SHVC decoder diagram

Figure 6 shows a two layer spatial scalability SHVC decoder architecture consisting of a base layer (BL) decoder and an enhancement layer (EL) decoder. The base layer decoder is identical to a single HEVC decoder. The reconstructed picture from the base layer is then up-sampled and put into the EL DPB as a long term reference picture and used along with the EL temporal reference pictures for enhancement layer decoding.

NOTE: In the SHVC spec, the up-sampled reference layer picture is not specified as being stored in the DPB. Decoders that would store an entire up-sampled reference layer picture can store it in a memory buffer that is conceptually not considered as part of the DPB.

Furthermore, the base layer codec in Figure 2 and Figure 3 can be operated as a black box because SHVC EL codec only needs the reconstructed BL pictures. This makes it easier to support different codecs for the base layer, which is also referred to as hybrid codec scalability. Such feature allows previous generation codecs to be used in the BL for backward compatibility, and the more efficient HEVC codec is used in the EL to improve coding performance.

5.4.4 Upsampling filter

In SHVC, the upsampling filter is defined as an 8-tap filter for luma resampling, and a 4-tap filter for chroma resampling. The basic design enables the use of arbitrary upsampling ratios, in which filters for all 16 phase positions would be necessary. Table 4 shows the filter coefficients of 8-tap luma resampling filters and Table 5 shows the filter coefficients of the 4-tap chroma resampling filters. In HEVC, the luma and chroma sample interpolation for fractional sample positions use the similar type of filters with the same number of taps.

Table 4: 16-phase luma resampling filter

phase p	**interpolation filter coefficients**							
	f_L[p, 0]	**f_L[p, 1]**	**f_L[p, 2]**	**f_L[p, 3]**	**f_L[p, 4]**	**f_L[p, 5]**	**f_L[p, 6]**	**f_L[p, 7]**
0	0	0	0	64	0	0	0	0
1	0	1	−3	63	4	−2	1	0
2	−1	2	−5	62	8	−3	1	0
3	−1	3	−8	60	13	−4	1	0
4	−1	4	−10	58	17	−5	1	0
5	−1	4	−11	52	26	−8	3	−1
6	−1	3	−9	47	31	−10	4	−1
7	−1	4	−11	45	34	−10	4	−1
8	−1	4	−11	40	40	−11	4	−1
9	−1	4	−10	34	45	−11	4	−1
10	−1	4	−10	31	47	−9	3	−1
11	−1	3	−8	26	52	−11	4	−1
12	0	1	−5	17	58	−10	4	−1
13	0	1	−4	13	60	−8	3	−1
14	0	1	−3	8	62	−5	2	−1
15	0	1	−2	4	63	−3	1	0

Table 5: 16-phase chroma resampling filter

phase p	interpolation filter coefficients			
	f_C[p, 0]	f_C[p, 1]	f_C[p, 2]	f_C[p, 3]
0	0	64	0	0
1	−2	62	4	0
2	−2	58	10	−2
3	−4	56	14	−2
4	−4	54	16	−2
5	−6	52	20	−2
6	−6	46	28	−4
7	−4	42	30	−4
8	−4	36	36	−4
9	−4	30	42	−4
10	−4	28	46	−6
11	−2	20	52	−6
12	−2	16	54	−4
13	−2	14	56	−4
14	−2	10	58	−2
15	0	4	62	−2

5.4.5 Inter-layer texture prediction

The upsampling process above enables the projection of reference layer reconstructed samples to the enhancement layer resolution. SHVC requires an EL decoder to insert the up-sampled reference layer picture into the EL DPB as inter-layer reference picture (ILP) for texture prediction. The ILP is signaled in reference picture list (RPL) for reference in the same manner as usually in inter prediction.

The process for constructing the RPL at the decoder is relatively straightforward. First, an initial RPL is constructed in the same way as in HEVC version 1, the short-term reference pictures and long-term reference pictures identified in the bitstream are added to the RPL, and the upsampled base layer picture is inserted after the short-term reference pictures that have smaller values of picture order counts than the current picture and is marked as a long term reference picture.

Again, this is consistent with the HEVC except that the initial RPL contain the upsampled base layer picture and any additional reference layer pictures when present.

Such reference index signaling based approach improves the coding flexibility and efficiency since the enhancement layer encoder can signal a prediction from either temporal reference picture, or base layer upsampled reference picture, or both with weighted prediction. To limit memory bandwidth and complexity, SHVC specifies a bitstream restriction that the motion vector will be zero when referencing the upsampled reference layer samples. This simplifies the codec design, especially for implementations that might perform the up-sampling on the fly as part of prediction process, rather than upsampling whole reference pictures in advance.

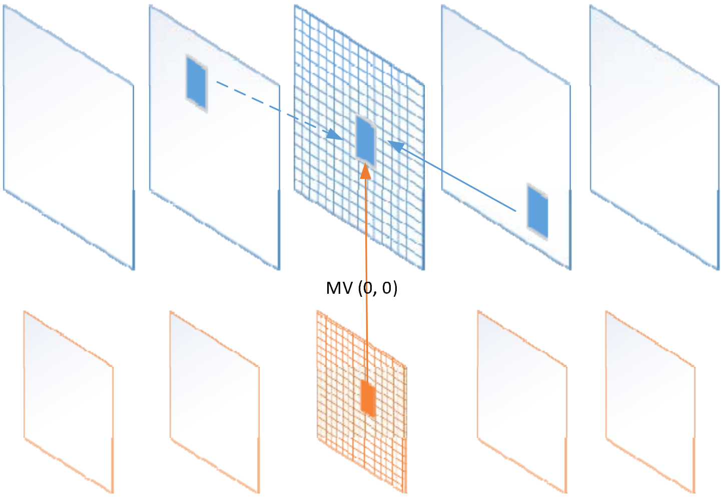

Figure 7: SHVC inter-layer texture prediction

5.4.6 Inter-layer motion prediction

SHVC uses the reference layer motion information when coding EL motion vectors (MVs) by making use of the existing temporal motion vector prediction (TMVP) process of HEVC version 1.

In HEVC, TMVP is used to predict motion information for a current block from a co-located block in the reference picture (Figure 8). The motion information consists of inter prediction mode, reference indices, luma MVs and reference picture order counts (POCs) of the co-located block. The motion information is compressed and stored on a 16x16 luma lock basis which reduces the worst-case memory size and bandwidth requirements for storing the reference layer motion information.

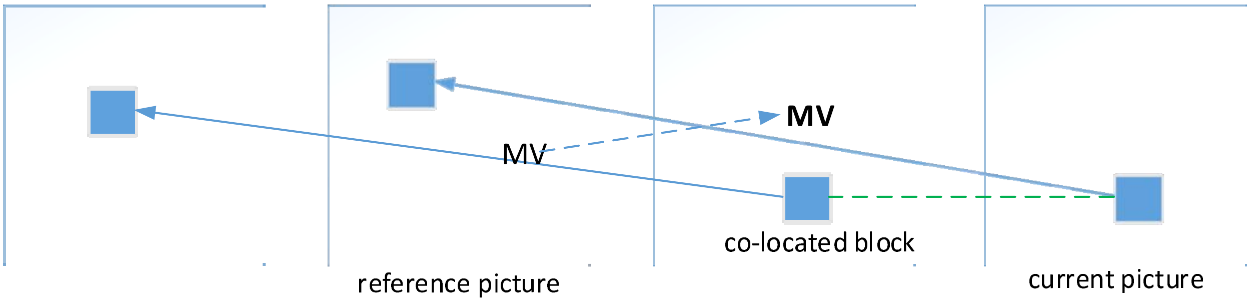

Figure 8: HEVC TMVP

SHVC inter-layer motion prediction maps BL's motion information to EL's resolution and the mapped motion information is also stored in units of 16x16 luma samples. Once the EL co-located position is determined and the corresponding motion information from the co-located BL block is available, a scaling operation is applied to those BL motion vectors to account for the upsampling ratio. The scaled BL motion information is then used as reference to predict the EL's motion information as shown in Figure 9. The inter-layer motion prediction provides a means to leverage a significant amount of reference layer information without changing the block level design of an HEVC codec. When the base layer is using a previous generation codec such as H264/AVC or MPEG-2, the inter-layer motion prediction is disabled since the BL motion information may not be available.

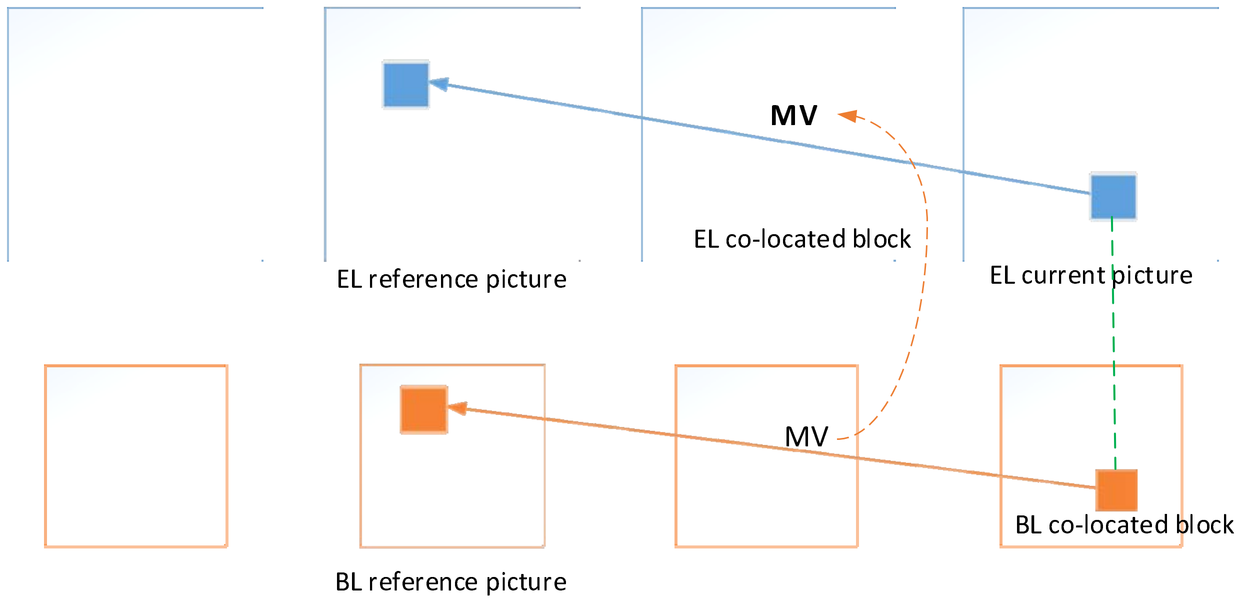

Figure 9: SHVC inter-layer motion prediction

5.4.7 Conclusion

SHVC adopts a multi-loop design framework with only high-level syntax changes relative to its base codec. Such design would ease SHVC implementation and allow re-using existing HEVC implementations. The SHVC coding modules such as up-sampling filter, inter-layer texture prediction and inter-layer motion prediction leverage a significant amount of reference layer information without changing the block level design of an HEVC codec.

The computational complexity and memory access would be similar between HEVC simulcast and SHVC from the encoder perspective. Multiple HEVC simulcast bitstreams or multiple layer SHVC bitstreams would be generated to support multiparty video conferencing, and the complexity of each layer's HEVC encoder or SHVC encoder would be similar based on the previous technical complexity analysis.

From the decoder perspective, the decoding complexity of the base layer would be exactly the same between HEVC simulcast decoder (of the lowest resolution) and SHVC base layer decoder. When decoding the high resolution layer, SHVC base layer decoder and enhancement layer decoder are both needed in order to output enhancement layer pictures, the extra computational complexity of SHVC decoder comparing to HEVC single layer decoder is determined by the base layer decoder complexity. For 2x spatial scalability, SHVC decoder complexity is estimated to be around 1.25x comparing to HEVC single layer decoder when the enhancement layer is the output layer.

6 Use cases

6.1 Multi-stream Multiparty Video Conferencing (MMVC)

6.1.1 The heterogeneous-device MMVC use case

The first MMVC use case considers video conferencing with multiple participating endpoints, e.g. special VC terminals, laptops, tablets and smartphones, with different decoding and render capabilities. This use case is thus referred to as the heterogeneous-device MMVC use case. The Multimedia Resource Function Processor (MRFP) as the conference focus makes connections between multiple video conferencing endpoints and receives video streams from each endpoint, and forwards a set of appropriate video streams to each endpoint.

The following assumptions are made for this use case:

1) It is assumed that there are two terminal classes, the high-end terminal devices that are capable of encoding/sending and decoding/receiving a high video resolution, e.g. 1080p@30fps, and the low-end terminal

devices that are capable of encoding/sending and decoding/receiving a medium video resolution, e.g. 720p@30fps.

2) Each UE other than the current active speaker displays a full video of the current active speaker and a thumbnail video of other participants. The video resolution of the full video is the lower of the current active speaker's sending capability and this UE's receiving capability. The video resolution of each thumbnail video is of a low video solution, e.g. 240p@15fps.

3) The current active speaker UE displays a full video of the previous active speaker and a thumbnail video of other participants. The video resolution of the full video is the lower of the previous active speaker's sending capability and the active speaker UE's receiving capability.

4) It is assumed that the thumbnail video of each participant is encoded as a single video bitstream (i.e. not as the base layer of a scalable video bitstream).

Figure 10 shows an example of this use case. Four endpoints (A, B, C and D) are the participants in one video conference, where participants A and D are high-end devices, B and C and are low-end devices, A is the current active speaker, and B is the previous active speaker.

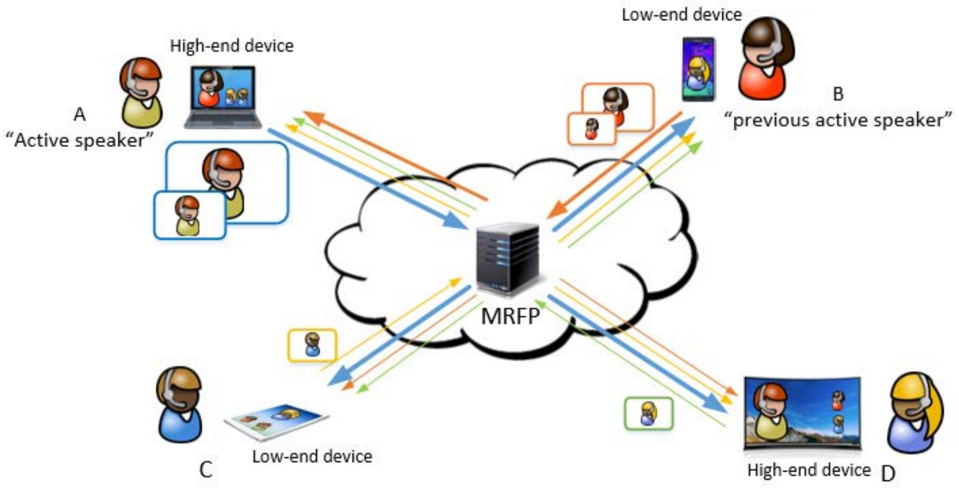

Figure 10: The heterogeneous-device MMVC use case

The signal flows are as follows:

Participant A (the current active speaker) sends a high video resolution (as main video for participant D) and a medium video resolution (as main video for participants B and C) to the MRFP, and receives (via the MRFP) the medium resolution full video from participant B and the low resolution thumbnail videos from participants C and D.

Participant B (the previous active speaker) sends a medium video resolution (as main video for participant A) and a low resolution thumbnail video (for participants C and D) to the MRFP, and receives (via the MRFP) the medium resolution full video from participant A and the low resolution thumbnail videos from participants C and D. The medium resolution video is encoded as a single video bitstream, and the thumbnail video is also encoded as a single video bitstream.

Participant C sends a low resolution thumbnail video (for participants A, B and D) to the MRFP, and receives (via the MRFP) the medium resolution full video from participant A and the low resolution thumbnail videos from participants B and D. The thumbnail video is encoded and transmitted as a single HEVC bitstream.

Participant D sends a low resolution thumbnail video (for participants A, B and C) to the MRFP, and receives (via the MRFP) the high resolution full video from participant A and the low resolution thumbnail videos from participants C and D. The thumbnail video is encoded and transmitted as a single HEVC bitstream.

The metrics for comparison of the potential solutions for this use case are:

1) Total uplink bitrate for each participant UE

2) Total downlink bitrate for each participant UE

3) Quality for participant UE

4) Decoding complexity for each participant UE

5) Encoding complexity for each participant UE

6.1.2 The heterogeneous-bandwidth MMVC use case

The second MMVC use case considers video conferencing with multiple participating endpoints with different access network bandwidths, e.g. as shown in Figure 11. This use case is thus referred to as the heterogeneous-bandwidth MMVC use case. Similarly as in the heterogeneous-device MMVC use case, the MRFP makes connections between multiple video conferencing endpoints and receives video streams from each endpoint, and forwards a set of appropriate video streams to each endpoint.

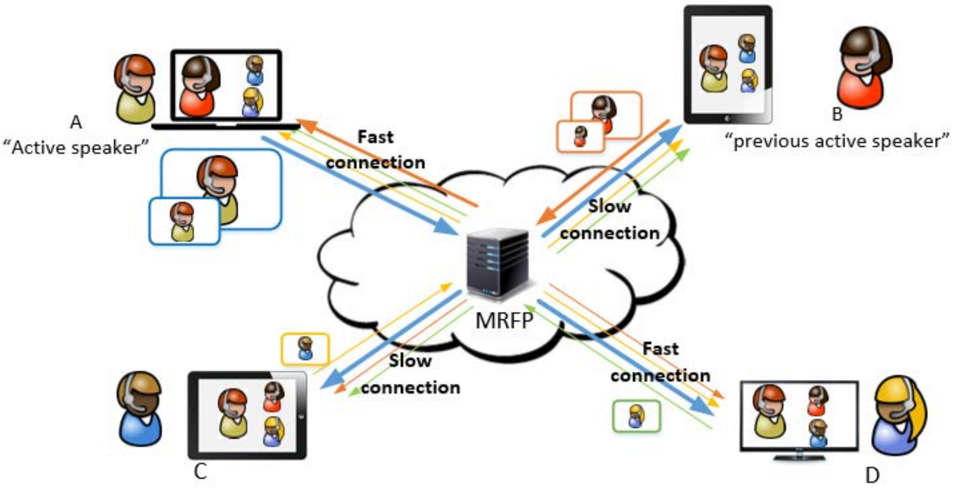

Figure 11: The heterogeneous-bandwidth MMVC use case

The following assumption is made for this use case:

```
1) It is assumed that there are two access network classes, the fast connection
that is capable of sending and receiving a high video resolution, e.g.
1080p@30fps, and the slow connection that is capable of sending and receiving a
medium video resolution, e.g. 720p@30fps.
2) It is also assumed that the sending and decoding capabilities of each
terminal's access network are within the terminal's encoding and decoding
capabilities.
```

With these two assumptions, then the 2nd, 3rd and 4th assumptions of the heterogeneous-device MMVC use case apply herein. With the same topology in Figure 11 as in Figure 10, the same signal flows as well as the metrics as described for the heterogeneous-device MMVC use case also apply herein.

6.1.3 Solutions for the MMVC use cases

6.1.3.0 General

The solutions from video coding point of view apply to both MMVC use cases in the same manner. Therefore, in the discussion of the solutions, no difference is made to which use case the solutions apply.

6.1.3.1 HEVC simulcast

HEVC simulcast can be applied as a video coding solution to the MMVC use cases. In this solution, each full or thumbnail video as described in the signal flows is an independently coded HEVC single-layer bitstream.

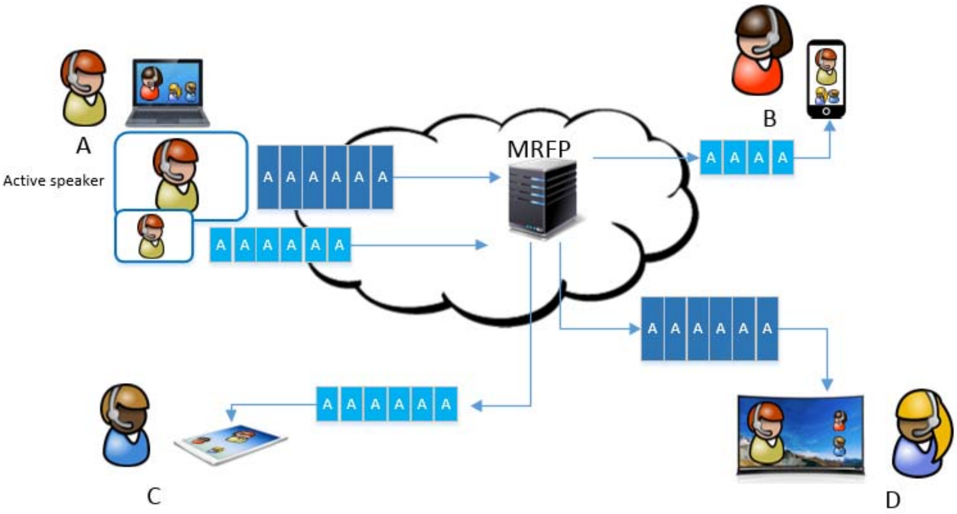

Figure 12: Use of HEVC simulcast in the MMVC use cases

Figure 12 shows the two independent HEVC single-layer bitstreams that are sent from participant A and received (through the MRFP) by other participants. Both bitstreams are sent to the MRFP, which forwards the high resolution bitstream to participant D and the medium resolution bitstream to participants B and C.

6.1.3.2 SHVC

Another video coding solution for the MMVC use cases is to apply SHVC. Multiple spatial resolutions can be encoded into one SHVC bitstream. In this solution, for any participant, if both the high and medium resolutions need to be sent simultaneously (e.g. participant A), the participant UE encodes the two solutions into two layers of one SHVC bitstream and sends the bitstreams to the MRFP; otherwise each full or thumbnail video as described in the signal flows is an independently coded HEVC single-layer bitstream.

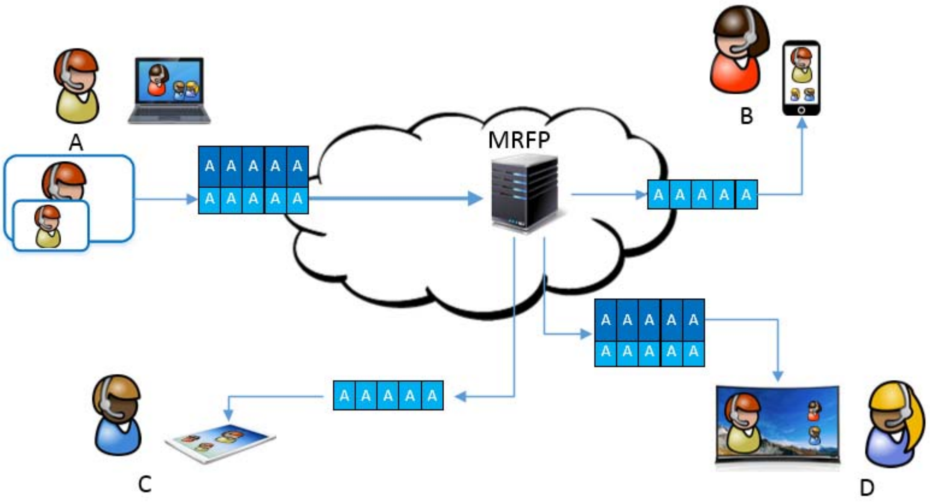

Figure 13: Use of SHVC in the MMVC use cases

Figure 13 shows the two layers of the SHVC bitstream that are sent from participant A and received (through the MRFP) by other participants. Both layers of the bitstream are sent to the MRFP, which forwards the both layers to participant D but only the base layer to participants B and C.

6.1.4 Comparison of the solutions

6.1.4.1 Uplink bandwidth

The uplink connection between UE A and MRFP includes one 1080p@30fps resolution video and one 720p@30fps resolution video.

The uplink connection between UE B and MRFP includes one 720p@30fps resolution video and one 240p@15fps resolution video. Due to the resolution difference, UE B may deploy simulcast to transmit 720p@30fps and 240p@15fps video streams.

The uplink connection between UE C and MRFP includes one 240p@15fps video.

The uplink connection between UE D and MRFP includes one 240p@15fps video.

Table 6 and Table 7 show the uplink luma BD-rate comparison between HEVC simulcast and SHVC with different deltaQP values, based on the test conditions described in clause 7.

Table 6: Uplink BD-rate comparison (SHVC vs. Simulcast, BL720p/EL1080p, DeltaQP=0)

	A	B	C	D
Kimono	-21.5%	0%	0%	0%
ParkScene	-13.1%	0%	0%	0%
Cactus	-18.8%	0%	0%	0%
BasketballDrive	-24.2%	0%	0%	0%
BQTerrace	-8.5%	0%	0%	0%
Average saving	-17.22%	0%	0%	0%

Table 7: Uplink BD-rate comparison (SHVC vs. Simulcast, BL720p/EL1080p, DeltaQP=2)

	A	B	C	D
Kimono	-35.5%	0%	0%	0%
ParkScene	-22.7%	0%	0%	0%
Cactus	-29.7%	0%	0%	0%
BasketballDrive	-33.6%	0%	0%	0%
BQTerrace	-15.2%	0%	0%	0%
Average saving	-27.34%	0%	0%	0%

6.1.4.2 Downlink bandwidth

The downlink connection between UE A and MRFP includes one 720P@30fps single layer stream and two 240p@15fps single layer stream.

The downlink connection between UE B and MRFP includes one 720P@30fps single layer stream and two 240p@15fps single layer streams.

The downlink connection between UE C and MRFP includes one 720P@30fps single layer stream and two 240p@15fps single layer streams.

For HEVC simulcast, the downlink connection between UE D and MRFP includes one 1080p@30fps single layer stream and two 240p@15fps single layer streams.

For SHVC, the downlink connection between UE D and MRFP includes one 1.5x spatial scalability SHVC stream (base layer 720p@30fps and enhancement layer 1080p@30fps) and two 240p@15fps single layer streams.

Table 8 and Table 9 show the downlink luma BD-rate comparison between HEVC simulcast and SHVC with different deltaQP values, based on the test conditions described in clause 7 for MMVC for the IRAP-alignment scenarios.

Table 8: Downlink BD-rate comparison (SHVC vs. Simulcast, deltaQP = 0)

	A	B	C	D
Kimono	0%	0%	0%	25.1%
ParkScene	0%	0%	0%	31.4%
Cactus	0%	0%	0%	25.2%
BasketballDrive	0%	0%	0%	16.5%
BQTerrace	0%	0%	0%	21.1%
Average cost	0%	0%	0%	23.86%

Table 9: Downlink BD-rate comparison (SHVC vs. Simulcast, deltaQP = 2)

	A	B	C	D
Kimono	0%	0%	0%	14.6%
ParkScene	0%	0%	0%	33.2%
Cactus	0%	0%	0%	22.7%
BasketballDrive	0%	0%	0%	13.8%
BQTerrace	0%	0%	0%	33.4%
Average cost	0%	0%	0%	23.54%

6.1.4.3 Decoding complexity

The decoding complexity of UE A, B and C is the same for HEVC simulcast and SHVC since each UE decodes the same amount of single layer streams (one 720p@30fps and two 240p@15fps). For simulcast, UE D decodes one 1080p@30fps video streams and two 240p@15fps single layer video streams; while for SHVC, UE D decodes one 1.5x spatial scalability SHVC stream and two 240p@15fps single layer video streams. The SHVC decoding complexity of UE D increases around 40% comparing to HEVC simulcast because the base layer video (720p@30fps) has to be decoded in order to output enhancement layer video (1080p@30fps).

6.1.4.4 Encoding complexity

The encoding complexity of UE B, C and D is the same for HEVC simulcast and SHVC since each UE encodes the same amount of single layer streams. For HEVC simulcast, UE A encodes one 1080p@30fps video streams and one 720p@30fps single layer video streams; while for SHVC, UE A encodes one 1.5x spatial scalability SHVC stream. Compared to simulcast, the complexity of SHVC encoding at UE A is typically less than that of simulcast encoding. This is because SHVC places zero motion constraint on the inter layer reference picture. When the inter layer reference picture provides sufficiently good prediction signal (without the need for motion estimation), early termination is typically applied at the encoder, and the need for motion estimation of the temporal reference pictures is avoided, leading to lower encoding complexity.

6.2 MBMS

6.2.1 The differentiated-service MBMS use case

This use case considers MBMS with subscription based differentiated video services. This use case is thus referred to as the differentiated-service MBMS use case. It should be reasonable to assume that two different classes of video services may be provided (as providing more classes would be heavy for any broadcast system), e.g. the normal video service of 720p@30fps and the premium video service of 1080p@30fps. UEs may subscribe to either of the two services because of its decoding and rendering capabilities, network access conditions, power saving strategies, price, and/or other considerations. UEs receiving the normal service receives and renders the lower quality video with lower resolution, and UEs receiving the premium service receives and renders the higher quality video with higher resolution.

Due to the use of the broadcast mode in eMBMS, all bits required for both services are assumed to be transmitted on all the network paths, from the content provider to the BM-SC, from the BM-SC to MBMS-GW, from MBMS-GW to eNodeB, as well as the air interface between eNodeB and UEs, as shown in Figure 14.

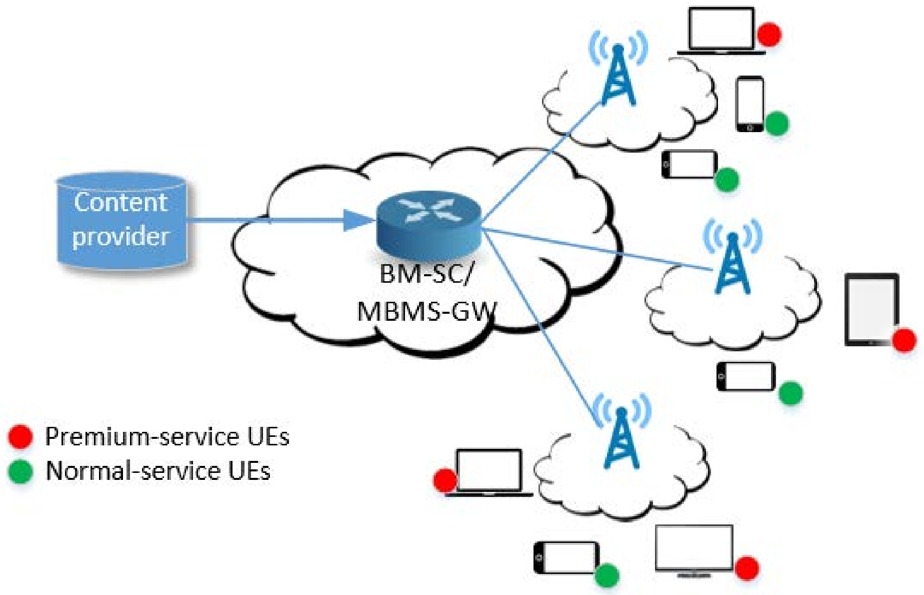

Figure 14: An MBMS use case with premium and normal video services

The metrics for comparison of the potential solutions for this use case are:

1) Bandwidth used for transmission of the video data (in the network links from the content provider to the BM-SC, and all the way to the UEs)

2) Quality for normal-service UEs and premium-service UEs

3) Decoding complexity for normal-service UEs and premium-service UEs

4) Encoding complexity

6.2.2 Solutions for the MBMS use case

6.2.2.1 HEVC simulcast

HEVC simulcast can be applied as a video coding solution to the MMVC use cases. In this solution, each full or thumbnail video as described in the signal flows is an independently coded HEVC single-layer bitstream.

One solution from video coding point of view for the MBMS use case with premium and normal video services is to use HEVC simulcast, where two independently encoded HEVC bitstreams representing the same video content, but with different spatial resolutions, are transmitted (from the content provider to the BM-SC, and all the way to the UEs), as shown in Figure 15. The two HEVC bitstreams are associated with two different FLUTE sessions of the same MBMS User Service. Each premium-service UE receives, decodes, and renders the video bitstream with the higher resolution only, while each normal-service UE receives, decodes, and renders the video bitstream with the lower resolution only.

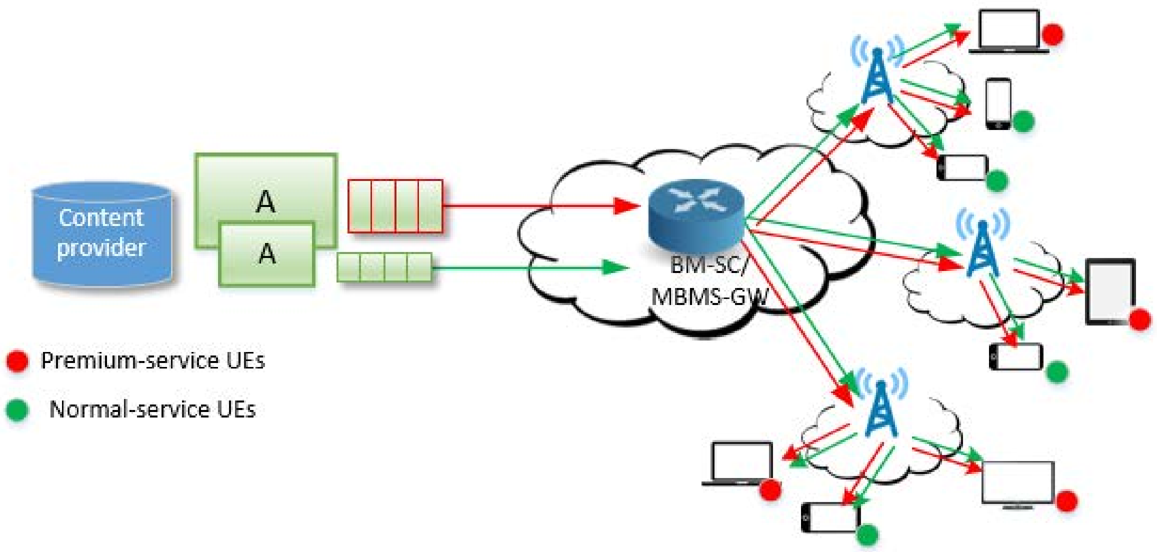

Figure 15: Use of HEVC simulcast in the differentiated-service MBMS use case

6.2.2.2 SHVC

Another solution from video coding point of view for the MBMS use case with premium and normal video services is to use SHVC, where one encoded SHVC bitstream with two layers of different spatial resolutions, is transmitted (from the content provider to the BM-SC, and all the way to the UEs), as shown in Figure 16. The sub-bitstreams of the two layers of the SHVC bitstream are associated with two different FLUTE sessions of the same MBMS User Service. Each premium-service UE receives and decodes both layers and renders the higher layer, while each normal-service UE receives, decodes, and renders the base layer only.

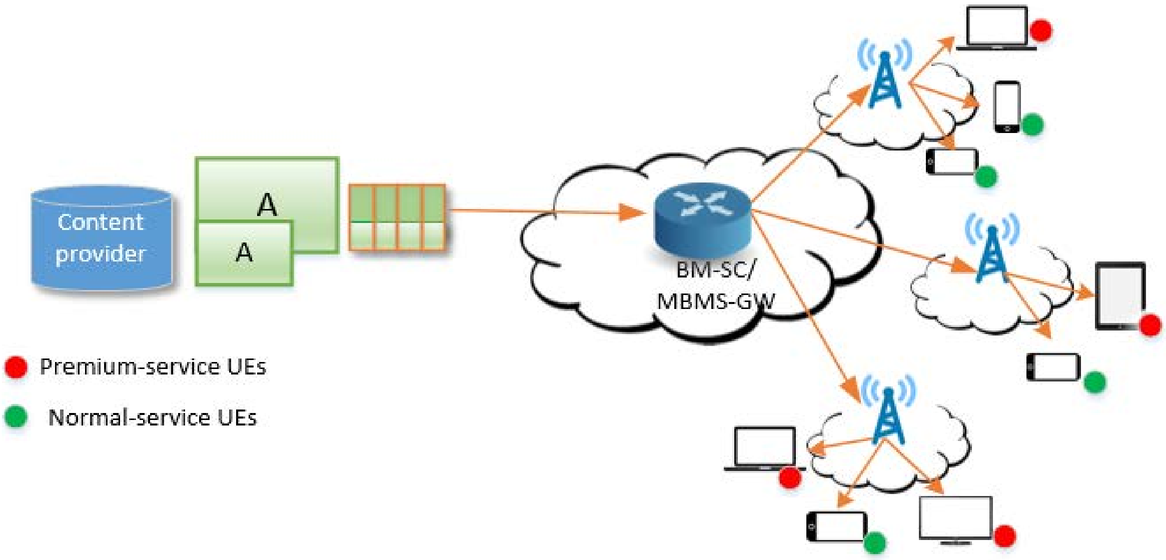

Figure 16: Use of SHVC in the differentiated-service MBMS use case

6.2.3 Comparison of the solutions

6.2.3.1 Transmission bandwidth

For solution with HEVC simulcast, the bandwidth for transmission from the content provider to the BM-SC, and all the way to the UEs is the bandwidth required for transmitting one HEVC coded 1080p@30fps stream and one HEVC coded 720p@30fps stream. For solution with SHVC, the bandwidth for transmission from content provider to BM-SC is the bandwidth required for transmitting one SHVC stream containing two layers (i.e., 720p@30fps for base layer and 1080p@30fps).

The simulation results is reported in TR 26.948 [18] for MBMS are relevant. For the cross-layer RAP non-aligned case, the average bandwidth decrease (i.e. based on BD-rate decrease) for SHVC comparing to HEVC simulcast was around 32.9% for 1.5x spatial scalability and 20% for 2x spatial scalability, and the max gain was up to 40.6%. For the cross-layer RAP aligned case, the average bandwidth decrease was around 31.9% for 1.5x spatial scalability and 18.7% for 2x spatial scalability, and the max gain was up to 40.5%. The results are tabulated in the following tables:

Table 10: Test results for IRAP aligned Class B (BL 720p – EL 1080p)

Test Sequences	QP delta BL & EL	BD-Rate Comparison		
		SHVC Vs. Simulcast		
		Y	U	V
Kimono	0	-28.4%	-20.9%	-18.7%
ParkScene	0	-19.5%	-15.3%	-15.1%
Cactus	0	-23.4%	-19.2%	-12.5%
BasketballDrive	0	-26.5%	-15.5%	-15.9%
BQTerrace	0	-14.9%	4.9%	10.4%
Average		-22.5%	-13.2%	-10.4%
Kimono	2	-40.5%	-34.5%	-33.3%
ParkScene	2	-28.9%	-24.6%	-25.2%
Cactus	2	-32.7%	-30.5%	-27.5%
BasketballDrive	2	-36.1%	-30.3%	-29.8%
BQTerrace	2	-21.5%	-11.2%	-10.5%
Average		-31.9%	-26.2%	-25.3%

Table 11: Test results for IRAP aligned Class B (BL 540p – EL 1080p)

Test Sequences	QP Delta BL & EL	BD-Rate Comparison		
		SHVC Vs. Simulcast		
		Y	U	V
Kimono	0	-19.2%	-11.9%	-9.5%
ParkScene	0	-10.2%	-8.5%	-8.7%
Cactus	0	-13.7%	-9.7%	-3.9%
BasketballDrive	0	-16.6%	-4.9%	-6.1%
BQTerrace	0	-7.7%	1.7%	6.2%
Average		-13.5%	-6.7%	-4.4%
Kimono	2	-27.0%	-18.9%	-16.8%
ParkScene	2	-14.8%	-12.0%	-12.1%
Cactus	2	-18.4%	-14.4%	-10.3%
BasketballDrive	2	-23.0%	-12.5%	-13.0%
BQTerrace	2	-10.2%	-2.1%	0.1%
Average		-18.7%	-12.0%	-10.4%

Table 12: Test results for IRAP non-aligned Class B (BL 720p – EL 1080p)

Test Sequences	QP Diff of BL & EL	BD-Rate Comparison		
		SHVC Vs. Simulcast		
		Y	U	V
Kimono	0	-28.4%	-20.7%	-18.3%
ParkScene	0	-20.2%	-15.9%	-15.2%
Cactus	0	-25.4%	-21.4%	-14.5%
BasketballDrive	0	-26.8%	-15.8%	-16.2%
BQTerrace	0	-16.6%	3.0%	7.5%
Average		-23.5%	-14.2%	-11.3%
Kimono	2	-40.6%	-34.5%	-33.1%
ParkScene	2	-29.6%	-25.2%	-25.4%
Cactus	2	-34.3%	-32.2%	-29.0%
BasketballDrive	2	-36.4%	-30.5%	-30.1%
BQTerrace	2	-23.4%	-13.1%	-13.5%
Average		-32.9%	-27.1%	-26.2%

Table 13: Test results for IRAP non-aligned Class B (BL 540p – EL 1080p)

Test Sequences	QP Diff of BL & EL	BD-Rate Comparison		
		SHVC Vs. Simulcast		
		Y	U	V
Kimono	0	-18.9%	-11.0%	-8.5%
ParkScene	0	-11.7%	-9.6%	-9.0%
Cactus	0	-16.3%	-12.8%	-6.4%
BasketballDrive	0	-17.0%	-5.0%	-6.5%
BQTerrace	0	-10.1%	-0.1%	3.4%
Average		-14.8%	-7.7%	-5.4%
Kimono	2	-26.8%	-18.3%	-16.0%
ParkScene	2	-15.9%	-13.1%	-12.7%
Cactus	2	-20.6%	-17.0%	-12.5%
BasketballDrive	2	-23.4%	-12.7%	-13.3%
BQTerrace	2	-13.2%	-4.8%	-3.5%
Average		-20.0%	-13.2%	-11.6%

6.2.3.2 Decoding complexity

Decoding complexity or overhead at UEs depends on how many layers an UE needs to decode. For solution with HEVC simulcast, an UE needs to decode one layer stream, i.e. either stream of 720p@30fps or stream of 1080p@30fps. Decoding complexity for UEs receiving normal-service when the solution with SHVC is used can be assumed the same as when solution with HEVC simulcast is used because UEs receiving normal-service can ignore coded data for enhancement layer.

The decoding complexity overhead for UEs receiving premium-service when the SHVC solution is used is roughly the percentage of the number of samples in the lower resolution video relative to that in the higher resolution video.

6.2.3.3 Encoding complexity

For the solution with HEVC simulcast, the content provider has to encode one 1080p@30fps video stream and one 720p@30fps video stream; on the other hand, for the solution with SHVC, the content provider has to encode one stream with two layers (i.e., 720p@30fps for base layer and 1080p@30fps). Compared to simulcast, the complexity of SHVC encoding at UE A is typically less than that of simulcast encoding. This is because SHVC places zero motion constraint on the inter layer reference picture. When the inter layer reference picture provides sufficiently good prediction signal (without the need for motion estimation), early termination is typically applied at the encoder, and the need for motion estimation of the temporal reference pictures is avoided, leading to lower encoding complexity.

6.3 3GP-DASH

6.3.1 The 3GP-DASH use case

This use case considers providing 3GP-DASH video streaming services to multiple end user devices. A diverse of end user devices could be with different display capabilities and network access conditions. Each end user device may prefer receiving a different quality of a content, possibly with a different resolution, and request the chosen video content from the origin server, involving caches between the origin server and the UEs. During a session, a UE may also adaptive switch to segments of different representations of different bitrates and qualities, and possibly also different spatial resolutions, to adapt to the dynamic network conditions. As shown in Figure 17, a video content is encoded into multiple video streams in different representations providing different levels of resolutions or qualities, e.g. as 4 Representations of resolutions 360p@30fps, 720p@30fps, 1080p@30fps and 1080p@60fps in an Adaptation Set. Copies of the streams may be stored in the caches and directly served to the UEs.

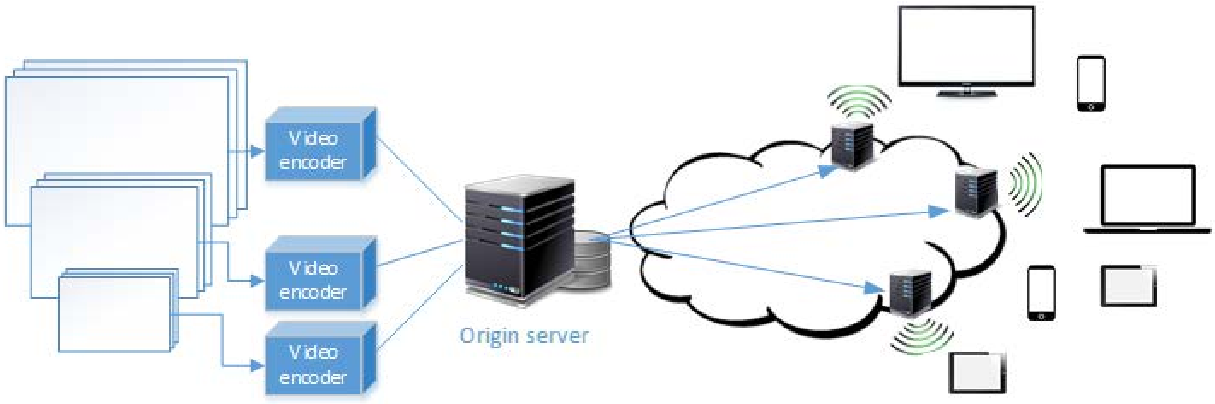

Figure 17: 3GP-DASH use case

The metrics for comparison of the potential solutions for this use case are:

1) Bandwidth used for outgoing video transmission from the origin server to (the first level) caches

2) Bandwidth used for incoming video transmission to the UEs

3) Video quality received by the UEs

4) Decoding complexity of UEs

5) Encoding complexity

6.3.2 Solutions for the 3GP-DASH use case

6.3.2.1 HEVC simulcast

One solution from video coding point of view for the 3GP-DASH use case is to use HEVC simulcast, where each resolution or quality representation can be encoded into an independent HEVC single-layer bitstream and stored on the origin server and probably also caches in 3GP-DASH segments. Based on the client requests, the corresponding representation segments are delivered to the client. Each end user device needs to decode an HEVC single-layer bitstream.

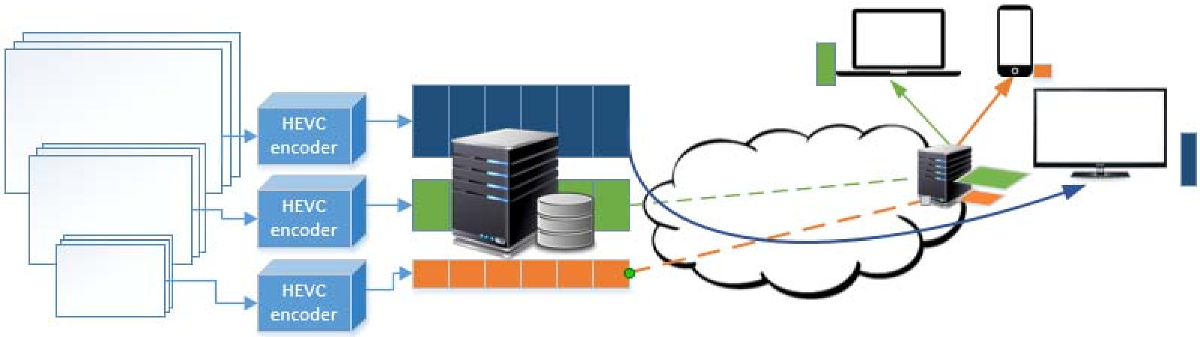

Figure 18: HEVC simulcast in 3GP-DASH use case

Figure 18 shows such HEVC simulcast example, 3 representation resolutions are encoded into 3 bitstreams and stored in the origin server. The medium and low resolution stream copies can be replicated on the cache, but not the high resolution stream due to a storage size limit of the cache. The server or the cache may stream the corresponding representation segment to the client based on the client's streaming request.

6.3.2.2 SHVC

Another solution from video coding point of view for the 3GP-DASH use case is to use SHVC, where multiple resolutions or quality representations can be encoded into multi-layer SHVC bitstreams. Herein each layer can be encapsulated as one 3GP-DASH Representation. A client wanting a particular resolution or quality can request segments of that Representation and all other Representations it depends on (i.e. request the desired layer and all layers the desired layer depends on). The request layer and all its dependent layers will then be sent to the client and the client can decode the bitstream and output the desired layer.

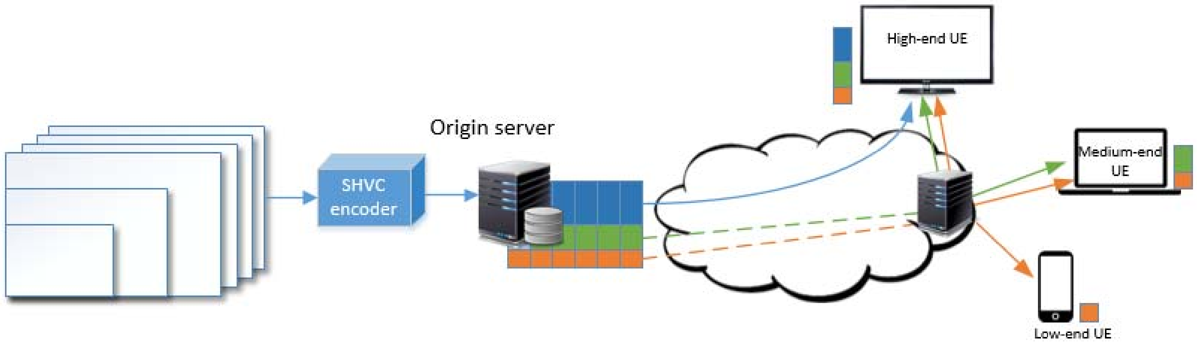

Figure 19: SHVC in 3GP-DASH

Figure 19 shows an example of using SHVC in the 3GP-DASH use case. A single SHVC stream is stored in the origin server and replicated in the edge cache supporting 3 spatial resolution representations. Each client may request the base layer representation for phone device, medium layer representation on laptop and high representation for TV display from the edge cache.

6.3.3 Comparison of the solutions

6.3.3.0 General

For comparison of solutions with HEVC simulcast and SHVC, simulations with three representation of spatial resolution 360p, 720p and 1080p have been conducted.

For outgoing transmission bandwidth, SHVC solution requires less bandwidth for transmitting the encoded streams from origin server to cache and to UEs. The bandwidth reduction varies from 9.22% to 10.52% for transmitting both 360p and 720p resolution streams and from 23.34% to 23.62% for transmitting 360p, 720p and 1080p resolution streams.

For incoming transmission bandwidth, SHVC solution has data overhead for UEs when receiving medium to high resolution representation. The overhead varies from 20.4% to 22.1% when receiving representation with highest resolution 720p and from 24.9% to 26.88% when receiving representation with highest resolution 1080p.

6.3.3.1 Outgoing transmission bandwidth

6.3.3.1.0 General

Comparison of bandwidth used for transmission for solution with HEVC simulcast and SHVC can be analysed based on BD-rate difference between the two solutions. As described in the use-case, there are two outgoing transmissions, that is, from origin server to cache and from origin server to UEs.

6.3.3.1.1 From origin server to caches

For transmission from origin server to the caches, the transmission only involve low and medium resolution representation. Table 14 and Table 15 tabulate BD-rate reduction for coding of 720p resolution that SHVC can achieved when compared to HEVC simulcast.

Table 14: BD-rate decrease for coding of 720p given by SHVC over HEVC simulcast for case of IRAP aligned.

IRAP Aligned	SHVC vs. HEVC Simulcast		
	Y	U	V
Kimono	-17.3%	-11.9%	-10.1%
ParkScene	-5.6%	-6.4%	-4.7%
Cactus	-8.5%	-4.5%	-3.0%
BasketballDrive	-13.7%	-3.8%	-6.3%
BQTerrace	-1.0%	-0.7%	1.9%
Average	-9.22%	-5.46%	-4.44%

Table 15: BD-rate decrease for coding of 720p given by SHVC over HEVC simulcast for case of IRAP non-aligned

IRAP Non-Aligned	SHVC vs. HEVC Simulcast		
	Y	U	V
Kimono	-16.8%	-10.7%	-8.8%
ParkScene	-7.2%	-7.5%	-6.0%
Cactus	-10.2%	-6.3%	-4.7%
BasketballDrive	-13.9%	-3.4%	-6.2%
BQTerrace	-4.5%	-3.3%	-1.2%
Average	-10.52%	-6.24%	-5.38%

6.3.3.1.2 From origin server to UEs

For transmission from origin server to the UEs, the transmission only involves high resolution representation. Table 16 and Table 17 tabulate BD-rate reduction that SHVC can achieve when compared to HEVC simulcast.

Table 16: BD-rate decrease for coding of 1080p given by SHVC over HEVC simulcast for case of IRAP aligned.

IRAP Aligned	SHVC vs. HEVC Simulcast		
	Y	U	V
Kimono	-30.0%	-21.4%	-18.7%
ParkScene	-19.5%	-14.7%	-14.0%
Cactus	-23.8%	-18.1%	-9.7%
BasketballDrive	-27.5%	-11.4%	-13.7%
BQTerrace	-15.9%	8.4%	17.1%
Average	-23.34%	-11.44%	-7.80%

Table 17: BD-rate decrease for coding of 1080p given by SHVC over HEVC simulcast for case of IRAP non-aligned

IRAP Non-Aligned	SHVC vs. HEVC Simulcast		
	Y	U	V
Kimono	-30.0%	-20.7%	-17.9%
ParkScene	-20.3%	-15.1%	-14.4%
Cactus	-24.1%	-18.3%	-9.7%
BasketballDrive	-27.3%	-10.7%	-12.9%
BQTerrace	-16.4%	8.2%	17.0%
Average	-23.62%	-11.32%	-7.58%

6.3.3.2 Incoming transmission bandwidth

As described in the use-case, depending on which representation an UE chooses and depending on which solution is used, an UE may receive data from origin server, cache or from both origin server and cache.

When an UE chooses the low resolution representation (i.e. 360p resolution), the incoming transmission bandwidth is the same for HEVC simulcast and SHVC solution. When an UE chooses the medium resolution representation (i.e. 720p), with HEVC simulcast solution, the UE receives a single layer of 720p video resolution; whereas with SHVC solution, the UE receives two layers (base layer 360p and enhancement layer 720p). Likewise, when an UE chooses the high resolution representation (i.e. 1080p resolution), with HEVC simulcast solution, the UE receives a single layer of 1080p; whereas with SHVC solution, the UE receives three layers of 360p, 720p and 1080p.

The SHVC bandwidth overhead comparing to HEVC simulcast solution bandwidth is tabulated in Table 18 to Table 21.

Table 18: BD-rate overhead for choosing medium resolution representation (720p) when SHVC solution is used.

IRAP Aligned	SHVC vs. HEVC Simulcast		
	Y	U	V
Kimono	17.3%	25.0%	27.5%
ParkScene	23.2%	22.1%	24.2%
Cactus	24.7%	30.1%	32.2%
BasketballDrive	19.7%	33.4%	29.9%
BQTerrace	25.6%	25.7%	29.0%
Average	22.10%	27.26%	28.56%

Table 19: BD-rate overhead for choosing medium resolution representation (720p) when SHVC solution is used

IRAP Non-Aligned	SHVC vs. HEVC Simulcast		
	Y	U	V
Kimono	18.0%	26.7%	29.4%
ParkScene	21.1%	20.6%	22.6%
Cactus	22.3%	27.6%	29.8%
BasketballDrive	19.4%	33.9%	30.0%
BQTerrace	21.2%	22.4%	25.1%
Average	20.40%	26.24%	27.38%

Table 20: BD-rate overhead for choosing high resolution representation (1080p) when SHVC solution is used.

IRAP Aligned	SHVC vs. HEVC Simulcast		
	Y	U	V
Kimono	28.6%	44.6%	49.6%
ParkScene	34.7%	42.6%	43.6%
Cactus	30.9%	39.4%	56.1%
BasketballDrive	25.7%	54.3%	50.1%
BQTerrace	14.5%	52.1%	64.9%
Average	26.88%	46.60%	52.86%

Table 21: BD-rate overhead for choosing high resolution representation (1080p) when SHVC solution is used

IRAP Non-Aligned	SHVC vs. HEVC Simulcast		
	Y	U	V
Kimono	28.2%	45.2%	50.5%
ParkScene	30.6%	39.2%	40.2%
Cactus	27.7%	36.4%	52.9%
BasketballDrive	25.2%	54.6%	50.6%
BQTerrace	12.8%	49.9%	62.7%
Average	24.90%	45.06%	51.38%

6.3.3.3 Decoding complexity

Decoding complexity is mainly proportional to the resolution(s) of the video represented in the bitstream. For HEVC simulcast solution, only one single layer stream needs to be decoded, i.e. either stream of 360p@30fps, 720p@30fps, 1080p@30fps or stream of 1080p@60fps. For SHVC solution, the decoding complexity depends on the resolution of each layer needs to be decoded in order to output the highest layer video resolution.

6.3.3.4 Encoding complexity

For HEVC simulcast solution, the content provider has to encode streams for each representations (i.e. 360p@30fps, 720p@30fps, 1080p@30fps and 1080@60fps). For SHVC solution, the content provider has to encode streams with multiple layers in which each layer is associated with one representations. Compared to simulcast, the complexity of SHVC encoding at UE A is typically less than that of simulcast encoding. This is because SHVC places zero motion constraint on the inter layer reference picture. When the inter layer reference picture provides sufficiently good prediction signal (without the need for motion estimation), early termination is typically applied at the encoder, and the need for motion estimation of the temporal reference pictures is avoided, leading to lower encoding complexity.

7 Test cases, conditions, and results

7.1 Test cases and conditions

7.1.1 General conditions

The following test conditions apply for multi-stream multiparty video conferencing (MMVC), MBMS and 3GP-DASH tests:

Codecs (profiles): HEVC Main profile vs Scalable Main profile

Test sequences, resolutions and frame rates for MMVC and MBMS:

Two sets of the test sequences are used, as follows:

720p@30fps / 1080p@30fps, corresponding to the Class E and B sequences as listed in Table 2 of TR 26.906 (the relevant parts of the table is copied below for convenience). For the sequences Kimono and ParkScene, for which the original sequences were only of 24 fps, the frame rate to be used is 24 fps for both resolutions. For the sequences Cactus and BasketballDrive, for which the original sequences were of 50 fps, the frame rate to be used is 25 fps. For the sequences BQTerrace, FourPeople, Johnny and KristenAndSara, for which the original sequences were of 60 fps, the frame rate to be used is 30 fps. For the sequences which the original sequence are of 50fps or 60 fps, the frames with odd indices (assuming the initial index being 0) are dropped to achieve the half frame rate.

Table 22: Class B and E test sequences for MMVC and MBMS tests

Class	Sequence	Spatial resolution	Frame rate
Class B	Kimono	1920x1080	24 fps
	ParkScene	1920x1080	24 fps
	Cactus	1920x1080	50 fps
	BasketballDrive	1920x1080	50 fps
	BQTerrace	1920x1080	60 fps
Class E	Kimono_720p	1280x720	24 fps
	ParkScene_720p	1280x720	24 fps
	Cactus_720p	1280x720	50 fps
	BasketballDrive_720p	1280x720	50 fps
	BQTerrace_720p	1280x720	60 fps

720p@30fps / 640x360p@30fps, corresponding to the Class VC-E sequences listed in Table 5 of TR 26.906 (the table is copied below for convenience) and their sub-sampled (both spatially and temporally) version.

Table 23: Class VC-E test sequences for MMVC tests

Class	Sequence	Spatial resolution	Frame rate
Class VC-E	FourPeople	1280x720	60 fps
	Johnny	1280x720	60 fps
	KristenAndSara	1280x720	60 fps

RAP distance for MMVC and MBMS: 2 seconds (i.e. one RAP every 64 pictures when frame rate is 30 fps and every 48 pictures when frame rate is 25 fps)

For SHVC encoding, two options are tested, where the first option is with cross-layer aligned RAPs with RAP distance of 2 seconds for both layers, and the second option is with RAP distance of 2 seconds for the base layer, and longer RAP distance of 4 seconds for the enhancement layer (i.e. one RAP every 128 pictures when frame rate is 30 fps and every 96 pictures when frame rate is 25 fps).

QP Configuration

Fixed QP configuration is used without rate control to avoid uncertainty due to different rate control algorithms. Cascaded QP setting (e.g. higher QP for P pictures than I pictures, higher QP for B pictures than P pictures is allowed.

When temporal level is used, higher QP for higher temporal level than lower temporal level in hierarchical coding structures) is allowed.

For coding of enhancement layer: Two sets of delta QP values, deltaQP = {0, +2}. The delta QP value specifies the difference between initial EL QP and initial BL QP for collocated pictures in the two layers. For example, when deltaQP = 0, for each picture of a particular resolution, the same QP that was used in base-layer is used; when deltaQP = 2, for each picture in enhancement layer, the QP that was used in base-layer plus 2 is used.

Rate-distortion optimized quantization. Rate-distortion optimized quantization is disabled.

7.1.2 MMVC

For multi-stream multiparty video conferencing, the video bitstreams should be encoded with low-delay coding structure where the decoding order of pictures is identical to the presentation order to minimize the delay introduced by the codec. To enable late tuning-in, insertion of frequent random access points (RAPs) is needed. A key point that needs to be determined for multi-stream multiparty video conferencing is whether simulcast of multiple single-layer HEVC bitstreams or multiple layers per SHVC should be used.

With the above points in mind, the following test conditions are specified for testing of potential video codecs for multi-stream multiparty video conferencing:

Input test sequences, resolutions and frame rates are as listed in clause 7.1.1

Encoding settings

For single-layer coding (including coding of the base layer in multi-layer coding): similar as for MTSI tests specified in TR 26.906 (with the exception that temporal scalability is not used, for simplicity), as follows:

QP configuration

For Class B sequences, the following QP set is used {25, 28, 31, 34}.

For Class VC-E sequences, the following QP set is used {19, 22, 25, 28}.

Number of reference pictures in the reference picture list is set equal to 2.

GOP and prediction structures

The IPPP coding structure, wherein the first picture in each random access point period is an IDR picture and the rest are P pictures, and the decoding order equals the output order, is used.

The previous two pictures in decoding order are always used for prediction.

Temporal scalability is not enabled.

Motion vector search range: The motion vector search range, in units of integer luma samples, is restricted to 32 in both directions.

MTU size matching and multiple slices are allowed. The size of each slice in a picture is set to 1200 bytes, with the exception that the last slice in each picture is allowed to have a smaller size.

For SHVC coding of the enhancement layer, the following encoding settings are used:

QP configuration

For Class B sequences, the following QP set is used {25, 28, 31, 34}.

For Class VC-E sequences, the following QP set is used {19, 22, 25, 28}.

Number of reference pictures in the reference picture list is set equal to 3.

GOP and prediction structure structures

The IPPP coding structure, wherein the first picture in each random access point period is an IDR picture and the rest are P pictures, and the decoding order equals the output order, is used.

If there is no lower layer picture in the same access unit, the previous two pictures of the same layer in decoding order are always used for prediction. Otherwise, the previous two picture of the same layer in decoding order and the lower layer picture in the same access units are used for prediction (except for an IRAP picture only inter-layer prediction is used).

Temporal scalability is not enabled.

Motion vector search range: The motion vector search range, in units of integer luma samples, is restricted to 32 in both directions.

MTU size matching and multiple slices are allowed. The size of each slice in a picture is set to 1200 bytes, with the exception that the last slice in each picture is allowed to have a smaller size.

7.1.3 MBMS

For MBMS, the video bitstreams should be encoded with the so-called random access coding structure to achieve the highest compression efficiency. To enable stream or layer switching in DASH or late tuning-in and channel switching in MBMS, insertion of frequent random access points (RAPs) is needed. Two different scenarios for enhancement should be tested: enhancement of video spatial resolution and enhancement of quality.

With the above points in mind, the following test conditions are specified for testing of potential video codecs for multi-stream multiparty video conferencing:

Input test sequences from Class B, resolutions and frame rates are as listed in clause 7.1.1

Encoding settings

For single-layer coding (including coding of the base layer in multi-layer coding):

QP configuration: For Class B sequences, the following QP set is used {22, 25, 28, 31}.

Number of reference pictures in each reference picture lists (for forward prediction and backward prediction) is set equal to 2.

GOP and Prediction structures

GOP size is 8.

The hierarchical B-picture structure as used for each layer as in the HEVC random access common test condition.

IRAP picture type is CRA except for the first IRAP picture, for which IDR is used.

Temporal scalability is not enabled.

Motion vector search range: The motion vector search range, in units of integer luma samples, is restricted to 64 in both directions.

MTU size matching is not enabled.

For SHVC coding of the enhancement layer, the following encoding settings are used:

QP configuration: For Class B sequences, the following QP set is used {22, 25, 28, 31}

Number of reference pictures in each reference picture lists for base layer is set equal to 2 and for enhancement layer is set equal to 3 (2 from the same layer and 1 from the base layer).

GOP and Prediction structures

GOP size is 8.

The hierarchical B-picture structure as used in the SHVC random access common test condition.

IRAP picture type is CRA except for the first IRAP picture, for which IDR is used.

Temporal scalability is not enabled.

Motion vector search range: The motion vector search range, in units of integer luma samples, is restricted to 64 in both directions.

MTU size matching is not enabled.

7.1.4 3GP-DASH

For 3GP-DASH, the video bitstreams should be encoded with the following features:

To achieve highest compression efficiency, random access coding structure is used.

To enable layer switching in DASH or late tuning-in and channel switching, insertion of frequent random access points (RAPs) is needed.

To enable frequent switching to higher layer / representation, higher layer may have more frequent RAPs.

With the above points in mind, the following test conditions are specified for testing of potential video codecs for multi-stream multiparty video conferencing:

Input test sequences, resolutions and frame rates are as follows:

Input test sequences are test sequences from class B sequences.

Three layers with the following spatial resolutions 1080p / 720p / 360p

Frame rates are either 24, 25 or 30 fps depending on the sequences. For the sequences Cactus, BasketballDrive and BQTerrace for which the original sequence are of 50fps or 60 fps, the frames with odd indices (assuming the initial index being 0) are dropped to achieve the half frame rate.

Table 24: Test sequences for 3GP-DASH test

Sequence	Spatial resolution	Frame rate
Kimono	1920x1080	24
	1280x720	24
	640x360	24
ParkScene	1920x1080	24
	1280x720	24
	640x360	24
Cactus	1920x1080	25
	1280x720	25
	640x360	25
BasketballDrive	1920x1080	25
	1280x720	25
	640x360	25
BQTerrace	1920x1080	30
	1280x720	30
	640x360	30

Encoding settings

For single-layer coding (including coding of the base layer in multi-layer coding):

QP configuration: For Class B sequences, the following QP set is used {22, 25, 28, 31}.

Number of reference pictures in each reference picture lists (for forward prediction and backward prediction) is set equal to 2.

Temporal scalability is enabled.

GOP and Prediction structures

GOP size is 8.

The hierarchical B-picture structure as used for each layer as in the HEVC random access common test condition.

IRAP picture type is CRA except for the first IRAP picture, for which IDR is used.

RAP distance:

For base layer: every 4 seconds (i.e. one RAP every 128 pictures when frame rate is 30 fps and every 96 pictures when frame rate is 24 or 25 fps).

For enhancement layers, two options are tested:

Option 1: every 4 seconds (i.e. one RAP every 128 pictures when frame rate is 30 fps and every 96 pictures when frame rate is 24 or 25 fps).

Option 2: every 2 seconds (i.e. one RAP every 64 pictures when frame rate is 30 fps and every 48 pictures when frame rate is 24 or 25 fps).

Inter-layer prediction: No inter-layer prediction is used.

Motion vector search range: The motion vector search range, in units of integer luma samples, is restricted to 64 in both directions.

MTU size matching is not enabled.

For SHVC coding of the enhancement layer, the following encoding settings are used:

QP configuration: For Class B sequences, the following QP set is used {22, 25, 28, 31}.

Number of reference pictures in each reference picture lists for base layer is set equal to 2 and for enhancement layer is set equal to 3 (2 from the same layer and 1 from the base layer).

Temporal scalability

Temporal scalability is enabled.

GOP and Prediction structures

GOP size is 8.

The hierarchical B-picture structure as used in the SHVC random access common test condition.

IRAP picture type is CRA except for the first IRAP picture, for which IDR is used.

RAP distance:

For base layer: 4 seconds (i.e. one RAP every 128 pictures when frame rate is 30 fps and every 96 pictures when frame rate is 24 or 25 fps).

For enhancement layers, two options are tested:

Option 1: 4 seconds (i.e. one RAP every 128 pictures when frame rate is 30 fps and every 96 pictures when frame rate is 24 or 25 fps).

Option 2: 2 seconds (i.e. one RAP every 64 pictures when frame rate is 30 fps and every 48 pictures when frame rate is 24 or 25 fps).

Inter-layer prediction (ILP)

Linear dependency structure is used for ILP, that is, pictures in layer *n*, where *n > 0*, may use only collocated pictures from layer *n – 1* as reference for ILP.

Inter-layer prediction (ILP) is used with the following constraints:

Option 1: No constraint. All lower layer pictures are used for ILP references.

Option 2: Only IRAP pictures and pictures at temporal sub-layer 0 are used for ILP references.

Option 3: Only pictures up to temporal sub-layer 1 are used for ILP references.

Option 4: Only pictures up to temporal sub-layer 2 are used for ILP references.

Motion vector search range: The motion vector search range, in units of integer luma samples, is restricted to 64 in both directions.

MTU size matching is not enabled.

7.2 Test results

7.2.1 MMVC

7.2.1.1 General

In this clause, simulation results for the multi-stream multiparty video conferencing service are provided, comparing SHVC vs HEVC simulcast. A particular value of the BD-rate decrease of SHVC comparing simulcast indicates how much less bandwidth, in percentage, is needed for transmission of the two-layer SHVC bitstream compared to transmission of both HEVC single-layer bitstreams, on average for the same quality of the higher resolution video. The comparison indicates the difference of the bandwidth requirements for SHVC vs simulcast in the network link between a sender and bitstream-switching MCU in the multi-stream multiparty video conferencing service.

7.2.1.2 Results for aligned IRAP pictures case

BD-rate results for cross layer IRAP aligned case is presented in Table 25 and Table 26.

For 1.5x spatial scalability, SHVC has overall 27.34% BD-rate decrease comparing simulcast when deltaQP is 2, and the max gain can be up to 35.5%.

For 2x spatial scalability scenario, SHVC has overall 8.03% BD-rate decrease comparing simulcast when the deltaQP is 2, and the max gain can be up to 10.6%.

Table 25: MMVC IRAP 1.5x aligned results (SHVC vs. simulcast)

	1.5x spatial scalability (Class-B)			
	deltaQP	**Y**	**U**	**V**
Kimono	0	-21.5%	-21.4%	-20.4%
ParkScene	0	-13.1%	-10.0%	-9.1%
Cactus	0	-18.8%	-14.3%	-10.9%
BasketballDrive	0	-24.2%	-17.4%	-16.9%
BQTerrace	0	-8.5%	5.7%	13.6%
Average gain SHVC vs. Simulcast		-17.22%	-11.48%	-8.74%
Kimono	2	-35.5%	-36.3%	-34.8%
ParkScene	2	-22.7%	-18.1%	-18.0%
Cactus	2	-29.7%	-27.1%	-25.3%
BasketballDrive	2	-33.6%	-29.6%	-28.7%
BQTerrace	2	-15.2%	-3.2%	-2.2%
Average gain SHVC vs. Simulcast		-27.34%	-22.86%	-21.8%

Table 26: MMVC IRAP 2x aligned results (SHVC vs. simulcast)

	2x spatial scalability (Class-VC_E)			
	deltaQP	**Y**	**U**	**V**
FourPeople	0	-7.7%	-2.7%	-0.1%
Johnny	0	-2.8%	4.3%	7.2%
KristenAndSara	0	-6.3%	-2.6%	-0.1%
Average gain SHVC vs. Simulcast		-5.6%	-0.33%	2.33%
FourPeople	2	-10.6%	-4.9%	-1.8%
Johnny	2	-4.5%	6.2%	7.3%
KristenAndSara	2	-9.0%	-3.2%	-2.8%
Average gain SHVC vs. Simulcast		-8.03%	-0.63%	0.9%

7.2.1.3 Results for IRAP non-aligned test case

BD-rate results for cross-layer RAP non-aligned case is presented in Table 27 and Table 28.

For 1.5x spatial scalability scenario, SHVC has overall 27.9% BD-rate decrease comparing simulcast when deltaQP is 2, and the max gain can be up to 35.6%.

For 2x spatial scalability scenario, SHVC has overall 10.76% BD-rate decrease comparing simulcast when deltaQP is 2, and the max gain can be up to 12.8%.

Table 27: MMVC IRAP 1.5X non-aligned results (SHVC vs. simulcast)

	1.5x spatial scalability (Class-B)			
	deltaQP	**Y**	**U**	**V**
Kimono	0	-21.6%	-21.4%	-20.3%
ParkScene	0	-14.1%	-10.5%	-9.0%
Cactus	0	-20.1%	-15.7%	-12.3%
BasketballDrive	0	-24.4%	-17.2%	-16.7%
BQTerrace	0	-9.4%	3.6%	9.6%
Average gain SHVC vs. Simulcast		-17.92%	-12.24%	-9.74%
Kimono	2	-35.6%	-36.4%	-34.6%
ParkScene	2	-23.4%	-18.3%	-17.5%
Cactus	2	-30.7%	-28.2%	-26.4%
BasketballDrive	2	-33.7%	-29.6%	-28.5%
BQTerrace	2	-16.3%	-5.0%	-6.1%
Average gain SHVC vs. Simulcast		-27.94%	-23.5%	-22.62%

Table 28: MMVC IRAP 2X non-aligned results (SHVC vs. simulcast)

		2x spatial scalability (Class-VC_E)		
	deltaQP	**Y**	**U**	**V**
FourPeople	0	-9.6%	-5.6%	-2.7%
Johnny	0	-5.8%	0.1%	2.3%
KristenAndSara	0	-8.3%	-5.7%	-2.9%
Average gain SHVC vs. Simulcast		-7.9%	-3.73%	-1.1%
FourPeople	2	-12.8%	-7.7%	-4.5%
Johnny	2	-8.2%	1.7%	1.6%
KristenAndSara	2	-11.3%	-6.0%	-5.7%
Average gain SHVC vs. Simulcast		-10.76%	-4%	-2.86%

7.2.1.4 First set of additional results

Additional down-sampled sequences for Class VC-E were tested where the base layer is 960x540 and enhancement layer is class VC-E.

With the same RAP aligned test conditions, SHVC has overall 17% BD-rate decrease comparing simulcast when deltaQP is 2, and the max gain can be up to 19.8%.

With the same non-RAP aligned test conditions, SHVC has overall 18.76% the BD-rate decrease comparing simulcast when deltaQP is 2, and the max gain can be up to 21.2%.

Table 29: MMVC IRAP aligned results class-VC_E (BL540P/EL720P)

	1.5x spatial scalability (Class-VC_E)			
	deltaQP	Y	U	V
FourPeople	0	-11.6%	-6.6%	-5.1%
Johnny	0	-7.3%	-0.3%	1.9%
KristenAndSara	0	-12.1%	-9.3%	-8.2%
Average gain SHVC vs. Simulcast		-10.33%	-5.4%	-3.8%
FourPeople	2	-17.4%	-16.5%	-16.7%
Johnny	2	-13.8%	-13.6%	-15.6%
KristenAndSara	2	-19.8%	-20.1%	-22.1%
Average gain SHVC vs. Simulcast		-17%	-16.73%	-18.13%

Table 30: MMVC IRAP non-aligned results class-VC_E (BL540P/EL720P)

	1.5x spatial scalability (Class-VC_E)			
	deltaQP	Y	U	V
FourPeople	0	-13.0%	-9.2%	-7.1%
Johnny	0	-9.7%	-4.6%	-2.6%
KristenAndSara	0	-13.7%	-11.7%	-10.8%
Average gain SHVC vs. Simulcast		-12.13%	-8.5%	-6.83%
FourPeople	2	-18.8%	-18.5%	-18.4%
Johnny	2	-16.3%	-16.7%	-18.5%
KristenAndSara	2	-21.2%	-21.7%	-23.7%
Average gain SHVC vs. Simulcast		-18.76%	-18.96%	-20.2%

7.2.1.5 Second set of additional results

This clause presents additional simulation results for the use case where the previous active speaker sends both medium and thumbnail video, but uses SHVC for encoding. The results show one scenario with 240p resolution thumbnail (i.e., 3x in both width and height relative to the 720p resolution) and another scenario with 360p thumbnail (i.e., 2x in both width and height relative to the 720p resolution).

The Class E test sequences as listed in Table 22 were used for 2x and 3x spatial scalability MMVC tests between the medium-resolution main video and the thumbnail video. For spatial 3x scalable video coding between the medium-resolution main video and the thumbnail video, Class-E (1280x720) sequences were used as the enhancement layer and the corresponding subsampled sequences of 424x240 were used as the base layer. For spatial 2x scalable video coding between the medium-resolution main video and the thumbnail video, Class- E (1280x720) sequences were used as enhancement layer and the corresponding subsampled 640x360 sequences were used as base layer.

For both 3x and 2x spatial scalability cases, the QP settings for the base layer were {25, 28, 31, 34}, and the QPs used for the enhancement layer were {27, 30, 33, 36} and {29, 32, 35, 38}, respectively.

The following test conditions were used for the simulations:

1) The motion vector search range is 64.

2) Up to 2 active temporal reference pictures are used for inter prediction of each coding picture.

3) Cross-layer picture type aligned where the base layer and enhancement layer share the same RAP distance (approximately 2 seconds).

The SHM9.0 reference software was used for the simulation.

For the 3x spatial scalability scenario, SHVC has on average 6.7% uplink BD-rate saving and 13.1% downlink BD-rate cost comparing simulcast. The sequence-specific results are shown in the table below.

Table 31: MMVC 3X (720p/240p) spatial scalability results (SHVC vs. simulcast)

Sequences	Uplink BD-rate saving	Downlink BD-rate cost
Kimono	12%	13%
ParkScene	5%	12%
Cactus	6%	14%
BasketballDrive	8%	12%
BQTerace	2%	14%
Average	**6.7%**	**13.1%**

For 2x spatial scalability scenario, SHVC has on average 18.3% uplink BD-rate saving and 19.2% downlink BD-rate cost comparing simulcast. The sequence-specific results are shown in the table below.

Table 32: MMVC 2X (720p/360p) spatial scalability results (SHVC vs. simulcast)

Sequences	Uplink BD-rate saving	Downlink BD-rate cost
Kimono	28%	12%
ParkScene	17%	20%
Cactus	18%	20%
BasketballDrive	20%	14%
BQTerace	9%	30%
Average	**18.3%**	**19.2%**

Even though the bandwidth saving using SHVC for 3x ratio is less than that for 2x ratio, further investigation shows that the total uplink bandwidth usage between 2x and 3x is quite similar when SHVC is used.

Table 33 illustrates the details of the base layer (BL) bitrate, enhancement layer (EL) bitrate, BL quality and EL quality between 2x and 3x spatial scalability. It shows that the enhancement layer (720P) picture quality is almost the same (0.14 dB difference) for 2x and 3x scalability. Further, the overall bitrates of both layers are also approximately the same (1.57% difference) between 2x and 3x scalability. In other words, SHVC may be used as a tool by the previous active speaker (B) to provide higher resolution thumbnail (360p) instead of low resolution thumbnail (240p) at similar quality and similar uplink bandwidth.

Table 33: MMVC 2X (720p/360p) vs. 3X (720p/240p) comparison

			3x (reference)				2x					
	BL QP	**EL QP**	**EL rate kbps**	**BL rate kbps**	**EL+BL rate**	**EL PSNR**	**EL rate kbps**	**BL rate kbps**	**EL+BL rate**	**EL PSNR**	**Total Δbitrate**	**EL ΔPSNR**
Kimono	25	27	1595.53	425.20	2020.73	39.43	1122.21	830.64	1952.85	39.20	-3.36%	-0.23
	28	30	1013.95	276.44	1290.38	37.52	709.08	539.81	1248.89	37.29	-3.22%	-0.23
	31	33	625.92	176.42	802.33	35.65	427.72	346.09	773.81	35.43	-3.55%	-0.22
	34	36	410.92	110.16	521.09	33.98	291.34	216.00	507.34	33.78	-2.64%	-0.20
	25	29	1095.99	425.20	1521.19	38.11	572.66	830.64	1403.30	37.90	-7.75%	-0.21
	28	32	673.08	276.44	949.52	36.21	319.89	539.81	859.70	35.98	-9.46%	-0.23
	31	35	428.38	176.42	604.80	34.47	190.88	346.09	536.97	34.22	-11.21%	-0.25
	34	38	278.34	110.16	388.50	32.85	127.39	216.00	343.39	32.59	-11.61%	-0.26
ParkScene	25	27	2190.67	319.43	2510.11	36.77	1864.29	826.26	2690.55	36.70	7.19%	-0.07
	28	30	1329.03	208.27	1537.30	34.89	1097.99	523.13	1621.12	34.81	5.45%	-0.08
	31	33	784.47	135.48	919.95	33.10	624.01	328.87	952.88	33.00	3.58%	-0.10
	34	36	474.93	85.20	560.14	31.46	374.75	200.58	575.32	31.35	2.71%	-0.11
	25	29	1542.12	319.43	1861.55	35.49	1172.16	826.26	1998.42	35.39	7.35%	-0.10
	28	32	910.71	208.27	1118.98	33.66	639.47	523.13	1162.60	33.54	3.90%	-0.12
	31	35	539.92	135.48	675.40	31.96	349.03	328.87	677.91	31.82	0.37%	-0.14
	34	38	319.41	85.20	404.62	30.39	194.14	200.58	394.72	30.23	-2.45%	-0.16
Cactus	25	27	2888.89	528.40	3417.28	37.22	2401.35	1130.71	3532.07	37.09	3.36%	-0.12
	28	30	1849.04	365.98	2215.02	35.36	1505.61	763.81	2269.42	35.20	2.46%	-0.15
	31	33	1158.39	249.96	1408.35	33.50	918.78	510.58	1429.37	33.34	1.49%	-0.16
	34	36	767.70	167.09	934.79	31.86	609.73	334.29	944.02	31.69	0.99%	-0.17
	25	29	2083.81	528.40	2612.20	35.95	1522.23	1130.71	2652.95	35.77	1.56%	-0.17
	28	32	1298.58	365.98	1664.56	34.07	905.20	763.81	1669.01	33.90	0.27%	-0.17
	31	35	834.74	249.96	1084.70	32.35	550.96	510.58	1061.54	32.16	-2.14%	-0.19
	34	38	547.04	167.09	714.13	30.74	352.37	334.29	686.67	30.53	-3.85%	-0.21
Basketball	25	27	2699.12	509.06	3208.18	38.49	2225.02	1019.61	3244.63	38.38	1.14%	-0.12
	28	30	1755.88	355.22	2111.09	36.67	1410.98	704.50	2115.48	36.53	0.21%	-0.14
	31	33	1123.02	245.72	1368.73	34.83	873.35	483.40	1356.75	34.68	-0.88%	-0.15
	34	36	771.77	169.17	940.94	33.21	601.97	327.30	929.28	33.05	-1.24%	-0.16
	25	29	1959.66	509.06	2468.72	37.24	1449.48	1019.61	2469.10	37.15	0.02%	-0.09
	28	32	1242.72	355.22	1597.93	35.39	870.31	704.50	1574.81	35.30	-1.45%	-0.09
	31	35	822.24	245.72	1067.96	33.69	547.37	483.40	1030.77	33.58	-3.48%	-0.11
	34	38	560.24	169.17	729.41	32.09	367.16	327.30	694.46	31.95	-4.79%	-0.14
BQTerrace	25	27	2473.98	306.75	2780.73	36.52	2288.29	841.09	3129.38	36.50	12.54%	-0.02
	28	30	1439.46	194.54	1634.00	34.92	1309.08	522.95	1832.02	34.88	12.12%	-0.04
	31	33	861.50	127.37	988.87	33.29	765.35	330.20	1095.56	33.22	10.79%	-0.07
	34	36	541.77	83.01	624.78	31.70	474.68	208.00	682.68	31.59	9.27%	-0.10
	25	29	1704.54	306.75	2011.30	35.45	1489.65	841.09	2330.74	35.40	15.88%	-0.05
	28	32	1008.11	194.54	1202.65	33.83	850.13	522.95	1373.08	33.75	14.17%	-0.08
	31	35	623.26	127.37	750.63	32.21	502.34	330.20	832.55	32.09	10.91%	-0.13
	34	38	389.54	83.01	472.55	30.58	302.59	208.00	510.59	30.41	8.05%	-0.17

Average	**1.57%**	**-0.14**

7.2.1.6 Additional analysis for comparing SHVC and HEVC simulcast

7.2.1.6.1 Introduction

This clause presents additional cost and benefit analysis taking into account the varying numbers of premium UEs and regular UEs in an MMVC session, which may be different than the exact numbers in the example descriptions of the use cases.

Referring to Figure 10, this clause presents the cost/benefit analysis of SHVC vs. simulcast by taking into account the numbers of premium and regular UEs in an MMVC session, for the following cases:

- Case A: the current active speaker sends two video resolutions (one "high" @ 1080p + one "medium" @720p), coded using either SHVC or HEVC simulcast. The previous active speaker sends one video resolution ("medium" @720p) and one thumbnail (@240p) video, coded using HEVC simulcast. All other UEs send only one thumbnail video, coded using HEVC.
- Case B: the current active speaker sends one video resolution ("medium" @720p). The previous active speaker sends one video resolution ("medium" @720p) and one thumbnail (@240p), coded using either SHVC or HEVC simulcast. All other UEs send one thumbnail video, coded using HEVC. It is assumed in this case that the UEs who may be the previous active speaker support SHVC.
- Case C: the same as Case A, except that the previous active speaker also uses SHVC or HEVC simulcast to code the main video ("medium" @720p) and the thumbnail (@240p) video. In other words, Case C is a combination of Case A and Case B.

7.2.1.6.2 Uplink vs downlink transmission cost

In order to analyse the balance between uplink saving and downlink penalty using SHVC, an important factor, one that reflects the relative cost of uplink transmission and downlink transmission, needs to be determined. This is because the 3G and 4G LTE wireless channels (in fact as well as most wired channels) are often asymmetric in practice. Specifically, available uplink transmission bandwidth is often less than available downlink transmission bandwidth. For the analysis in the following two subclauses, it is assumed that the relative cost of uplink transmission is 2.375 times that of downlink transmission, based on averaging the related parameters of some mobile operators. It should be noted that this factor heavily affects the result of the analysis. For example, if a different value of the factor is assumed, then the result can be different.

7.2.1.6.3 Case by case cost/benefit analysis of SHVC vs HEVC simulcast

7.2.1.6.3.1 General

First the following notations are defined to be used in the analysis:

- **N**: total # of UEs in the MMVC session
- **H**: single layer rate of "premium" quality main video
- **M**: single layer rate of "medium" quality main video
- **T**: single layer rate of thumbnail video
- **a**: performance gain of SHVC over simulcast
- **b**: performance penalty of SHVC versus single layer
- **f**: a weighting factor that reflects the cost of uplink traffic relative to that of downlink traffic, as discussed in Clause 7.2.1.6.2.

7.2.1.6.3.2 Case A

The following additional notations are defined for Case A:

- **N_0**: total # of premium UEs receiving high quality video (not counting current active speaker)
- **N_1**: total # of regular UEs and current active speaker

Note that the reason why the current active speaker is not included in N_0 but included in N_1 is that the current active speaker receives medium quality video M from the past active speaker (due to the capabilities of the current and previous active speakers). Later in this clause, the case when the previous active speaker sends H instead of M is analysed (assuming both current and previous active speakers have high capability), which will show that basically the same conclusion can be drawn, regardless of whether the current active speaker is included in N_0 or N_1.

Table 34 lists the uplink and downlink bandwidth consumption depending on whether the UE is the active speaker, the last active speaker, any other UE, whether the UE is a premium or regular UE, and whether the UE uses simulcast or SHVC.

Table 34: Uplink and downlink traffic using simulcast and SHVC for each UE type for Case A (where the previous active speaker sends M)

	UE type	Simulcast	SHVC
Uplink	Active speaker	H + M	(H + M) * (1-a)
	Last active speaker	M + T	M + T
	All other UEs	(N-2) * T	(N-2) * T
Downlink	Premium UEs	N_0 * (H + T * (N-2))	N_0 * (H * (1+b) + T * (N-2))
	Normal UEs and active speaker	N_1 * (M + T * (N-2))	N_1 * (M + T * (N-2))

Factoring in f, the relative cost of uplink vs downlink transmission, T_simul, the total bandwidth using simulcast, is derived as follows:

$T\_simul = f*(H + M + M + T + (N-2)*T) + N_0*(H + T * (N-2)) + N_1*(M + T * (N-2))$

And the following for T_shvc, the total bandwidth using SHVC:

$T\_shvc = f*((H + M)*(1-a) + M + T + (N-2)*T) + N_0*(H*(1+b) + T * (N-2)) + N_1*(M + T * (N-2))$

Subtracting T_shvc from T_simul and simplifying the difference, the following is obtained:

$T\_simul – T\_shvc$

$= (f*(H + M) + N_0*H) – (f*((H + M)*(1-a)) + N_0*H*(1+b))$

$= f * (H + M) * a – N_0 * H * b$

Substituting a = 0.27, b = 0.24, M ≈ 0.5H (see the latest TR26.948 in Tdoc S4-151084), and f = 2.375 (see Clause 7.2.1.6.2), the following is obtained:

$T\_simul – T\_shvc$

$\approx 0.64 (H + M) – 0.24 * N_0 * H$

$\approx 0.96 H – 0.24 * N_0 * H$

For SHVC to have overall bandwidth reduction over simulcast, that is, $T\_simul – T\_shvc \geq 0$, there should be 4 or fewer premium UEs in the MMVC session, i.e. $N_0 \leq 4$ should be true. Otherwise, if $N_0 > 4$, then SHVC results in higher overall bandwidth than HEVC simulcast.

The above only considered the case when the previous active speaker is a regular UE sending normal ("medium" @720p) quality video (M). Deeper investigation shows that similar conclusion holds when the previous active speaker is a premium UE sending high quality video (H), as discussed below.

The following additional notations are defined:

- N_0: total # of premium UEs receiving high quality video (this will include current active speaker)

- N_1: total # of regular UEs (this will not include current active speaker)

Table 35: Uplink and downlink traffic using simulcast and SHVC for each UE type for Case A (last active speaker sends H)

	UE type	Simulcast	SHVC
Uplink	Active speaker	H + M	(H + M) * (1-a)
	Last active speaker	H + T	H + T
	All other UEs	(N-2) * T	(N-2) * T
Downlink	Premium UEs, not including current active speaker	(N_0-1) * (H + T * (N-2))	(N_0-1) * (H * (1+b) + T * (N-2))
	Current active speaker	H + T * (N-2)	H + T * (N-2)
	Normal UEs	N_1 * (M + T * (N-2))	N_1 * (M + T * (N-2))

Factoring in f, the relative cost of uplink vs downlink transmission, T_simul, the total bandwidth using simulcast, is derived as follows:

T_simul = f*(H + M + H + T + (N-2)*T) + N_0*(H + T * (N-2)) + N_1*(M + T * (N-2))

And the following for T_shvc, the total bandwidth using SHVC:

T_shvc = f*((H + M)*(1-a) + H + T + (N-2)*T) + (N_0-1)*(H*(1+b) + T * (N-2)) + (H + T * (N-2)) + N_1*(M + T * (N-2))

Subtracting T_shvc from T_simul and simplifying the difference, the following is obtained:

T_simul – T_shvc

= (f*(H + M) + (N_0-1)*H) – (f*(H + M)*(1-a) + (N_0-1)*H*(1+b))

= f * (H + M) * a – (N_0 – 1) * H * b

Substituting a = 0.27, b = 0.24, M ≈ 0.5H (see the latest TR26.948 in Tdoc S4-151084), and f = 2.375 (see Clause 7.2.1.6.2), the following is obtained:

T_simul – T_shvc

≈ 0.64 (H + M) – 0.24 * (N_0 – 1) * H

≈ 0.96 H – 0.24 * (N_0 – 1) * H

For SHVC to have overall bandwidth reduction over simulcast, that is, T_simul – T_shvc ≥ 0, there should be 5 or fewer premium UEs in the MMVC session, i.e. $N_0 \leq 5$ should be true. Otherwise, if $N_0 > 5$, then SHVC results in higher overall bandwidth than HEVC simulcast. Note that this is basically the same conclusion as above, since in the above analysis N_0 does not include the current active speaker, whereas N_0 here includes the current active speaker.

To summarize, *for Case A, the conclusion is that,* under the assumptions for this analysis, *SHVC outperforms simulcast if, excluding the current active speaker, an MMVC session has 4 or fewer premium UEs.*

7.2.1.6.3.3 Case B

No additional notations are necessary for Case B. Table 36 shows uplink/downlink bandwidth depending on the UE type and codec choice.

Table 36: Uplink and downlink traffic using simulcast and SHVC for each UE type for Case B

	UE type	Simulcast	SHVC
Uplink	Active speaker	M	M
	Last active speaker	M + T	(M + T) * (1-a)
	All other UEs	(N-2) * T	(N-2) * T
Downlink	Current active speaker	M + T * (N-2)	M * (1+b) + T * (N-2)
	All other UEs	(N-1) * (M + T * (N-2))	(N-1) * (M + T * (N-2))

Factoring in f, the relative cost of uplink vs downlink transmission, T_simul, the total bandwidth using simulcast, is derived as follows:

T_simul = f*(M + M + T + (N-2)*T) + (M + T * (N-2)) + (N-1)*(M + T * (N-2))

And the following for T_shvc, the total bandwidth using SHVC:

T_SHVC = f*(M + (M + T)*(1-a) + (N-2)*T) + (M*(1+b) + T * (N-2)) + (N-1)*(M + T * (N-2))

Subtracting T_SHVC from T_simul and simplifying the difference, the following is obtained:

T_simul – T_shvc

= (f*(M+T) + M) – ((f*(M+T)*(1-a) + M*(1+b))

= f * (M+T) * a – M * b

Substituting a = 0.067, b = 0.13, T ≈ 0.2M (see Tdoc S4-151317), and f = 2.375 (see Clause 7.2.1.6.2), the following is obtained:

T_simul – T_shvc

≈ 0.16 (M + T) – 0.13 * M

≈ 0.19 M – 0.13 M

≈ 0.06 M

Note that the above T_simul – T_shvc is always greater than 0. In other words, under the assumptions for this analysis, *SHVC always slightly outperforms simulcast for Case B.*

7.2.1.6.3.4 Case C

The following additional notations are defined for Case C:

- N_0: total # of premium UEs receiving high quality video (not counting current active speaker)
- N_1: total # of regular UEs and current active speaker

And substituting "a" and "b" as defined above with the following:

- a_1: performance gain of SHVC over simulcast when coding the two video resolutions of the current active speaker (H and M)
- b_1: performance penalty of SHVC versus single layer when coding the two video resolutions of the current active speaker (H and M)
- a_2: performance gain of SHVC over simulcast when coding the two video resolutions of the past active speaker (M and T)

- b_2: performance penalty of SHVC versus single layer when coding the two video resolutions of the past active speaker (M and T)

In this clause only the case when the current active speaker is not included in N_0 is analysed, as the conclusion holds regardless of whether the current active speaker is included in N_0 or not (as seen above in Clause 7.2.1.6.3.1).

Table 37: Uplink and downlink traffic using simulcast and SHVC for each UE type for Case C

	UE type	Simulcast	SHVC
Uplink	Active speaker	H + M	(H + M) * (1-a_1)
	Last active speaker	M + T	(M + T) * (1-a_2)
	All other UEs	(N-2) * T	(N-2) * T
Downlink	Premium UEs	N_0 * (H + T * (N-2))	N_0 * (H * (1+b_1) + T * (N-2))
	Current active speaker	M + T * (N-2)	M * (1+b_2) + T * (N-2)
	All other UEs	(N_1-1)* (M + T * (N-2))	(N_1-1) * (M + T * (N-2))

Factoring in f, the relative cost of uplink vs downlink transmission, T_simul, the total bandwidth using simulcast, is derived as follows:

T_simul = f*(H + M + M + T + (N-2)*T) + N_0*(H + T * (N-2)) + M + T * (N-2) + (N_1-1)*(M + T * (N-2))

And the following for T_shvc, the total bandwidth using SHVC:

T_SHVC = f*((H + M)*(1- a_1) + (M + T)*(1- a_2)+ (N-2)*T) + N_0*(H*(1+b_1) + T * (N-2)) + M * (1+b_2) + T * (N-2) + (N_1-1)*(M + T * (N-2))

Subtracting T_shvc from T_simul and simplifying the difference, the following is obtained:

T_simul – T_shvc

= f*(H + M + M + T) + N_0*H + M – (f*((H + M)*(1- a_1) + (M + T)*(1- a_2)) + N_0*H*(1+b_1) + M*(1+b_2))

= f * ((H+M) * a_1 + (M+T)*a_2) – N_0*H*b_1 – M * b_2

Substituting a_1 = 0.27, b_1 = 0.24, M ≈ 0.5H (see the latest TR26.948 in Tdoc S4-151084), a_2 = 0.067, b_2 = 0.13, T ≈ 0.2M ≈ 0.1H (see Tdoc S4-151317), and f = 2.375 (see Clause 7.2.1.6.2), the following is obtained:

T_simul – T_shvc

≈ 2.375 * (1.5H * 0.27 + 0.6H * 0.067) – N_0 *H* 0.24 – 0.5*H*0.13

≈ 1.057H – N_0*H* 0.24 – 0.065*H

≈ 0.99H – N_0*H* 0.24

For SHVC to have overall bandwidth reduction over simulcast, that is, T_simul – T_shvc ≥ 0, there should be 4.13 or fewer premium UEs in the MMVC session, i.e. $N_0 \leq 4.13$ should be true. Otherwise, if $N_0 > 4.13$, then SHVC results in higher overall bandwidth than HEVC simulcast.

Therefore, *the conclusion for Case C is similar to that for Case A.*

7.2.2 MBMS

7.2.2.0 General

In this clause, simulation results for the MBMS service are provided, comparing SHVC vs HEVC simulcast. It is assumed in this use case that all the premium UEs support SHVC. A particular value of the BD-rate decrease of SHVC comparing simulcast indicates how much less bandwidth, in percentage, is needed for transmission of the two-layer SHVC bitstream compared to transmission of both HEVC single-layer bitstreams, on average for the same quality of the higher resolution video. The comparison indicates the difference of the bandwidth requirements for SHVC vs simulcast in the network link between the Content Provider and the BM-SC, as well as the network link between the BM-SC and GGSN (for GPRS) or MBMS-GW (for EPS) in the MBMS service when different qualities of the same video content are provided.

7.2.2.1 Results for aligned IRAP pictures case

Table 38 and Table 39 show the summary of the test results. For the given test condition, SHVC provides BD-rate gains up to 40.5% for Kimono sequences with QP different of 2 between the first layer (i.e. QP set of 22, 25, 28, 31) and the second layer (i.e. QP set of 24, 27, 30, 33). The overall average gain is 31.9% for QP delta 2.

Table 38: MBMS IRAP aligned results class B (BL 720p – EL 1080p)

Test Sequences	deltaQP	BD-Rate Comparison		
		SHVC Vs. Simulcast		
		Y	U	V
Kimono	0	-28.4%	-20.9%	-18.7%
ParkScene	0	-19.5%	-15.3%	-15.1%
Cactus	0	-23.4%	-19.2%	-12.5%
BasketballDrive	0	-26.5%	-15.5%	-15.9%
BQTerrace	0	-14.9%	4.9%	10.4%
Average		-22.5%	-13.2%	-10.4%
Kimono	2	-40.5%	-34.5%	-33.3%
ParkScene	2	-28.9%	-24.6%	-25.2%
Cactus	2	-32.7%	-30.5%	-27.5%
BasketballDrive	2	-36.1%	-30.3%	-29.8%
BQTerrace	2	-21.5%	-11.2%	-10.5%
Average		-31.9%	-26.2%	-25.3%

For the configuration where the first layer and the second layer have spatial resolutions of 540p and 1080p, respectively, the performance of SHVC drops. The BD-rate gain is up to 27%. The overall average gain is 18.7% for QP delta 2. This is an expected result as a higher spatial resolution ratio between the first layer and the second layer means lower cross-layer correlation.

Table 39: MBMS IRAP aligned results class B (BL 540p – EL 1080p)

Test Sequences	deltaQP	BD-Rate Comparison		
		SHVC Vs. Simulcast		
		Y	U	V
Kimono	0	-19.2%	-11.9%	-9.5%
ParkScene	0	-10.2%	-8.5%	-8.7%
Cactus	0	-13.7%	-9.7%	-3.9%
BasketballDrive	0	-16.6%	-4.9%	-6.1%
BQTerrace	0	-7.7%	1.7%	6.2%
Average		-13.5%	-6.7%	-4.4%
Kimono	2	-27.0%	-18.9%	-16.8%
ParkScene	2	-14.8%	-12.0%	-12.1%
Cactus	2	-18.4%	-14.4%	-10.3%
BasketballDrive	2	-23.0%	-12.5%	-13.0%
BQTerrace	2	-10.2%	-2.1%	0.1%
Average		-18.7%	-12.0%	-10.4%

7.2.2.2 Results for non-aligned IRAP pictures case

SHVC supports a so-call step-wise up-switching feature by allowing IRAP pictures in the second layer occurs less frequently than IRAP pictures in the first layer. For example, IRAP pictures in the first layer occur every 2 seconds whereas IRAP pictures in the second layer occur every 4 seconds. This feature can improve the coding efficiency of SHVC while still allowing the same joining time interval (e.g. every 2 seconds) compared to simulcast.

Figure 20 illustrates the effect of step-wise up-switching when a user changes channel at the worst case scenario. In the best case scenario, the first immediate IRAP pictures in both layers after a user switches channel are aligned. In the worst case scenario, the first immediate IRAP pictures in both layers after a user switches channel are not aligned, the user would see base service for a short period of time until the next IRAP occurs in the second layer.

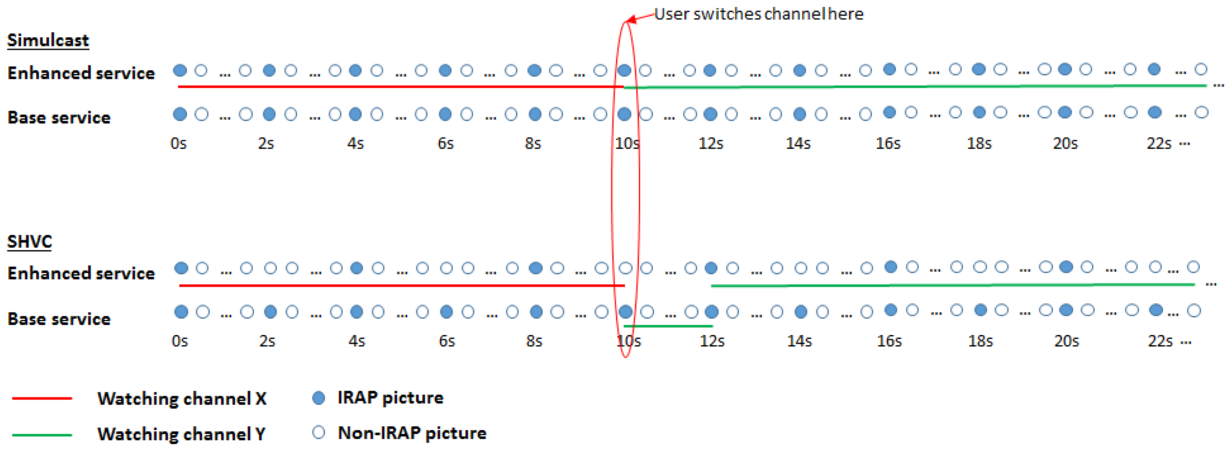

Figure 20: Effect of step-wise up-switching to channel switching

Table 40 and Table 41 show the summary of the test results. For the given test condition, SHVC provides BD-rate gains up to 40.6%. The overall average gain is 32.9% for QP delta 2.

In similar trend as reported for IRAP aligned case, for the configuration where the first layer and the second layer have spatial resolutions of 540p and 1080p, respectively, the performance of SHVC drops. The BD-rate gain is up to 26.8%. The overall average gain is 20% for QP delta 2.

Table 40: MBMS IRAP non-aligned Class B (BL 720p – EL 1080p)

Test Sequences	deltaQP	BD-Rate Comparison		
		SHVC Vs. Simulcast		
		Y	U	V
Kimono	0	-28.4%	-20.7%	-18.3%
ParkScene	0	-20.2%	-15.9%	-15.2%
Cactus	0	-25.4%	-21.4%	-14.5%
BasketballDrive	0	-26.8%	-15.8%	-16.2%
BQTerrace	0	-16.6%	3.0%	7.5%
Average		-23.5%	-14.2%	-11.3%
Kimono	2	-40.6%	-34.5%	-33.1%
ParkScene	2	-29.6%	-25.2%	-25.4%
Cactus	2	-34.3%	-32.2%	-29.0%
BasketballDrive	2	-36.4%	-30.5%	-30.1%
BQTerrace	2	-23.4%	-13.1%	-13.5%
Average		-32.9%	-27.1%	-26.2%

Table 41: MBMS IRAP non-aligned Class B (BL 540p – EL 1080p)

Test Sequences	deltaQP	BD-Rate Comparison		
		SHVC Vs. Simulcast		
		Y	U	V
Kimono	0	-18.9%	-11.0%	-8.5%
ParkScene	0	-11.7%	-9.6%	-9.0%
Cactus	0	-16.3%	-12.8%	-6.4%
BasketballDrive	0	-17.0%	-5.0%	-6.5%
BQTerrace	0	-10.1%	-0.1%	3.4%
Average		-14.8%	-7.7%	-5.4%
Kimono	2	-26.8%	-18.3%	-16.0%
ParkScene	2	-15.9%	-13.1%	-12.7%
Cactus	2	-20.6%	-17.0%	-12.5%
BasketballDrive	2	-23.4%	-12.7%	-13.3%
BQTerrace	2	-13.2%	-4.8%	-3.5%
Average		-20.0%	-13.2%	-11.6%

7.2.3 3GP-DASH

7.2.3.0 General

In this clause, simulation results for the 3GP-DASH service are provided, comparing SHVC vs HEVC simulcast. A particular value of the BD-rate decrease of SHVC comparing simulcast indicates how much less bandwidth, in percentage, is needed for transmission of the three-layer SHVC bitstream compared to transmission of all three HEVC single-layer bitstreams, on average for the same quality of the highest resolution video. The comparison indicates the difference of the bandwidth requirements for SHVC vs simulcast in the network link between the origin server and the caches/proxies.

The temporal scalability is enabled in 3GP-DASH test. There are 4 temporal sub-layers of the hierarchic-B coding structure with GOP length 8 as illustrated in Figure 21. The results of using different number of temporal sub-layers for the inter-layer prediction are provided.

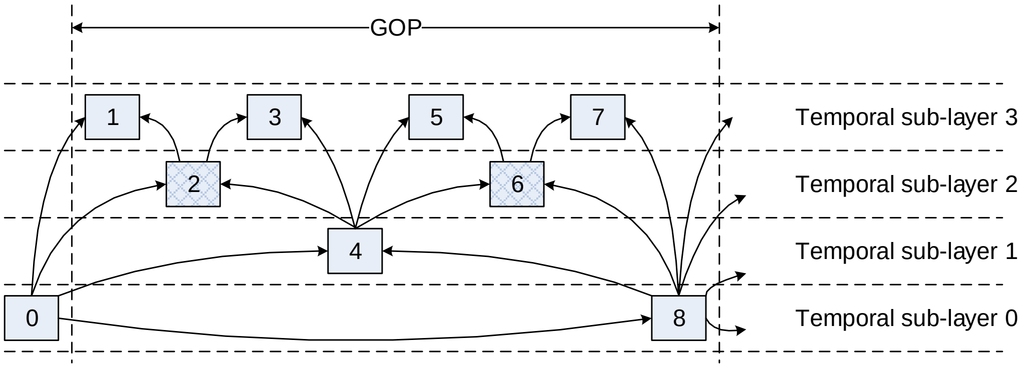

Figure 21: Temporal sub-layers of hierarchic-B coding structure

There are 2 IRAP distance options. For cross-layer IRAP aligned option, both base layer and enhancement layer share the same IRAP distance (approximately 4 seconds). For the cross-layer IRAP non-aligned option, the base layer IRAP distance is approximately 4 seconds while the enhancement layer IRAP distance is 2 seconds in order to support quick up-switch to the high quality video. Each IRAP picture type is CRA except for the first IRAP which uses IDR.

Each enhancement layer uses it immediate lower layer as its reference layer for the inter-layer prediction. E.g. the 1st enhancement layer uses base layer as its reference layer, and 2nd enhancement layer uses the 1st enhancement layer as its reference layer.

7.2.3.1 Results for aligned IRAP pictures case

BD-rate results for cross-layer RAP aligned cases are presented in Table 42 to Table 45.

The overall BD-rate decrease of SHVC comparing to simulcast is between 18.1% and 31.94% depending on the deltaQP value and number of temporal sub-layers to be used for inter-layer prediction. The max gain can be up to 41.5%.

Table 42 shows the results when temporal sub-layer 0 can be used for inter-layer prediction (ILP).

Table 42: 3GP-DASH IRAP aligned simulation results (option 2)

IRAP Aligned	SHVC vs. Simulcast			
	deltaQP	Y	U	V
Kimono	0	-24.8%	-19.0%	-19.7%
ParkScene	0	-17.3%	-14.5%	-14.5%
Cactus	0	-18.3%	-14.0%	-10.5%
BasketballDrive	0	-16.6%	-10.9%	-13.5%
BQTerrace	0	-13.5%	0.4%	10.3%
Average		-18.10%	-11.60%	-9.58%
Kimono	2	-31.8%	-27.1%	-27.8%
ParkScene	2	-24.1%	-21.9%	-22.0%
Cactus	2	-23.9%	-21.1%	-18.4%
BasketballDrive	2	-21.8%	-17.0%	-19.7%
BQTerrace	2	-19.2%	-9.9%	-4.2%
Average		-24.16%	-19.40%	-18.42%

Table 43 shows the results when temporal sub-layer 0 and 1 can be used for ILP.

Table 43: 3GP-DASH IRAP aligned simulation results (option 3)

IRAP Aligned	SHVC vs. Simulcast			
	deltaQP	Y	U	V
Kimono	0	-28.4%	-21.3%	-20.8%
ParkScene	0	-18.7%	-14.8%	-14.5%
Cactus	0	-21.5%	-16.6%	-11.0%
BasketballDrive	0	-21.1%	-12.3%	-15.1%
BQTerrace	0	-14.9%	3.7%	13.8%
Average		-20.92%	-12.26%	-9.52%
Kimono	2	-36.8%	-29.4%	-28.5%
ParkScene	2	-26.1%	-22.1%	-22.0%
Cactus	2	-28.2%	-24.3%	-20.5%
BasketballDrive	2	-27.9%	-20.1%	-22.5%
BQTerrace	2	-21.0%	-8.8%	-4.0%
Average		-28.00%	-20.94%	-19.50%

Table 44 shows the results when temporal sub-layer 0, 1 and 2 can be used for ILP.

Table 44: 3GP-DASH IRAP aligned simulation results (option 4)

IRAP Aligned	SHVC vs. Simulcast			
	deltaQP	**Y**	**U**	**V**
Kimono	0	-30.4%	-22.6%	-20.9%
ParkScene	0	-19.6%	-15.0%	-14.4%
Cactus	0	-23.8%	-18.6%	-11.1%
BasketballDrive	0	-25.9%	-13.5%	-16.2%
BQTerrace	0	-16.0%	7.0%	16.0%
Average		-23.14%	-12.54%	-9.32%
Kimono	2	-40.6%	-31.8%	-29.9%
ParkScene	2	-27.6%	-22.8%	-22.5%
Cactus	2	-31.4%	-26.9%	-22.5%
BasketballDrive	2	-34.1%	-23.4%	-25.5%
BQTerrace	2	-22.1%	-8.0%	-4.4%
Average		-31.16%	-22.58%	-20.96%

Table 45 shows the results when all pictures of the reference layers can be used for ILP.

Table 45: 3GP-DASH IRAP aligned simulation results (option 1)

IRAP Aligned	SHVC vs. Simulcast			
	deltaQP	**Y**	**U**	**V**
Kimono	0	-30.0%	-21.4%	-18.7%
ParkScene	0	-19.5%	-14.7%	-14.0%
Cactus	0	-23.8%	-18.1%	-9.7%
BasketballDrive	0	-27.5%	-11.4%	-13.7%
BQTerrace	0	-15.9%	8.4%	17.1%
Average		-23.34%	-11.44%	-7.80%
Kimono	2	-41.5%	-32.7%	-30.0%
ParkScene	2	-27.7%	-23.0%	-22.7%
Cactus	2	-32.1%	-27.6%	-22.9%
BasketballDrive	2	-36.6%	-24.2%	-25.5%
BQTerrace	2	-21.8%	-7.6%	-4.0%
Average		-31.94%	-23.02%	-21.02%

7.2.3.2 Results for non-aligned IRAP pictures case

BD-rate results for cross-layer RAP aligned cases are presented in Table 46 to Table 49.

The overall BD-rate decrease of SHVC comparing to simulcast is between 18.56% and 32.20% depending on the deltaQP value and number of temporal sub-layers to be used for inter-layer prediction. The max gain can be up to 41.2%.

Table 46: 3GP-DASH IRAP non-aligned simulation results (option 2)

IRAP Non-Aligned	SHVC vs. Simulcast			
	deltaQP	**Y**	**U**	**V**
Kimono	0	-24.9%	-18.7%	-19.3%
ParkScene	0	-18.3%	-15.1%	-15.1%
Cactus	0	-19.0%	-14.6%	-11.2%
BasketballDrive	0	-16.6%	-10.7%	-13.3%
BQTerrace	0	-14.0%	-0.4%	9.4%
Average		-18.56%	-11.90%	-9.90%
Kimono	2	-31.6%	-26.3%	-26.9%
ParkScene	2	-25.1%	-22.4%	-22.5%
Cactus	2	-25.0%	-21.8%	-19.2%
BasketballDrive	2	-21.8%	-16.8%	-19.4%
BQTerrace	2	-20.1%	-10.5%	-5.4%
Average		-24.72%	-19.56%	-18.68%

Table 47: 3GP-DASH IRAP non-aligned simulation results (option 3)

IRAP Non-Aligned	SHVC vs. Simulcast			
	deltaQP	Y	U	V
Kimono	0	-28.3%	-20.8%	-20.2%
ParkScene	0	-19.6%	-15.3%	-15.0%
Cactus	0	-22.0%	-16.9%	-11.4%
BasketballDrive	0	-21.0%	-11.9%	-14.6%
BQTerrace	0	-15.4%	3.3%	13.5%
Average		-21.26%	-12.32%	-9.54%
Kimono	2	-36.6%	-28.4%	-27.3%
ParkScene	2	-27.0%	-22.4%	-22.3%
Cactus	2	-28.9%	-24.6%	-20.9%
BasketballDrive	2	-27.8%	-19.7%	-22.1%
BQTerrace	2	-21.7%	-9.1%	-5.2%
Average		-28.40%	-20.84%	-19.56%

Table 48: 3GP-DASH IRAP non-aligned simulation results (option 4)

IRAP Non-Aligned	SHVC vs. Simulcast			
	deltaQP	Y	U	V
Kimono	0	-30.4%	-22.0%	-20.2%
ParkScene	0	-20.5%	-15.4%	-14.7%
Cactus	0	-24.2%	-18.8%	-11.4%
BasketballDrive	0	-25.8%	-13.0%	-15.7%
BQTerrace	0	-16.4%	6.6%	15.5%
Average		-23.46%	-12.52%	-9.30%
Kimono	2	-40.4%	-30.7%	-28.5%
ParkScene	2	-28.4%	-22.9%	-22.6%
Cactus	2	-31.8%	-27.1%	-22.8%
BasketballDrive	2	-34.0%	-22.8%	-24.9%
BQTerrace	2	-22.8%	-8.4%	-5.4%
Average		-31.48%	-22.38%	-20.84%

Table 49: 3GP-DASH IRAP non-aligned simulation results (option 1)

IRAP Non-Aligned	SHVC vs. Simulcast			
	deltaQP	Y	U	V
Kimono	0	-30.0%	-20.7%	-17.9%
ParkScene	0	-20.3%	-15.1%	-14.4%
Cactus	0	-24.1%	-18.3%	-9.7%
BasketballDrive	0	-27.3%	-10.7%	-12.9%
BQTerrace	0	-16.4%	8.2%	17.0%
Average		-23.62%	-11.32%	-7.58%
Kimono	2	-41.2%	-31.5%	-28.6%
ParkScene	2	-28.4%	-23.0%	-22.6%
Cactus	2	-32.5%	-27.6%	-22.9%
BasketballDrive	2	-36.4%	-23.5%	-24.7%
BQTerrace	2	-22.5%	-7.9%	-4.9%
Average		-32.20%	-22.70%	-20.74%

8 Conclusions

8.1 Introduction

The technical report provides deep technical analyses of different new use cases in the context of 3GPP multimedia services. Based on the analyses and the results, SHVC can provide technical benefits in different scenarios and circumstances. Therefore, whenever new use cases and scenarios, either those documented in the TR or new ones, are considered within emerging 3GPP services, SHVC may be an attractive candidate to be considered for such cases.

The remainder of this clause provides the conclusions of the use cases, test results and analyses for multi-stream multiparty video conferencing (MMVC), IMS based telepresence, MBMS and 3GP-DASH.

8.2 MMVC and IMS based telepresence

Both HEVC simulcast and SHVC can be used as the video codec solution in the multi-stream multiparty video conferencing (MMVC) service, in the context of the MTSI service, for the heterogeneous-device MMVC use case and the heterogeneous-bandwidth MMVC use case described in clauses 6.1.1 and 6.1.2, respectively. These use cases are currently not fully supported by TS 26.114 [4].

As shown in clause 6.1.4, using SHVC can reduce uplink bandwidth for UEs that send different versions of its video for display as the main video (as opposed to thumbnail video) by other UEs, at the cost of increased downlink bandwidth and decoding complexity for UEs that display a high resolution main video. For these use cases, uplink bandwidth reduction is an important benefit. As reported in clause 6.1.4, an average uplink bandwidth saving of about 27% can be achieved, and the corresponding reported average downlink bandwidth cost for those UEs affected was about 24%. The increase in decoding complexity for those UEs affected is roughly the percentage of the number of samples in the lower resolution video relative to that in the higher resolution video. The complexity of SHVC encoding is typically less than that of simulcast encoding (see clause 6.1.4.4).

When the previous active speaker sends both medium and thumbnail video, but uses SHVC for encoding with the thumbnail video being the base layer, additional gain for SHVC is achieved. For the scenario with 240p resolution thumbnail (i.e., 3x in both width and height relative to the 720p resolution), the average BD-rate decrease for SHVC comparing to HEVC simulcast was around 6.7%, and the BD-rate increase for SHVC comparing to HEVC single layer coding (720p) was around 13.1%. For the scenario with 360p resolution thumbnail (i.e., 2x in both width and height relative to the 720p resolution), the average BD-rate decrease for SHVC comparing to HEVC simulcast was around 18.3%, and the BD-rate increase for SHVC comparing to HEVC single layer coding (720p) was around 19.2%.

The cost/benefit of using SHVC vs. simulcast by taking into account the numbers of premium and regular UEs in an MMVC session was analysed in Clause 7.2.1.6. Three cases were considered; the conclusion for each of these three cases is summarized as follows:

- Case A: a high resolution and a medium resolution video for the current active speaker are coded using either SHVC or HEVC simulcast. For this case, if there are 4 or fewer premium UEs (not including the current active speaker) in the MMVC session, then SHVC outperforms simulcast under the assumptions for the analysis.
- Case B: a medium resolution video and a thumbnail video for the previous active speaker are coded using either SHVC or HEVC simulcast. For this case, SHVC always slightly outperforms simulcast under the assumptions for the analysis.
- Case C: a combination of Case A and Case B, where a high resolution and a medium resolution video for the current active speaker are coded using either SHVC or HEVC simulcast, and a medium resolution video and a thumbnail video for the previous active speaker are coded using either SHVC or HEVC simulcast. For this case, if there are 4.13 or fewer premium UEs (not including the current active speaker) in the MMVC session, then SHVC outperforms simulcast under the assumptions for the analysis.

The MMVC use cases and the above conclusion also apply to the IMS based telepresence service. These use cases are currently not fully supported by TR 26.923 [17].

8.3 MBMS

Both HEVC simulcast and SHVC can be used as the video codec solution in the MBMS service for the differentiated-service use case described in clause 6.2.1. This use case is currently not fully supported by TS 26.346 [3].

As shown in clause 6.2.3, using SHVC can reduce bandwidth for transmitting the encoded video streams from the content provider, through BM-SC, MBMS-GW and eNodeB, all the way to the UEs, at the cost of increased decoding complexity for UEs that consume the premium service. As reported in clause 6.2.3 Table 12 (for IRAP non-aligned Class B (BL 720p – EL 1080p)), an average bandwidth saving of 32.9% can be achieved. The increase in decoding complexity for UEs that consume the premium service is roughly the percentage of the number of samples in the lower resolution video relative to that in the higher resolution video. The complexity of SHVC encoding is typically less than that of simulcast encoding (see clause 6.2.3.3).

8.4 3GP-DASH

Both HEVC simulcast and SHVC can be used as the video codec solution in the 3GP-DASH service for the use case described in clause 6.3.1. This use case is currently not fully supported by TS 26.247 [1].

As shown in clause 6.3.3, using SHVC can reduce bandwidth for transmitting the encoded streams from the origin server to caches, at the cost of increased downlink bandwidth and decoding complexity for UEs rendering any video Representation beyond the lowest video Representation. As reported in clause 6.3.3, an average bandwidth saving of about 24% can be achieved, and the corresponding reported average downlink bandwidth cost was about 25%. The increase in decoding complexity is roughly the percentage of the sum of the number of samples in the videos of the lower Representations relative to the number of samples in the video of the highest Representation. The complexity of SHVC encoding is typically less than that of simulcast encoding (see clause 6.3.3.4).

Annex A:
Change history

Change history							
Date	**TSG #**	**TSG Doc.**	**CR**	**Rev**	**Subject/Comment**	**Old**	**New**
09-2015	69	SP-150457			Presented to TSG SA#69 (for information)		1.0.0
12-2015	70	SP-150667			Presented to TSG SA#70 (for approval)	1.0.0	2.0.0
12-2015	70				Approved at TSG SA#70	2.0.0	13.0.0

Change history							
Date	**Meeting**	**TDoc**	**CR**	**Rev**	**Cat**	**Subject/Comment**	**New version**
2017-03	75					Version for Release 14	14.0.0
2018-06	80					Version for Release 15	15.0.0
2020-07	-	-	-	-	-	Update to Rel-16 version (MCC)	**16.0.0**
2022-04	-	-	-	-	-	Update to Rel-17 version (MCC)	**17.0.0**

History

Document history		
V17.0.0	May 2022	Publication