|  |  |
| --- | --- |
| **ITU – Telecommunications Standardization Sector**STUDY GROUP 16 Question 6**Video Coding Experts Group (VCEG)**74th Meeting: 12-19 July 2024, Sapporo, JP | Document VCEG-BV02 |

|  |  |
| --- | --- |
| Question: | Q.6/SG16 (VCEG) |
| Source: | **Yan Ye (Alibaba), Marta Karczewicz (Qualcomm), Mary-Luc Champel (Xiaomi), Patrice Onno (Canon), Li Zhang (Bytedance), Xianglin Wang (Kwai), Dong Wang (OPPO), Zhuoyi Lyu (vivo), Yongkai Huo (Transsion), Ajay Luthra (Picsel Labs), Edouard François (InterDigital), Han Boon Teo (Panasonic), Yoshitaka Kidani (KDDI), Sung-Chang Lim (ETRI)** | Tel: Email: | yan.ye@alibaba-inc.commartak@qti.qualcomm.comchampelmaryluc@xiaomi.comPatrice.Onno@crf.canon.frlizhang.idm@bytedance.comxianglinwang@kwai.comwangdong7@oppo.comzhuoyi.lv@vivo.comyongkai.huo@transsion.comajay@picsellabs.comEdouard.Francois@interdigital.comhanboon.teo@sg.panasonic.comyo-kidani@kddi.comsclim@eti.re.kr |
| Title: | **Proposed timeline and requirements for the next generation video coding standard** |
| Purpose: | Proposal |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

VVC, the latest codec standard jointly developed by ITU-T and MPEG, was finalized in July 2020. Since then, JVET has been actively exploring coding technologies that can achieve higher compression performance compared to VVC. The latest enhanced compression model (ECM) version 13 shows that more than 25% of bitrate savings over VVC in terms of luma PSNR can be achieved in the random access configuration. Two rounds of subjective evaluations have shown that the ECM also provides subjective benefits. Considering regular standard development cycles and the current video industry landscape, this contribution proposes timeline for the development of the next generation video codec standard. This contribution also proposes a set of requirements for this next generation codec standard.

Mirror documents of this contribution have also been submitted to ISO/IEC SC 29/WG02 MPEG technical requirements as m68403 and JVET as JVET-AI0247.

# Introduction

JVET has been exploring next generation video coding technologies beyond the capability of VVC since 2021. The exploration activities are primarily conducted in the form of exploration experiments (EE) and Ad hoc Groups (AhG) following two main directions: enhanced compression model (ECM) and neural network video coding (NNVC). Both directions have been consistently delivering improved coding efficiency over VVC. Especially, for the ECM direction, Figure 1 plots the evolution of ECM’s performance (BD rate savings in terms of luma PSNR) in the random access (RA) configuration of the ECM CTC [1]. As shown, since ECM-2.0, ECM’s performance has been steadily increasing for every version of ECM. After integration of all tools adopted at the 34th JVET meeting, the latest ECM-13.0 achieves average coding performance gain of 25.5% for camera-captured content and 31-40% gain for screen content (class F and class TGM) in terms of luma PSNR in the RA configuration. Table 1 shows the detailed per-class performance of ECM-13.0.

****

Figure 1. ECM's compression performance vs. VVC in terms of luma PSNR in the RA configuration

Table 1 Coding performance in RA configuration of the ECM CTC

|  |  |
| --- | --- |
|  | Random Access Main 10 |
|  | Y | U | V | EncT | DecT |
| Class A1 | -26.04% | -21.59% | -33.88% | 946.3% | 892.5% |
| Class A2 | -29.16% | -31.81% | -36.71% | 887.1% | 1036.9% |
| Class B | -23.43% | -28.98% | -26.76% | 747.6% | 846.1% |
| Class C | -24.99% | -20.48% | -21.31% | 821.5% | 922.3% |
| Class E |  |  |  |  |  |
| Overall  | **-25.51%** | **-25.80%** | **-28.72%** | **831.6%** | **911.4%** |
| Class D | -25.69% | -21.07% | -22.42% | 785.7% | 1008.7% |
| Class F | -31.65% | -34.02% | -34.69% | 660.1% | 557.6% |
| Class TGM | -40.63% | -45.97% | -46.06% | 634.7% | 487.9% |

Two rounds of subjective evaluations had been performed to assess the subjective quality benefits that the ECM provides. The most recent round of assessments was performed in April 2024 using ECM-12.0 in the form of expert viewings and subjective viewing using naïve subjects [2]. Using the results from expert viewing in Rennes, Table 2 compares the BD-rate savings in terms of objective quality (PSNR) and subjective quality (MOS). As can be seen, except for a couple of outliers including one very fast-moving gaming sequence that is difficult to watch and rate, most of the sequences behave as expected, with MOS-based BD-rate numbers exceeding the PSNR-based BD-rate numbers. The formal subjective testing conducted in Rome using naïve subjects shows more gains than expert viewing, although BD-rate results are not readily available and therefore not included here.

Table 2 PSNR-based and MOS-based BD-rate savings of UHD and HD content

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **PSNR-based** | **MOS-based** |
| **UHD**  | **BeachMountain2** | -20.0% | -12.7% |
| **CrossRoad2s500** | -19.0% | -25.3% |
| **DrivingPOV3** | -24.4% | -28.8% |
| **Average** | **-22.2%** | **-22.3%** |
| **HD** | **Beatriz** | -14.8% | -21.5% |
| **DOTA2s360** | -12.2% | 1.9% |
| **GTAVs090** | -16.9% | -18.2% |
| **Average** | **-14.9%** | **-12.6%** |

JVET held a joint session with AG5 MPEG visual quality assessment, WG2 MPEG technical requirements, and ITU-T VCEG in Rennes to discuss the subjective results in [2]. As a result of the joint discussion, WG2 issued a recommendation in Section 2.1.2 of its output document no. 355 “Recommendations of the 15th MPEG WG2 meeting (MPEG 146)” [3] as follows:

*WG2 recommends MPEG members to submit use cases and requirements in the area of future video standardization beyond VVC.*

Fourteen companies active in JVET are submitting this contribution in response to this recommendation from WG2. Mirror documents are also submitted to ISO/IEC SC 29/WG02 MPEG technical requirements as m68403 and JVET as JVET-AI0247.

# Proposed timeline for the next generation codec standard

The following timeline is proposed:

* Call for Proposals (CfP) to be issued: April or July, 2025
* Responses to the Call to be evaluated: January or April, 2026
* Version 1 finalization: July or October, 2028

Rationales for this proposed timeline are provided below:

* ECM provides significant evidence that beyond-VVC-capability coding technologies exist. To us, this means that the Call for Evidence (CfE) stage could be considered optional, and therefore the proposed timeline does not include the CfE stage;
* Sufficient time, e.g. at least 6 months, should be given to proponents to prepare for their responses to the CfP. The proposed timeline provides at least 6 months of preparation time;
* The proposed finalization would create an 8-year gap between version 1 of VVC and that of the next generation standard. Considering the gap between AVC and HEVC was ~10 years and the gap between HEVC and VVC was ~7 years, we assert that the proposed timeline is in line with our usual practice;
* ITU-T VCEG and ISO/IEC MPEG are not the only “player in town” when it comes to developing video codec standards. The proposed timeline is aimed at maintaining an appropriate cycle such that the joint standards of ITU-T VCEG and ISO/IEC MPEG remain competitive in the marketplace in terms of providing superior compression performance and fulfilling industry’s needs.

# Proposed requirements

In this section, a set of requirements are proposed for the next generation video codec standard. The requirements are modified based on those defined for the VVC standard [4][5].

## Applications

*Requirement:*

* The next generation codec standard should be capable of serving the needs of numerous applications, including:
	+ Mobile communication and entertainment.
	+ On-demand streaming, download-and-play, and storage-media-based applications.
	+ Live streaming.
	+ Low-delay interactive communication, e.g. for remote collaboration.
	+ Ultra-low delay communication and entertainment, e.g. clouding gaming or autonomous driving.
	+ Broadcast.
	+ Digital cinema and large-screen digital imagery.
	+ Immersive applications, such as Virtual, augmented and mixed reality for communication, interaction and entertainment.
	+ Surveillance and smart home.
	+ Video editing.
	+ User-generated content.
	+ Film grain content.
	+ Screen content.
	+ Gaming content.
	+ AI-generated content.
	+ Stereoscopic and multi-view content.
	+ Machine analysis of video content.
	+ Internet of Things, industrial applications, automotive, robotics, and digital medicine.

*Remark:*

It is suggested that special care be taken to provide a good solution to film grain content.

## Compression Capability

*Requirement***:**

* The next generation codec standard should be capable of providing a bit rate reduction of at least 40% at similar subjective quality compared to VVC Main 10 for 4K and higher resolution video.
* The top priority is high coding efficiency for 4:2:0 8-bit and 10-bit video content at resolutions, frame rates and bit rates that are widely used in consumer applications.

*Rationale***:**

* Given the continued rapid expansion of video traffic volume, we consider a 40% rate reduction target to be sufficiently meaningful for a vast number of video use cases and applications in terms of bandwidth and storage cost reduction. Further, the 40% reduction target should be mainly targeted for 4K or higher resolutions as the main use cases and applications continue to migrate to higher resolutions. We expect other resolutions to also benefit from the increased compression capability of the next generation codec standard.
* Most of today’s consumer applications around the world rely on 8-bit and 10-bit video. Therefore, we propose that equal consideration be given to both formats.

## Complexity

Complexity refers to computational resource consumption (in terms of battery drainage, power consumption, computing cycles, memory capacity, memory bandwidth, etc.) based on typical computing architectures and parallelization mechanisms.

*Requirements:*

* The next generation codec’s complexity shall allow for feasible implementation at the expected time of usage.
* Real-time decoding shall be feasible at the expected time of usage.
* The encoder complexity should be capable of trading off complexity and coding efficiency. Real-time encoding with adequate coding efficiency advantage over existing standards shall be feasible at the expected time of usage. Non-real-time encoding with further improvement of coding efficiency should be feasible at the expected time of usage.

## Bit depth and colour sampling

*Requirements:*

* Commonly applied functionalities and formats that can be readily supported without major architectural modifications should be supported in the core coding design of v1, including the following:
	+ 4:2:0 colour sampling and monochrome
	+ Bit depths of 8 to 10
	+ Arbitrary picture size and frame rates (incl. variable frame rate)
	+ Variable picture size within a coded video sequence
	+ Panoramic projection
	+ Stereoscopic 3D, multi-view, depth, alpha, and auxiliary pictures
* The next generation codec standard shall be capable of representing video signals with bit depths of 8 and 10 in v1, and should be capable of representing video signals with bit depths ranging from 8 to 16.
* The next generation codec standard shall be capable of representing video signals with 4:2:0 and monochrome colour sampling in v1, and should be capable of representing video signals with colour samplings ranging up to 4:4:4.

*Rationale***:**

* Given the recent rapid development of virtual reality applications, we propose to give stereoscopic 3D and multi-view content and information closely associated with these applications, such as depth and alpha maps, a bigger emphasis in v1 of the next generation video codec standard.

## Resolutions, frame rates, etc.

*Requirements:*

* The next generation codec standard shall be capable of representing pictures and video signals with spatial resolution up to 8Kx4K, and should be capable of representing pictures and video signals with spatial resolution larger than 8Kx4K.
* The next generation codec standard shall be capable of representing video signals with spatial resolutions with a square or portrait aspect ratio since such formats are already widely used in popular smartphone applications and wearable devices such as smartwatches.
* The next generation codec standard shall be capable of representing video signals with temporal resolutions ranging up to 240 fps, and should be capable of representing video signals with temporal resolutions higher than 240 fps.
* The next generation codec standard shall be capable of representing still pictures.
* The next generation codec standard shall be capable of representing pictures with wide colour gamut and high dynamic range (e.g., ITU-R BT.2100) as well as conventional colour formats (e.g., ITU-R BT.709).
* The next generation codec standard shall be capable of coding progressively scanned video signals, and may support the coding of other scanning formats.

*Rationale:*

* The maximum “mandatory” frame rate in the VVC requirements document [4][5] was set to 120 fps. Today, smartphone cameras are already capable of capturing video up to 240fps. In addition, some video services allow upload of UGC content up to 240fps. While slow-motion capture can be used to reach even higher frame rates, at the moment it seems 240fps is the maximum camera sensors natively support. Therefore, it is proposed to make 240fps mandatory for the next generation video codec.

## Stereoscopic 3D and multi-view

*Requirement:*

* Version 1 of the next generation codec standard shall support stereoscopic 3D content, and should support multi-view content.

## End-to-end delay

Encoder processing time is the time distance used by an encoder starting from the input of the first video sample and ending with the output of the last bit representing that picture.

Decoder processing time is the time distance used by a decoder starting from the input of the first bit and ending with the output of the last video sample representing that picture.

Low end-to-end delay is a minimized time distance that is equal to the sum of the minimum encoder processing time and minimum decoder processing time, while representing the video signal at a desired coding efficiency.

Ultra-low end-to-end delay is a minimized time distance that is shorter than the sum of the minimum encoder processing time and minimum decoder processing time, while representing the video signal at a desired coding efficiency.

*Requirements:*

* The next generation codec standard shall be capable of low end-to-end delay operation, efficiently enabling interactive and conversational applications.
* The next generation codec standard should be capable of ultra-low end-to-end delay operation, efficiently enabling applications such as cloud gaming and virtual reality.

## Random access and "Trick mode" support

*Requirements***:**

* The next generation codec standard shall have support for random access points in the video bitstream for functionality such as channel switching and program chapter access.
* The next generation codec standard shall have support for pause, fast forward, normal speed reverse, and fast reverse access to a stored video bitstream.
* The next generation codec standard shall have support for spatial random access in the video bitstream for functionality such as extraction, transmission, and decoding of only one or more rectangular regions of the pictures, for efficiently enabling applications such as virtual reality, 360o video streaming, and region-of-interest stream adaptation.

## Support of additional channels

*Further discussion:*

* More and more devices are already able to capture RGBa rather than just RGB in order to enable use cases with transparency such as video overlay. Similarly, many applications are using RGBd format where depth can be either captured or generated, e.g. via AI-based algorithms. Existing codecs use auxiliary pictures to code these additional video channels. It is suggested that the next generation codec standard should be broadened to a “4-channel” concept so as to code alpha or depth together with the YUV channels in a more coherent manner.
* For some applications, the alpha information would need to be coded with different parameters (for instance, with a lower maximum QP or even with lossless coding) than those used to code the video channels. This is a capability that the next generation codec standard should offer.
* The next generation codec standard shall support bit depth of alpha and depth information from 1 bit to 16 bits and shall offer the possibility to code different dynamic ranges of alpha/depth with different precisions.

## Other requirements

Other requirements outlined in [4][5], including packet loss robustness, scalable extensions, could be further developed in a future version of the requirements document.

# References

1. M. Karczewicz, Y. Ye, “Common test conditions and evaluation procedures for enhanced compression tool testing “, JVET 34th meeting, Hannover, Germany, October 2023, Doc. JVET-AF2017 (available at <https://jvet-experts.org/doc_end_user/documents/32_Hannover/wg11/JVET-AF2017-v1.zip>).
2. M. Wien, V. Baroncini, G. Baroncini, “Preliminary report on subjective performance evaluation of the ECM”, JVET 34th meeting, Rennes, France, April 2024, Doc. JVET-AH0344 (available at <https://jvet-experts.org/doc_end_user/documents/34_Rennes/wg11/JVET-AH0344-v1.zip>).
3. Recommendations of the 15th MPEG WG2 meeting (MPEG 146), Doc. N00355 of ISO/IEC JTC1/SC29/WG 2 MPEG technical requirements, April 2024.
4. “Requirements for Future Video Coding (FVC)”, ITU-T SG16/Q6 VCEG 56th meeting, Torino, Italy, July 2017, Doc. VCEG-BD03 (available at <http://wftp3.itu.int/av-arch/video-site/1707_Tor/>).
5. “Requirements for a Future Video Coding Standard v5”, ISO/IEC JTC1/SC29/WG11 MPEG, 119th meeting, Torino, Italy, July 2017, Doc. N17074 (available at <https://dms.mpeg.expert/doc_end_user/documents/119_Torino/wg11/w17003.zip>).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_