Unified Communication Specification for H.264/MPEG-4 Part 10 AVC and SVC Modes Version 1.0b

This document is now managed by IMTC. Please note that any change that affects backwards compatibility requires a vote of approval by not less than two-thirds of its members.

The IPR disclosure links have been updated to point to the IMTC web site.

Version 1.0 – December 21, 2012

IPR Disclosures

In accordance with the UCI Forum IPR Policy, the following IPR disclosures have been filed:

Ericsson	http://portal.imtc.org/DesktopModules/Inventures_Document/FileDownload.aspx?ContentID=21762
Polycom, Inc.	http://portal.imtc.org/DesktopModules/Inventures_Document/FileDownload.aspx?ContentID=21763
Microsoft Corp.	http://portal.imtc.org/DesktopModules/Inventures_Document/FileDownload.aspx?ContentID=21764
Vidyo, Inc.	http://portal.imtc.org/DesktopModules/Inventures_Document/FileDownload.aspx?ContentID=21765

Table of Contents

1. Introduction

This document contains a specification for H.264/MPEG-4 Part 10 Advanced Video Coding (AVC) Modes for real-time point-to-point and multipoint conferencing products.

The goal of this specification is to support the use of AVC and SVC video in the entire gamut of unified communications applications that use video. Targeted application scenarios thus include low-end mobile phone video chat, all the way to high-end multi-monitor telepresence systems.

This specification assumes that the reader is familiar with the H.264/MPEG-4 Part 10 AVC standard specification and its Scalable Video Coding (SVC) extension specified in Annex G. In the following, the term 'AVC' refers to the H.264/MPEG-4 Part 10 specification excluding the scalability features of Annex G, whereas 'SVC' refers specifically to systems that use the scalability features.

2. Normative references

[1] ITU-T Rec. H.264 | ISO/IEC 14496-10 Advanced video coding for generic audiovisual services. The standard is available at http://www.itu.int/rec/T-REC-H.264. Unless otherwise specified, this document refers to the edition approved by ITU-T in January 2012 (posted at the ITU-T web site link above). Annex G of this specification contains the SVC extension.

[2] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119, March 1997.

3. Terminology

3.1. Abbreviations

For the purposes of this specification, the following abbreviations apply (those with an asterisk '*' are copied from the H.264/MPEG-4 Part 10 specification):

ASO	Arbitrary Slice Ordering
CABAC*	Context-based Adaptive Binary Arithmetic Coding
CAVLC*	Context-based Adaptive Variable Length Coding
FMO	Flexible Macroblock Ordering
IDR*	Instantaneous Decoding Refresh
MB*	Macroblock
NAL*	Network Abstraction Layer
POC	Picture Order Count
PPS	Picture Parameter Set
QP	Quantization Parameter
SEI*	Supplemental Enhancement Information
SPS	Sequence Parameter Set

SSPS	Subset Sequence Parameter Set
SVC*	Scalable Video Coding
UC	Unified Communications
VCL*	Video Coding Layer

3.2. Definition

For the purposes of this specification, the following definitions apply (those with an asterisk '*' are copied from the H.264/MPEG-4 Part 10 specification):

Bitstream*	A sequence of bits comprising NAL units that forms the representation of coded pictures and associated data forming one or more coded video sequences.
DID	dependency_id as defined in Annex G of the H.264/MPEG-4 Part 10 specification.
NAL unit	the basic encapsulation structure in H.264/MPEG-4 Part 10; a syntax structure containing an indication of the type of data to follow, followed by that data in the form of a byte sequence interspersed as necessary with emulation prevention bytes
PPSID	picture_parameter_set_id as defined in the H.264/MPEG-4 Part 10 specification.
PRID	priority_id as defined in Annex G of the H.264/MPEG-4 Part 10 specification.
QID	quality_id as defined in Annex G the H.264/MPEG-4 Part 10 specification.
Reference frame	a frame that may be used for inter prediction in the decoding process of subsequent frame(s) in decoding order.
SPSID	seq_parameter_set_id as defined in the H.264/MPEG-4 Part 10 specification.
SVC base layer	designates the bitstream in which all VLC NAL units with dependency_id or quality_id greater than zero are removed.
TID	temporal_id as defined in Annex G of the H.264/MPEG-4 Part 10 specification.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

4. General Properties

Encoders and decoders that conform to a particular UC Mode must meet the constraints defined in this specification. Mode and capability negotiation between encoders and decoders is assumed to be available, but it is outside the scope of this specification.

4.1. Mode Structure

Five UC modes are defined in this specification. They include AVC single layer, SVC with temporal scalability, SVC with temporal and quality scalability[1], SVC with temporal and spatial scalability, and SVC with all of temporal, quality, and spatial scalability. The intention of including modes with incremental scalability capabilities is to allow encoder chip and device manufacturers to gradually incorporate the necessary support into their devices.

The UC Modes are as follows.

- UC Mode 0: Non-scalable single layer AVC bitstream.
- UC Mode 1: SVC with temporal scalability using hierarchical P pictures.
- UC Mode 2q: SVC with temporal scalability using hierarchical P pictures and quality scalability.
- UC Mode 2s: SVC with temporal scalability using hierarchical P pictures and spatial scalability.
- UC Mode 3: SVC scalability with all of temporal scalability (using hierarchical P pictures), quality, and spatial scalability.

Encoders that conform to higher level modes shall include the capabilities of encoding bitstreams associated with lower level modes. For example, encoders that conform to UC Mode 2q must be able to generate a single layer AVC stream, i.e., UC Mode 0.

> NOTE: Change of UC Mode may be requested through signaling means that are outside the scope of this specification.

Figure 1 depicts the hierarchical structure and relationship for all UC Modes. Details of each UC Mode are elaborated in Section 5.

Encoders and decoders which can only operate at UC Mode 0 will not be subject to certification by the UCI Forum. Encoders and decoders will be tested, however, on their ability to encode and decode, respectively, UC Mode 0 streams.

[1] Quality scalability is also known as SNR scalability.

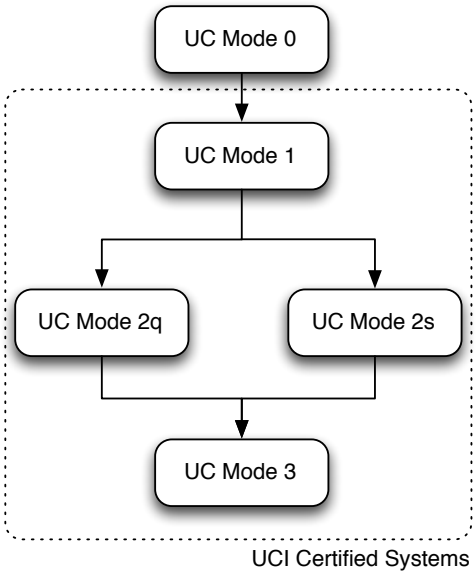

Figure 1. The hierarchical relationship of UC Modes

The UC Modes indicate properties of the bitstream that will be produced by an encoder or transmitted from a server without taking into account any adaptations that may happen dynamically at either the sender or during transport to the receiver. For example, a sender may elect to eliminate a layer to accommodate reduced available bit rate. Such adaptations are outside the scope of this specification.

4.2. Profiles

In order to facilitate interoperability among a large variety of UC systems, this specification includes both mandatory and optional profiles for video encoders. It is assumed that newer systems are capable of capability negotiation, and thus will enable the use of the optional and more sophisticated profiles.

Encoders and decoders conforming to this specification must support the Constrained Baseline profile for non-scalable bitstreams. Encoders and decoders may also support the Constrained High profile for non-scalable bitstreams, and it is recommended that they do so. This requirement refers to UC Mode 0 and – due to the hierarchical construction of UC Modes – to all five UC Modes.

Encoders conforming to UC Mode 0 may include NAL units of type 14 (referred to as prefix NAL units).

Encoders and decoders conforming to one of the scalable UC Modes (1, 2s, 2q, or 3) must support the Scalable Constrained Baseline profile. Encoders and decoders conforming to one of the scalable UC Modes may also support the Scalable Constrained High profile, and it is recommended that they do so.

> NOTE: Since decoders have to conform to one of the H.264/MPEG-4 Part 10 profiles indicated above they may have to support additional layering structures to those specified in this document (e.g., with additional temporal layers).

4.3. Resolutions and Frame Rates

Encoders and decoders that conform to this specification must be able to encode a bitstream with parameters as specified in this section.

NOTE: Since decoders have to conform to one of the H.264/MPEG-4 Part 10 profiles indicated in Section 4.1, they may have to support additional parameters to those specified in this document (e.g., additional spatial resolutions or temporal frame rates).

4.3.1. Spatial Resolutions

Encoders and decoders conforming to this specification must be able to encode video sequences using all of the following spatial resolutions (in pixels):

1280x720, 960x540, 848x480, 640x360, 480x270, 424x240, 320x180

720x1280, 540x960, 480x848, 360x640, 270x480, 240x424, 180x320

Note that these resolutions have 16:9 and 9:16 aspect ratios.

NOTE: When spatial scalability is supported, the vertical and horizontal resolution ratios between successive layers must be 1.5 or 2 as specified in [1].

NOTE: An encoder may have to apply macroblock-aligned cropping in some of the resolutions listed above when upscaling/downscaling the spatial resolutions to use 1.5 scalability, in order to maintain the picture aspect ratio. For example, when creating a lower spatial layer from 1280x720 that uses 1.5 scalability, the downsampled signal would have an 853x480 resolution. In order to covert to 848x480, 5 columns would have to be removed prior to encoding the lower spatial resolution.

NOTE: Only progressive video is allowed in the AVC and SVC profiles used in this specification.

A square sample aspect ratio (1:1) is required in this specification. Other sample aspect ratios may be supported in the future.

NOTE: In case of a change from portrait to landscape orientation, or vice versa, where an encoder modifies the macroblock scanning order, a new Sequence Parameter Set (SPS) must be present in the bitstream to indicate the changed picture size parameters.

4.3.2. Temporal Resolutions

Encoders and decoders conforming to this specification must be able to encode and decode video at 30 frames per second.

When using temporal scalability, encoders that conform to this specification must generate bitstreams with dyadic frame rates (i.e., the ratio of the frame rates between any two temporal layers is a power of 2).

Encoders should generate bitstreams with constant frame rates.

NOTE: When there is adaptation, for example due to lighting level variations or packet losses, the frame rate may not end up being constant, or even dyadic (except when an entire layer is eliminated). The ratio of frames rates among temporal layers that is signaled, however, must be dyadic.

4.4. Levels

Encoders and decoders conforming to this specification must be able to operate at level 3.1, with the bit rate limited to 4 Mbps.

NOTE: Encoders may operate at higher bit rates, depending on the capabilities negotiation with corresponding decoders.

Note that the resolutions, frame rates, and levels listed above are a minimum set of requirements. Encoders that conform to this specification may support other resolutions and/or frame rates, or a higher level value than 3.1 as long as the constraints specified in Section 4.3 are fulfilled. For a particular bitstream, encoders that conform to this specification shall use a level value that best describes the bitstream.

5. UC Mode Definitions and Capabilities

5.1. UC Mode 0: AVC

This mode addresses the configuration of encoders that do not use SVC. Such encoders typically use a flat temporal picture coding structure with only one temporal layer. This structure is shown in **Figure 2**.

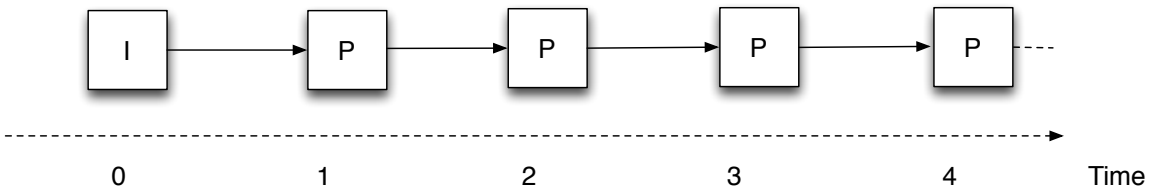

Figure 2. AVC single layer mode

In this figure, the arrows indicate prediction, and the letters 'I' and 'P' correspond to I and P pictures, respectively.

AVC allows the implementation of temporal scalability. Prefix NAL units (type 14) , however, which identify the temporal layer each slice belongs to, are defined in the SVC extension.

Mode 0 bitstreams may contain type 14 NAL units. This enables the direct conversion of UC Mode 1 through 3 streams to UC Mode 0 without re-encoding. The prefix NAL units shall be discarded by legacy AVC decoders that are not SVC-compatible (as prescribed by the AVC specification), so that the bitstream can still be decoded.

Encoders and decoders which can only operate at UC Mode 0 will not be subject to certification by the UCI Forum. UC Mode 1 and higher systems, however, shall be able to encode and decode UC Mode 0 bitstreams.

Table 1 provides an example bitstream structure for a UC Mode 0 stream with a single temporal layer, assuming a 360p 30fps resolution. Even-numbered access units are shown in shaded cells.

Table 1. Example UC Mode 0 Bitstream Structure for Stream 360p 30fps

NAL unit (type)	Relevant fields in the NAL	Description
SPS (7)	SPSID = 0	SPS of stream 360p 30Hz
PPS (8)	PPSID = 0, SPSID = 0	PPS of stream 360p 30Hz
Prefix (14)	TID = 0	Prefix NAL of stream 360p 30Hz
IDR slice (5)	PPSID = 0, POC = 0	Base layer IDR slice(s) in stream 360p 30Hz
Prefix (14)	TID = 0	Prefix NAL of stream 360p 30Hz
Non-IDR slice (1)	PPSID = 0, POC = 1	P slice(s) of stream 360p 30Hz
Prefix (14)	TID = 0	Prefix NAL of stream 360p 30Hz
Non-IDR slice (1)	PPSID = 0, POC = 2	P slice(s) of stream 360p 30Hz
Prefix (14)	TID = 0	Prefix NAL of stream 360p 30Hz
Non-IDR slice (1)	PPSID = 0, POC = 3	P slice(s) of stream 360p 30Hz
…	…	…

The maximum macroblock processing rate (MB/sec) for this stream is 27,600. This number must be smaller than or equal to the maximum macroblocks per second supported by the encoder.

5.2. UC Mode 1: SVC with Temporal Scalability

This mode addresses the configuration of encoders that use SVC with only temporal scalability. Encoders conforming to this mode may generate a minimum of two and a maximum of four temporal layers.

The hierarchical P prediction structure is used to achieve temporal scalability, with dyadic ratios of frame rates (i.e., the ratio of the frame rates between any two layers is a power of two). The frames in the enhancement layer(s) only use the immediately previous reconstructed frame in a lower layer as the reference frame. In support of this mode, each coded slice NAL unit (type 1 or 5) shall be immediately preceded by a prefix NAL unit (type 14).

Figures 3 through 5 depict the temporal picture coding structure for two, three, and four layer temporal scalability. The pictures are shown offset in the vertical dimension to indicate their association with a different temporal layer. Each picture is identified by a picture type letter ('I' or 'P') and a number indicating its temporal layer (0 through 3, depending on the number of layers available).

Figure 3 shows an example for the two-layer mode. If the maximum frame rate of the source is 30 fps, then layer 0, or the base layer, consists of pictures I0 and P0 and has a frame rate of 15 fps. Layer 1, or the enhancement layer has a frame rate of 15fps as well, and consists of pictures P1. Decoding of both layers 0 and 1 results in 30 fps.

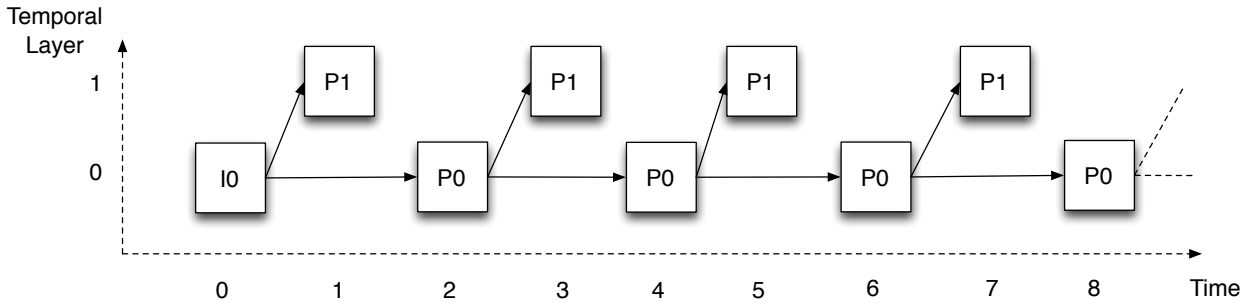

Figure 3. Two-layer temporal picture coding structure

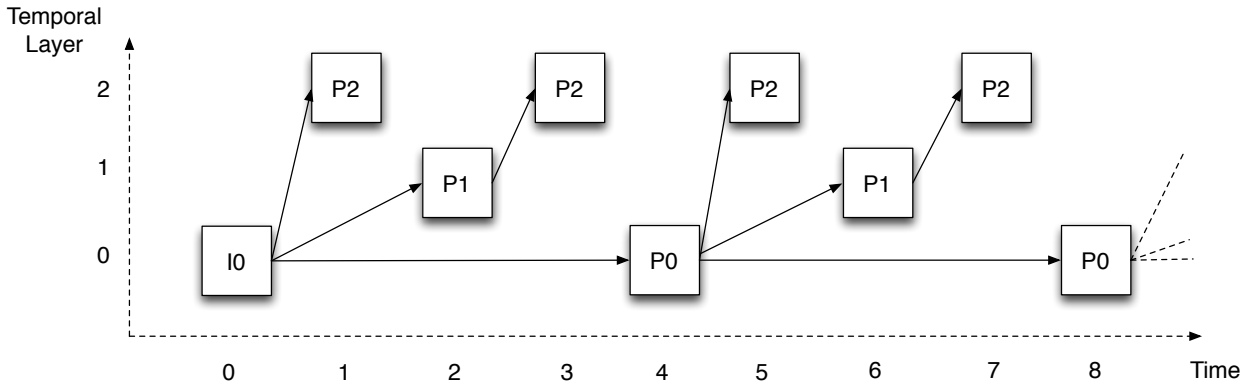

Figure 4. Three-layer temporal picture coding structure

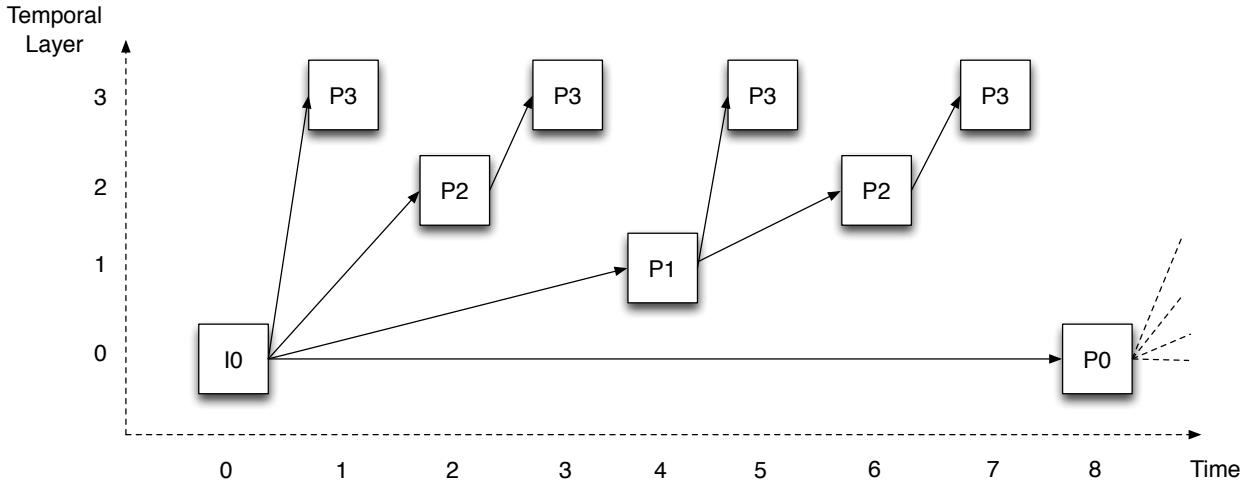

Figure 5. Four-layer temporal picture coding structure

Figure 4 shows an example of the three-layer mode. If the maximum frame rate of the source is 30 fps, then layer 0 consisting of pictures I0 and P0 has a frame rate of 7.5 fps, layer 1 consisting of pictures P1 has a frame rate of 7.5 fps, and layer 2 consisting of pictures P2 has a frame rate of 15 fps.

Finally, **Figure 5** shows an example of the four-layer mode. If the maximum frame rate of the source is 60 fps, then layer 0, has a frame rate of 7.5 fps. Layer 1, or the first temporal enhancement layer, has a frame rate of 7.5 fps as well. Decoding of layer 0 and layer 1 results in a video sequence at 15 fps. Layer 2, or the second temporal enhancement layer, has a frame rate of 15 fps. Decoding of layer 0, layer1, and layer 2 results in a video sequence at 30 fps. Finally, layer 3, or the third enhancement layer, has a frame rate of 30 fps. Decoding of layer 0, layer 1, layer 2 and layer 3 results in a video sequence at 60 fps.

UC Mode 1 bitstreams shall not use coded slice extension NAL units (type 20).

> NOTE: The coded slice extension NAL units (type 20) are only recognized by SVC decoders. While temporal scalability can be realized by such NAL units, this specification disallows the use of coded slice extension NAL units in UC Mode 1 so that AVC decoders can decode bitstreams at the highest possible frame rate.

For each prefix NAL, the value of temporal_id (TID) specifies the hierarchical dependency of a temporal layer relative to other layers, with 0 representing the base temporal layer, 1 the first temporal scalable layer, 2 the second temporal scalable layer, and so forth. The values of dependency_id (DID) and quality_id (QID) must be equal to 0. The values of no_inter_layer_pred_flag, discardable_flag and output_flag must be 1. The value of use_ref_base_pic_flag must be 0. The decoding order must be the same as the display order. The prefix NAL units shall be discarded by legacy AVC decoders that are not SVC-compatible, so that the bitstream can still be decoded.

Encoders that conform to this UC Mode must be able to generate bitstreams with at least two temporal layers.

> NOTE: From an encoder's perspective the differences between this mode and Mode 0 are: (1) the way the reference frame is used for P frames in support of hierarchical P structure; and (2) additional memory buffers for enhancement layers when three or four temporally scalable layers are supported.

Table 2. Example UC Mode 1 Bitstream Structure for Stream 360p 30fps

NAL unit (type)	Relevant fields in the NAL	Description
SPS (7)	SPSID = 0	SPS of stream 360p 30Hz
PPS (8)	PPSID = 0, SPSID= 0	PPS of stream 360p 30Hz
Prefix (14)	TID = 0	Prefix NAL of stream 360p 7.5Hz
IDR slice (5)	PPSID = 0, POC = 0	Base layer IDR slice(s) in stream 360p 7.5Hz
Prefix (14)	TID=2	Prefix NAL of stream 360p 30Hz
Non-IDR slice (1)	PPSID = 0, POC = 1	P slice(s) of stream 360p 30Hz
Prefix (14)	TID=1	Prefix NAL of stream 360p 15Hz
Non-IDR slice (1)	PPSID = 0, POC = 2	P slice(s) of stream 360p 15Hz
Prefix (14)	TID=2	Prefix NAL of stream 360p 30Hz
Non-IDR slice (1)	PPSID = 0, POC = 3	P slice(s) of stream 360p 30Hz
Prefix (14)	TID=0	Prefix NAL of stream 360p 7.5Hz
Non-IDR slice (1)	PPSID = 0, POC = 4	P slice(s) of stream 360p 7.5Hz
…	…	…

Encoders conforming to UC Mode 1 must be able to generate bitstreams in UC Mode 0 in addition to UC Mode 1.

Table 2 provides an example bitstream structure for a UC Mode 1 stream, assuming a 360p 30fps resolution. Even-numbered access units are shown in shaded cells.

The possible combinations of number of scalable layers, maximum frame rates, and resolutions are limited by the maximum macroblock processing rate (MB/sec) defined as a part of the hardware capabilities. The sum of macroblocks per second in the highest temporal scale must not exceed this limit. For example, Table 3 shows the macroblock processing rates for each layer in the above example.

Table 3. Macroblock Processing Rate Example

Temporal ID	MB/sec
0	6,900
1	6,900
2	13,800
Total:	27,600

In this example, the sum of macroblocks per second is 26,400 MB/sec. This number must be smaller than or equal to the maximum macroblocks per second supported by the encoder.

5.3. UC Mode 2q: SVC with Temporal and Quality Scalability

This mode addresses the configuration of encoders that use SVC with temporal and quality scalability. Encoders conforming to this mode may generate any combination of one to four temporal layers, with quality enhancement layers as specified below for each type of quality scalability. Quality scalability is applied to all the pictures to create two or more quality layers.

Three quality scalability modes are supported:

- Coarse-grain scalability (CGS) non-rewrite mode only with up to two enhancement layers,
- CGS rewrite mode with up to two enhancement layers, and
- Medium-grain scalability (MGS) with up to four enhancement layers.

NOTE: The rewrite mode is defined in the H.264/MPEG-4 Part 10 SVC specification. The encoded SVC SNR scalability bitstream utilizing the rewrite mode can be converted to an AVC bitstream without fully decoding the SVC bitstream.

Only one of CGS non-rewrite mode, CGS rewrite mode, or MGS may be used at the same time in the same coded video sequence.

Encoders conforming to this mode must be able to generate bitstreams with at least one quality enhancement layer and two temporal layers. The number of temporal layers follows the constraints

specified in Section 5.2. **Table 4** lists the allowed maximum number of quality enhancement layers in Mode 2q coded video bitstreams.

Table 4. Allowed maximum number of quality enhancement layers in Mode 2q[1]

		Quality			
		None	**CGS**	**CGS rewrite**	**MGS**
Temporal	**None**	[0]	-	-	-
	1	[1]	2	2	4
	2	[1]	2	2	4
	3	[1]	2	2	4
	4	[1]	2	2	4

1: Numbers is square brackets indicate demotion to a lower UC mode.

Encoders that conform to this mode must support the CGS non-rewrite mode.

NOTE: When the CGS mode is in use, the value of QID must be 0 in all SVC extension slice headers. The value of DID is 0 for the lowest quality layer, 1 for the first quality enhancement, and so forth. When the MGS mode is in use, the value of DID is 0 in all SVC extension slice headers. The QID is 0 for the lowest quality layer, 1 for the first quality enhancement sub-layer, and so forth. The values of TID must be the same as that in the corresponding lowest quality (QID=0) layer frame in all SVC extension slice headers.

Let DQId be the layer representation identifier of a layer of a coded video sequence, which is set equal to (DID << 4) + QID as defined in Eq. G-63 of [1] (Section G.7.4.1.1). DID and QID are the corresponding dependency_id and quality_id values associated with the particular layer representation.

In CGS scalability, in all SVC NAL units that have DID>0, ref_layer_dq_id shall have the value (DID-1)<<4.

In MGS scalability, in all SVC NAL units with QID>0, ref_layer_dq_id shall have the value DQId-1.

In all NAL units with DID=0 and QID=0 the value of no_inter_layer_pred_flag must be 1, and the values of output_flag, discardable_flag and use_ref_base_pic_flag must be 0. The value of use_ref_base_pic_flag must be 0 unless the slice is part of a reference base picture, where use_ref_base_pic_flag is set to 1. The values of discardable_flag and output_flag must be set to 1 in all slices associated with the layer representation that has the largest value of DQId in the coded video sequence and 0 otherwise.

Encoders conforming to Mode 2q must be able to generate bitstreams in Mode 0, in Mode 1, and in Mode 2q. The run-time configuration is negotiated between decoders and encoders using mechanisms that are outside the scope of this specification.

Figure 6 shows an example where temporal scalability with two layers is combined with quality scalability with one enhancement layer.

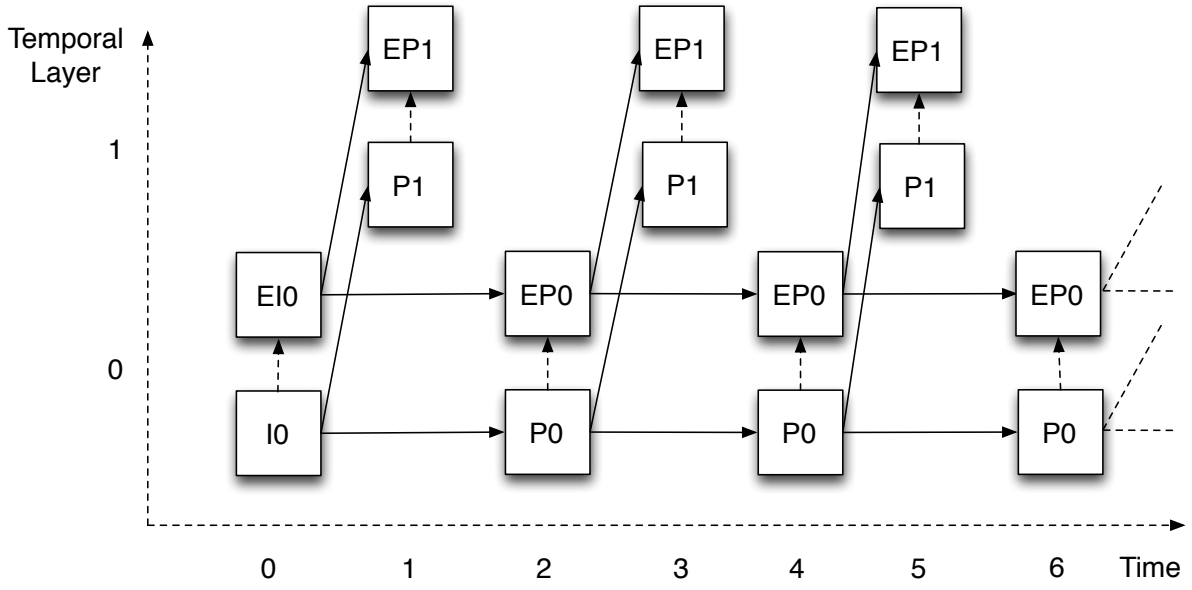

Figure 6. Example of 2-layer temporal combined with 2-layer CGS quality scalability (UC Mode 2q)

Solid arrows represent temporal prediction and reference, and dashed arrows represent inter-layer prediction and reference. Each picture is identified by a one- or two-character type indication, followed by its temporal layer. The type indications are I for intra, P for predicted, EI for intra enhancement, and EP for P picture enhancement. The temporal layers here are 0 and 1.

If we assume source material of 720p 30 fps and a fixed quantization stepsize (QP) with the two CGS quality layers using QPs of 38 (low quality) and 34 (high quality), the coding structure of **Figure 6** will provide the following 4 layers:

1. 720p 15fps with QP 38 as the base layer (TID=0, DID=0).
2. 720p 15fps with QP 38 as the temporal enhancement of the base layer (TID=1, DID=0) which, when combined with the base layer, provides 720p 30fps.
3. 720p 15fps with QP 34 as the quality enhancement layer of temporal layer 0 (TID=0, DID=1).
4. 720p 15fps with QP 34 as the quality enhancement layer of temporal layer 1 (TID=1, DID=1).

Table 5 shows the partitioning into two temporal and two CGS quality layers (4 total).

Table 5. Mode 2q Stream Example

		DID	
		0	1
TID	0	720p 15fps low quality	720p 15 fps high quality
	1	720 15fps (30 fps total)	720p 15fps (30 fps total) high quality

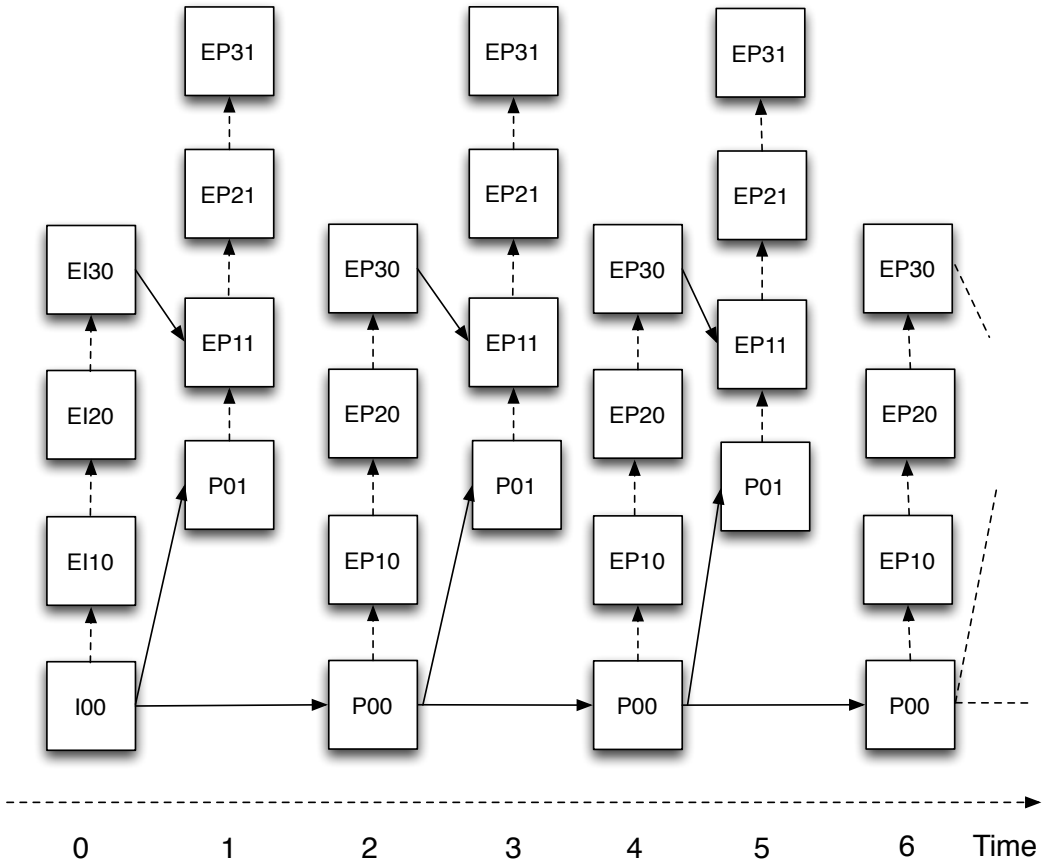

Figure 7. Example of 2-layer temporal combined with 4-layer MGS quality scalability (UC Mode 2q)

Figure 7 shows an example where temporal scalability with two layers is combined with MGS quality scalability with three enhancement layers. In this example, two digits follow the type indication. The first is the QID (0 through 3) and the second is the TID (0 or 1). Notice that the temporal reference for the lowest quality layer (QID=0) is always the lowest quality layer of the reference picture (e.g., both P01 and P00 are predicted from I00). The lowest quality layer thus is a valid AVC stream at the full temporal frame rate of the original video sequence. For quality layers higher than the lowest (QID>0), the temporal reference for the reference base pictures (pictures 0, 2, 4, etc.) is the lowest quality layer (QID=0) whereas for the other pictures (pictures 1, 3, 5, etc.) it is the highest available quality layer (QID=3). For example, for EP11 at time 1 the temporal reference is EI30.

Implementations that elect to support MGS must support the prediction structure shown in **Figure 7**, with the number of temporal and quality layers subject to the limitations provided in **Table 4**. Alternative configurations may be defined in the future. Signaling of the use of different prediction structures is outside the scope of this specification.

Table 6 illustrates the bitstream structure for the stream of **Table 5**. Even-numbered access units are shown in shaded cells.

Table 6. Mode 2q Bitstream Structure Example

NAL unit (type)	Relevant fields in the NAL	Description
SPS (7)	SPSID = 0	SPS of stream 720p 30Hz
PPS (8)	PPSID = 2, SPSID = 0	PPS of stream 720p 30Hz
SSPS (15)	SPSID = 1	Subset SPS of stream 720p 30Hz
PPS (8)	PPSID = 3, SPSID = 1	PPS of stream 720p 30Hz SVC
Prefix (14)	TID=0, DID=0	Prefix NAL of stream 720p 15Hz
IDR slice (5)	PPSID = 2, POC = 0	Base layer IDR slice(s) in stream 720p 15Hz
SVC ext slice (20)	PPSID = 3, TID=0, DID=1	EI slice(s) in stream 720p 15Hz
Prefix (14)	TID=1, DID=0	Prefix NAL of stream 720p 30Hz
Non-IDR slice (1)	PPSID = 2, POC = 1	P slice(s) of stream 720p 30Hz
SVC ext slice (20)	PPSID = 3, TID=1, DID=1	EP slice(s) in stream 720p 30Hz
Prefix (14)	TID=0, DID=0	Prefix NAL of stream 720p 15Hz
Non-IDR slice (1)	PPSID = 2, POC = 2	P slice(s) of stream 720p 15Hz
SVC ext slice (20)	PPSID = 3, TID=0, DID=1	EP slice(s) in stream 720p 15Hz
…	…	…

The possible combinations of number of scalable layers, maximum frame rates, and resolutions are limited by the maximum macroblock processing rate defined as a part of the encoder and decoder capabilities. The number of macroblocks per second for the highest quality layer in the highest temporal resolution must not exceed this limit. For example, **Table 7** shows the macroblock processing rates for each temporal resolution in the above example.

Table 7. Macroblock Processing Rate Example

		DID	
		0	1
TID	0	54,000	54,000
	1	54,000	54,000

In this example, the sum of macroblocks per second for the hiqhest quality layer is 108,000 MB/sec. This number must be smaller than or equal to the maximum macroblocks per second supported by the encoder or decoder.

5.4. UC Mode 2s: SVC with Temporal and Spatial Scalability

This mode addresses the configuration of encoders that use SVC with temporal and spatial scalability. Encoders conforming to this mode may generate any combination of one to four temporal layers, with two to four spatial layers (one to three enhancement layers).

Encoders conforming to this mode must be able to generate bitstreams with at least one spatial enhancement layer and two temporal layers. The number of temporal layers follows the constraints

specified in Section 5.2. **Table 8** lists the maximum allowed number of spatial enhancement layers in Mode 2s coded video bitstreams.

Table 8. Allowed maximum number of spatial enhancement layers in Mode 2s[1]

		Spatial
Temporal	**None**	-
	1	2
	2	2
	3	2
	4	2

Let DQId be the layer representation identifier of a layer of a coded video sequence, which is set equal to (DID << 4) + QID as defined in Eq. G-63 of [1] (Section G.7.4.1.1). DID and QID are the corresponding dependency_id and quality_id values associated with the particular layer representation.

In all SVC NAL units that have DID>0, ref_layer_dq_id shall have the value (DID-1)<<4.

In all NAL units with DID=0 and QID=0 the value of no_inter_layer_pred_flag must be 1, and the values of output_flag, discardable_flag and use_ref_base_pic_flag must be 0. The values of discardable_flag and output_flag must be set to 1 in all slices associated with the dependency representation that has the largest value of DID in the coded video sequence and 0 otherwise.

Encoders conforming to Mode 2s must be able to generate bitstreams in Mode 0, in Mode 1, and in Mode 2s. The run-time configuration is negotiated between decoders and encoders which is outside the scope of this specification.

Figure 8 shows an example where temporal scalability with two layers is combined with spatial scalability with one enhancement layer. Solid arrows represent temporal prediction and reference, and dashed arrows represent inter-layer prediction and reference. Each picture is identified by a one- or two-character type indication, followed by its temporal layer. The type indications are I for intra, P for predicted, EI for intra enhancement, and EP for P picture enhancement. The temporal layers here are 0 and 1.

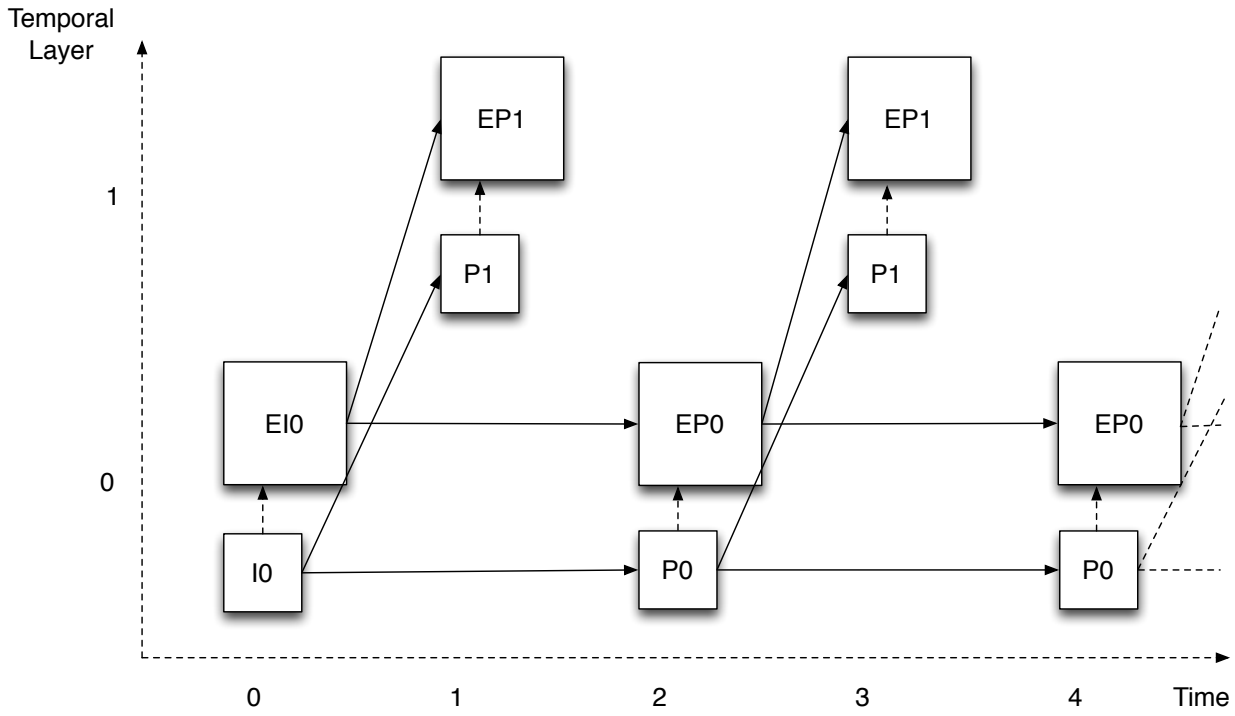

Figure 8. Example of 2-layer temporal combined with 2-layer spatial scalability (UC Mode 2s)

Assuming source material of 720p 30 fps as an example, and using a 2:1 resolution ratio, the coding structure of **Figure 8** will provide the following 4 layers:

1. 360p 15 fps as the base layer (TID=0, DID=0).
2. 360p 15 fps as the temporal enhancement of the base layer 1 (TID=1, DID=0) which, when combined with the base layer, provides 360p 30 fps.
3. 720p 15 fps as the spatial enhancement of temporal layer 0 (TID= 0, DID=1).
4. 720p 15 fps as the spatial enhancement of temporal layer 1 (TID=1, DID=1).

Table 9 shows the partitioning into two temporal layers and two spatial layers with 2:1 spatial scalability ratio (4 total).

Table 9. Mode 2s Stream Example

		DID	
		0	1
TID	0	360p 15fps low resolution	720p 15 fps high resolution
	1	360 15fps (30 fps total)	720p 15fps (30 fps total) high resolution

Table 10 illustrates the bitstream structure for the stream of **Table 9**. Even-numbered access units are shown in shaded cells.

Table 10. Mode 2s Bitstream Structure Example

NAL unit (type)	Relevant fields in the NAL	Description
SPS (7)	SPSID = 0	SPS of stream 360p 30Hz
PPS (8)	PPSID = 2, SPSID = 0	PPS of stream 360p 30Hz
SSPS (15)	SPSID = 1	Subset SPS of stream 720p 30Hz
PPS (8)	PPSID = 3, SPSID = 1	PPS of stream 720p 30Hz SVC scalable layer
Prefix (14)	TID=0, DID=0	Prefix NAL of stream 360p 15Hz
IDR slice (5)	PPSID = 2, POC = 0	Base layer IDR slice(s) in stream 360p 15Hz
SVC ext slice (20)	PPSID = 3, TID=0, DID=1	EI slice(s) in stream 720p 15Hz
Prefix (14)	TID=1, DID=0	Prefix NAL of stream 360p 30Hz
Non-IDR slice (1)	PPSID = 2, POC = 1	P slice(s) of stream 360p 30Hz
SVC ext slice (20)	PPSID = 3, TID=1, DID=1	EP slice(s) in stream 720p 30Hz
Prefix (14)	TID=0, DID=0	Prefix NAL of stream 360p 15Hz
Non-IDR slice (1)	PPSID = 2, POC = 2	P slice(s) of stream 360p 15Hz
SVC ext slice (20)	PPSID = 3, TID=0, DID=1	EP slice(s) in stream 720p 15Hz
…	…	…

The possible combinations of number of scalable layers, maximum frame rates, and resolutions are limited by the maximum macroblock processing rate defined as a part of the encoder and decoder capabilities. The number of macroblocks per second for the highest spatial layer in the highest temporal resolution must not exceed this limit. For example, **Table 11** shows the macroblock processing rates for each temporal resolution in the above example.

Table 11. Macroblock Processing Rate Example

		DID	
		0	1
TID	0	13,800	54,000
	1	13,800	54,000

In this example, the sum of macroblocks per second for the highest spatial layer is 108,000 MB/sec. This number must be smaller than or equal to the maximum macroblocks per second supported by the encoder or decoder.

5.5. UC Mode 3: SVC with Temporal, Quality, and Spatial Scalability

This mode addresses the configuration of encoders that use SVC with temporal, quality, and spatial scalability. **Table 12** lists the maximum allowed number of quality enhancement layers in Mode 3 coded video bitstreams.

Table 12. Allowed maximum number of quality enhancement layers in Mode 3[1, 2]

		Quality			
		None	**CGS**	**CGS rewrite**	**MGS**
Spatial	**None**	[0]	[2q]	[2q]	[2q]
	1	[2s]	2	2	4
	2	[2s]	2	2	4

1: Numbers in square brackets indicate demotion to a lower mode.

2: Temporal scalability must have 2 to 4 layers.

Let DQId be the layer representation identifier of a layer of a coded video sequence, which is set equal to (DID << 4) + QID as defined in Eq. G-63 of [1] (Section G.7.4.1.1). DID and QID are the corresponding dependency_id and quality_id values associated with the particular layer representation.

In all SVC NAL units that have DID>0 and QID =0, ref_layer_dq_id shall have the value (DID-1)<<4.

In all SVC NAL units that have QID>0, ref_layer_dq_id shall have the value DQId-1.

In all NAL units with DID=0 and QID=0 the value of no_inter_layer_pred_flag must be 1, and the values of output_flag, discardable_flag and use_ref_base_pic_flag must be 0. The value of use_ref_base_pic_flag must be 0 unless the slice is part of a reference base picture, where use_ref_base_pic_flag is set to 1. The values of discardable_flag and output_flag must be set to 1 in all slices associated with the layer representation that has the largest value of DQId in the coded video sequence and 0 otherwise.

Encoders conforming to Mode 3 must be able to generate bitstreams in Mode 0, Mode 1, Mode 2q, Mode 2s, and Mode 3. The run-time configuration is negotiated between decoders and encoders which is outside the scope of this specification.

Figure 9 shows an example with two temporal layers where a CGS quality enhancement layer is applied on a spatial enhancement layer.

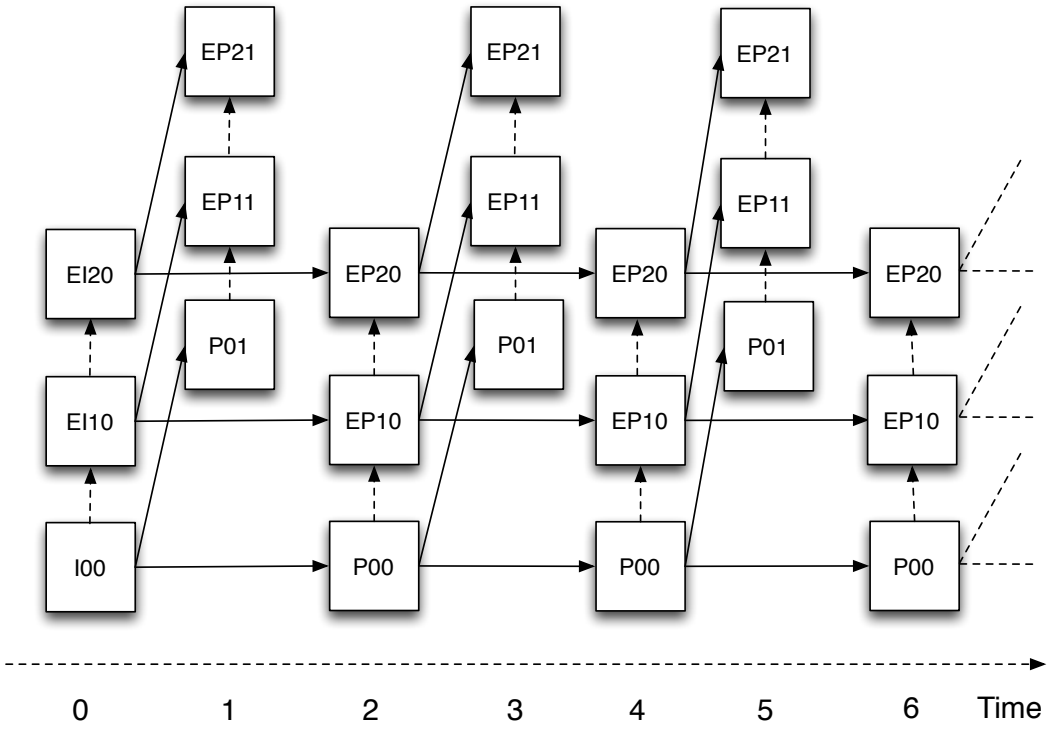

Figure 9. Example of 2-layer temporal combined with 2-layer spatial followed by a CGS quality enhancement layer (UC Mode 3)

Solid arrows represent temporal prediction and reference, and dashed arrows represent inter-layer prediction and reference. Each picture is identified by a one- or two-character type indication, followed by two-digit layer designation. The type indications are 'I' for intra, 'P' for predicted, 'EI' for intra enhancement, and 'EP' for P picture enhancement. The two-digit layer designation combines the DID as the first digit with the TID as the second digit. For example, EP11 is an enhancement layer picture with DID=1 and TID=1.

Note that from the figure it is not obvious what type of scalability each DID corresponds to. Indeed, each one of them can be either spatial or quality (CGS) scalability.

> NOTE: The requirement that an enhancement layer can only depend on a layer with DID one less than its own means that the arrows connecting layers vertically in a scalability diagram such as the one in **Figure 9** cannot skip a layer.

The same diagram of **Figure 9** could be used if the order of scalability types were different, for example a CGS quality enhancement layer followed by a spatial enhancement layer.

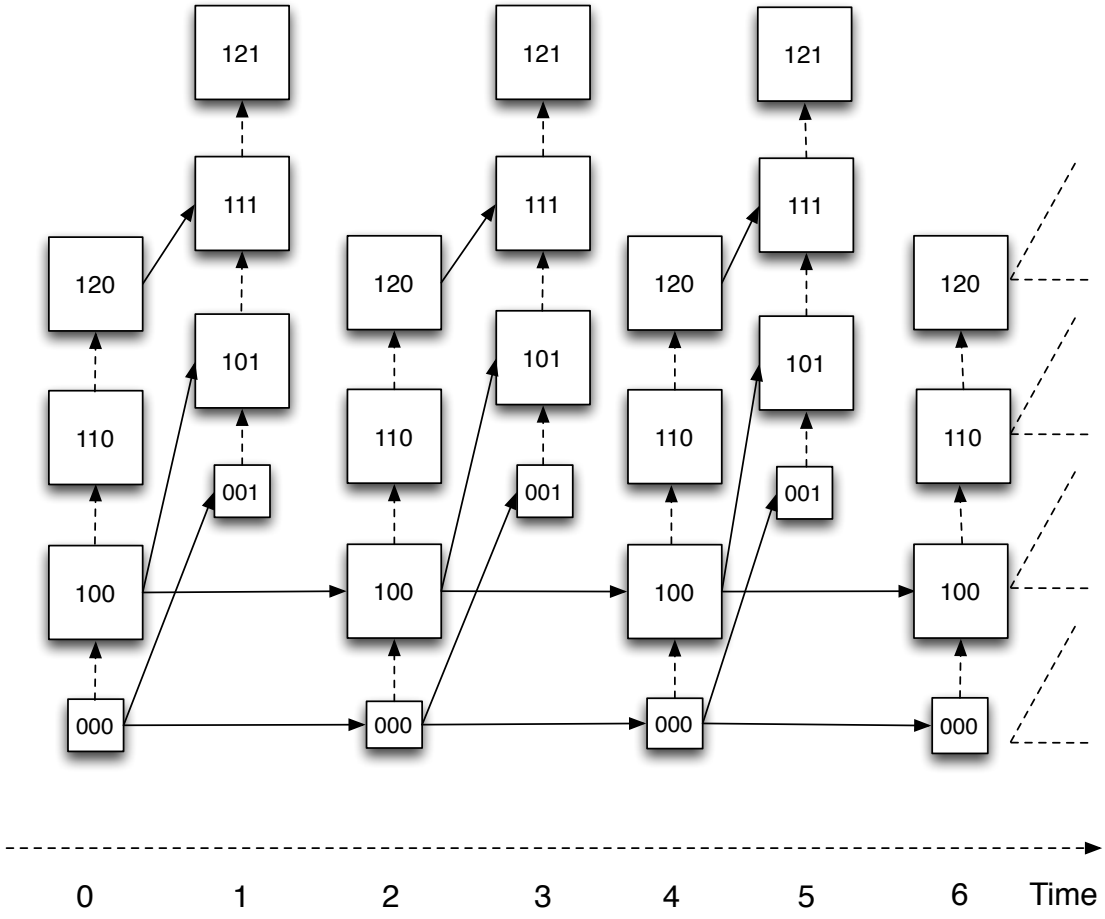

Figure 10. Example of 2-layer temporal combined with 2-layer spatial and 3-layer MGS quality scalability (UC Mode 3)

An MGS example with 2 temporal, 2 spatial, and 3 MGS quality layers is shown in **Figure 10**. In this example we explicitly show the spatial scalability components using different picture sizes. Solid arrows represent temporal prediction and reference, and dashed arrows represent inter-layer prediction and reference. Each picture is identified by a three-digit layer designation, corresponding to DID, QID, and TID, respectively. For example, "120" is a layer with DID=1, QID=2, and TID=0. As shown in the figure, the MGS layers are created after the spatial enhancement is applied (100 and 101).

We now identify the different types of layers available in the coding structure of **Figure 9**. We assume source material of 720p at 30fps, and further assume that that coding structure of **Figure 9** depicts quality scalability (DID=1), followed by spatial scalability (DID=2), followed again by quality scalability (DID=3). For quality scalability we will use two fixed quantizer settings (QP=38 and QP=34), and for spatial scalability a ratio of 2:1. We have the following 8 layers:

1. 360p 15fps with QP 38 as the base layer (DID=0, TID=0).
2. 360p 15fps with QP 38 as the temporal enhancement of the base layer (DID=0, TID=1) which, when combined with the base layer, provides 360p 30fps.
3. 360p 15fps with QP 34 as the quality enhancement layer to temporal layer 0 (DID=1, TID=0).

4. 360p 15fps with QP 34 as the quality enhancement layer to temporal layer 1 (DID=1, TID=1) which, when combined with the previous layer, provides 360p 30fps with QP 34.
5. 720p 15fps with QP 38 as the spatial enhancement layer to the quality enhancement layer of temporal layer 0 (DID=2, TID=0).
6. 720 p 15fps with QP38 as the spatial enhancement layer to the quality enhancement layer of temporal layer 1 (DID=2, TID=1).
7. 720p 15fps with QP 34 as the quality enhancement layer to the spatial enhancement layer of temporal layer 0 (DID=3, TID=0).
8. 720 p 15fps with QP34 as the quality enhancement layer to the spatial enhancement layer of temporal layer 1 (DID=3, TID=1).

Table 13 shows the partitioning into two temporal layers, two CGS quality layers (QPs 34 and 38), and two spatial resolutions (360p and 720p), for a total of 8 layers.

Table 13. Mode 3 Stream Example

		DID			
		0	1	2	3
TID	0	360p 15fps QP 38	360p 15fps QP 34	720p 15fps QP 38	720p 15fps QP 34
	1	360 15fps QP 38	360p 15fps QP 34	720p 15fps QP 38	720p 15fps QP 34

Table 14 illustrates the bitstream structure for the stream of **Table 13**. Even-numbered access units are shown in shaded cells.

Table 14. Mode 3 Bitstream Structure Example

NAL unit (type)	Relevant fields in the NAL	Description
SPS (7)	SPSID = 0	SPS of stream 360p 30Hz
PPS (8)	PPSID = 3, SPSID = 0	PPS of stream 360p 30Hz
SSPS (15)	SPSID = 1	Subset SPS of stream 360p 30 Hz SVC
PPS (8)	PPSID = 4, SPSID = 1	PPS of stream 360p 30 Hz SVC
SSPS (15)	SPSID = 2	Subset SPS of stream 720p 30Hz SVC
PPS (8)	PPSID = 5, SPSID = 2	PPS of stream 720p 30Hz SVC
Prefix (14)	PRID = 0, TID=0, DID=0	Prefix NAL of stream 360 15Hz
IDR slice (5)	PPSID = 3, POC = 0	Base layer IDR slice(s) in stream 360p 15Hz
SVC ext slice (20)	PPSID = 4, TID=0, DID=1	EI slice(s) in stream 360p 15Hz
SVC ext slice (20)	PPSID = 5, TID=0, DID=2	EI slice(s) in stream 720p 15Hz
SVC ext slice (20)	PPSID = 5, TID=0, DID=3	EI slice(s) in stream 720p 15Hz
Prefix (14)	TID=1, DID=0	Prefix NAL of stream 360p 30Hz
Non-IDR slice (1)	PPSID = 3, POC = 1	P slice(s) of stream 360p 30Hz
SVC ext slice (20)	PPSID = 4, TID=1, DID=1	EP slice(s) in stream 360p 30Hz
SVC ext slice (20)	PPSID = 5, TID=1, DID=2	EP slice(s) in stream 720p 30Hz
SVC ext slice (20)	PPSID = 5, TID=1, DID=3	EP slice(s) in stream 720p 30Hz

Prefix (14)	TID=0, DID=0	Prefix NAL of stream 360p 15Hz
Non-IDR slice (1)	PPSID = 3, POC = 2	P slice(s) in stream 360p 15Hz
SVC ext slice (20)	PPSID = 4, TID=0, DID=1	EP slice(s) in stream 360p 15Hz
SVC ext slice (20)	PPSID = 5, TID=0, DID=2	EP slice(s) in stream 720p 15Hz
SVC ext slice (20)	PPSID = 5, TID=0, DID=3	EP slice(s) in stream 720p 15Hz
…	…	…

The possible combinations of number of scalable layers, maximum frame rates, and resolutions are limited by the maximum macroblock processing rate (MB/sec) defined as a part of the encoder and decoder capabilities. The number of macroblocks per second for the highest quality layer in the highest spatial resolution and in the highest temporal resolution must not exceed this limit. For example, **Table 15** shows the macroblock processing rates for each temporal resolution in the above example.

Table 15. Macroblock Processing Rate Example

		DID			
		0	1	2	3
TID	0	13,800	13,800	54,000	54,000
	1	13,800	13,800	54,000	54,000

In this example, the sum of macroblocks per second for the highest spatial layer and highest quality layer is 108,000 MB/sec. This number must be smaller than or equal to the maximum macroblocks per second supported by the encoder or decoder.

6. Bitstream Priority Assignment

The NAL unit header extension introduced in Annex G of H.264 includes the 6-bit field priority_id (PRID) (Section G.7.3.1.1). The PRID does not affect the decoding process and its use and definition is left to applications.

This specification uses the PRID field to define the preferred ordering with which layers of a scalable bitstream should be removed, when necessary. The use of the PRID field allows encoders to provide an indication of the relative significance of individual layers, so that intermediate processing systems as well as decoders can make informed choices when they have to perform bitstream adaptation.

The PRID field values are assigned from the smallest to the highest, with a higher value indicating that the associated bitstream component is less significant than another bitstream component with a lower PRID value. In other words, the highest priority is given to the lowest PRID values.

> NOTE: As explained in Section G.8.8.1 of the H.264 specification, bitstreams conforming to one of the profiles defined in H.264 Annex G must contain at least one NAL unit with PRID, DID, TID, and QID, all equal to 0.

The PRID must be used in Modes 1, 2s, 2q, and 3, and should be used in Mode 0 bitstreams. PRID values are associated only with NAL units that contain coded slice data. NAL units where the PRID is not present should be treated as having the lowest value of PRID (i.e., 0).

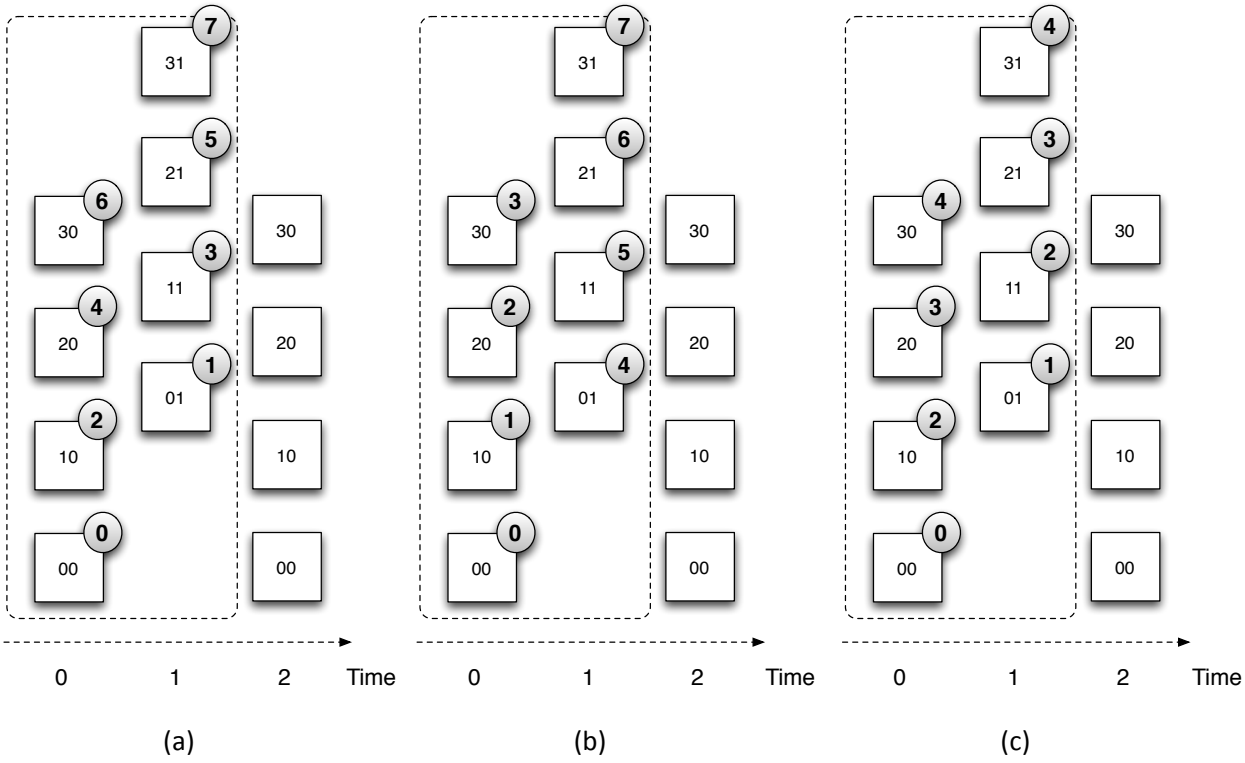

Figure 11. Examples of PRID assignments.

A specific process of assigning PRID's to combinations of DID, TID, and QID, implicitly defines a path in the three-dimensional scalability space created by the values of DID, TID, and QID within a particular bitstream. This path, from the highest value of priority_id to the lowest, is the one to be used when one wants to remove layers from the signal for adaptation or other purposes. The selection of a particular path depends on several factors, including operating conditions that may be known only during run-time. In some applications, for example, it may be preferrable to first remove spatial enhancement layers rather than reducing temporal resolution, whereas in others it may be preferable to reduce temporal resolution rather than spatial (e.g., content sharing).

Figure 11 depicts three examples of PRID assignment using a scalability structure with 2 temporal layers and 4 spatial or quality layers. The prediction references are not shown for simplicity. Each layer is identified by a two digit (DT) combination, with the first digit indicating the DID and the second digit indicating the TID.

In **Figure 11**(a) the assignment is performed across TID's, such that the path (starting from 0) first traverses all available TIDs before moving to the next DID (with TID=0). This design favors the

preservation of a high frame rate with the highest possible quality and would be an appropriate choice for regular video of participants.

In **Figure 11**(b) the assignment is performed within a TID, such that the path (again starting from 0) first traverses all DIDs for TID=0 before moving to scan all DIDs for TID=1. Here the design prefers to maintain high resolution even if it means that the frame rate may be reduced.

In both schemes (a) and (b) the granularity of the PRID assignment is at the level of individual bitstream components (distinct DID and TID combinations). In (a), for example, we can eliminate bistream component 31 thus reducing the quality just for the odd frames. This uneven elimination of bitstream components across frames may not be desirable in some applications. The scheme in **Figure 11**(c) is similar to the one in (a), but here the same PRID value is used across different temporal layers. This ensures that if a particular spatial or quality enhancement layer is removed from the video bitstream, it is removed from all pictures regardless of their TID.

This specification does not specify the process through which PRID assignment is performed.

> NOTE: Encoders should utilize the priority_id_setting_flag/priority_id_setting_uri which are included in the Scalability Information SEI message (G.13.1.1 [1]) to provide a description of the method used to calculate the PRID values.

* * *