



W H I T E P A P E R

SVC Uncovered

A short primer on the basics of Scalable Video Coding and its benefits

Table of Contents

1	Introduction	3
2	What Is SVC?	3
2.1	What Is Temporal Scalability?	4
2.2	What Is Spatial Scalability?	5
2.3	Combining Temporal and Spatial Scalability	6
3	Why SVC?	6
3.1	Video Conferencing Bridges	6
3.2	Error Resiliency	7
4	Conclusion.....	8

SVC Uncovered

1 Introduction

Scalable Video Coding (SVC) is an extension of the H.264/MPEG-4 AVC video compression standard, used today in most video communications endpoints. This nontechnical white paper will detail the components that comprise SVC, how these components work and the key benefits SVC can bring to video conferencing, including better video quality and reduction in latency (delays due to transcoding). Employing SVC also reduces infrastructure costs, as more of the computing load is handled by the endpoints rather than the more expensive MCUs (Multipoint Control Units).

AUTHOR PROFILE



Stefan Slivinski is the Manager of the Video Team at LifeSize, a division of Logitech. His team designs and develops all of the video algorithms that go into LifeSize's current and next generation video communications equipment.

His responsibilities over his ten-year career have included design and development of embedded video compression algorithms. Prior to joining LifeSize in 2005, he was at UB Video, developing video compression codecs for many of the major video conferencing OEM providers.

2 What Is SVC?

SVC stands for Scalable Video Coding. It is an enhancement within the H.264/MPEG-4 AVC video compression standard. The SVC concept itself is not new as it has actually been part of nearly every major video compression standard from H.263 to MPEG-4. In simpler terms, SVC provides the ability to encapsulate multiple compressed video sequences at various frame rates and resolutions in order to combine them into a single stream. While the video compression algorithm at the core of SVC's technology is still very much AVC, what SVC provides is a method to consolidate multiple video sequences at various frame rates and resolutions into one container and then takes it one step further, allowing these same video sequences to share information with each other in order to improve the quality of video.

SVC is made up of two components: 1) *temporal scalability* and 2) *spatial scalability*. (A third mode, *quality scalability*, also exists; however, for the purposes of this paper it is just a special case of spatial scalability.) Temporal scalability provides the ability to have multiple frame rates for a given resolution. Spatial scalability provides the ability to have multiple resolutions of a given video sequence. Each unique frame rate, or resolution, within the video sequence is referred to as a *layer*. The following sections are designed to further define temporal and spatial scalability and how they differ.

SVC Uncovered

2.1 What Is Temporal Scalability?

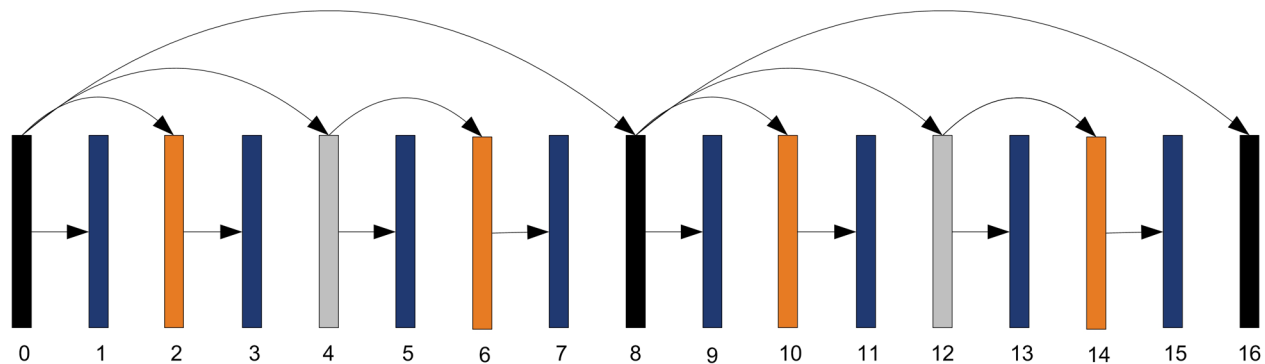


Illustration 1: Concept of temporal scalability

Temporal scalability is the ability to have multiple frame rates for the same video stream/resolution. This ability is not a new feature of SVC, as it is possible to do this with AVC, but SVC, being an enhanced form of AVC, simplifies how it’s done and makes it more obvious that there are multiple frame rates.

In *Illustration 1*, each rectangle represents a single frame from a video sequence. Assume that each frame represents 1/60th of a second, and therefore the frame rate of the video sequence is 60 frames per second (fps). The arrows show the dependencies between two frames. For example, in order to decode frame 4, frame 0 needs to be decoded also; therefore, frame 0 is a dependency of frame 4. Now consider frame 1; no frame is dependent on it, so it could be skipped and all remaining frames could be decoded. You could

then also skip all subsequent blue frames (frames 3, 5, 7, 9, 11, 13, and 15), which means you would then only decode half the frames, thus cutting the frame rate by a factor of 2 from 60 fps to 30 fps. Now that the blue frames are gone, you could skip the orange frames (frames 2, 6, 10, and 14) as no other frames are dependent on them anymore. This would again cut the frame rate in half, resulting in 15 fps. You could then skip the gray frames (frames 4 and 12) and be left with 7.5 fps. In summary, depending on which frames you choose to decode, this one stream could have 60 fps, 30 fps, 15 fps or 7.5 fps. That’s one video stream with four possible different frame rates — temporal scalability.

SVC Uncovered

2.2 What Is Spatial Scalability?

Spatial scalability is the ability to have two or more resolutions of the same video sequence within the same stream. This can be achieved easily by using one of the many container formats, whereby two or more completely separate video streams are combined. Then, the overall size of the container will be equal to the sum of all the streams within it (stream 1 + stream 2 + up to stream n). The difference is that SVC uses identical streams that happen to be at different resolutions. Fundamentally, the way traditional video compression works is by exploiting the fact that very little will change between two consecutive frames in time in order to reduce the amount of information necessary to represent that video sequence. The same holds true for two frames from the same instance in time but at slightly different resolutions. Spatial scalability uses information from different layers in order to reduce the overall size so that the combined size of independent streams ends up being potentially much smaller. This use of information between layers is referred to as interlayer prediction and forms the core of SVC. *Figure 1* shows a diagram of the interlayer dependencies of spatial scalability.

The three images in *Figure 1* represent three single video frames from the same point in time. Just as with temporal scalability, the arrows represent a dependency between the layers. Therefore, in order to decode layer 1, information from layer 0 is needed, and in order to decode layer 2, information is also needed from layer 1. The lowest resolution layer is often known as the base layer and is required by SVC to be fully interoperable with AVC, meaning it cannot use any components of SVC – ensuring that decoders capable of decoding only AVC could also decode the base layer of any SVC stream.

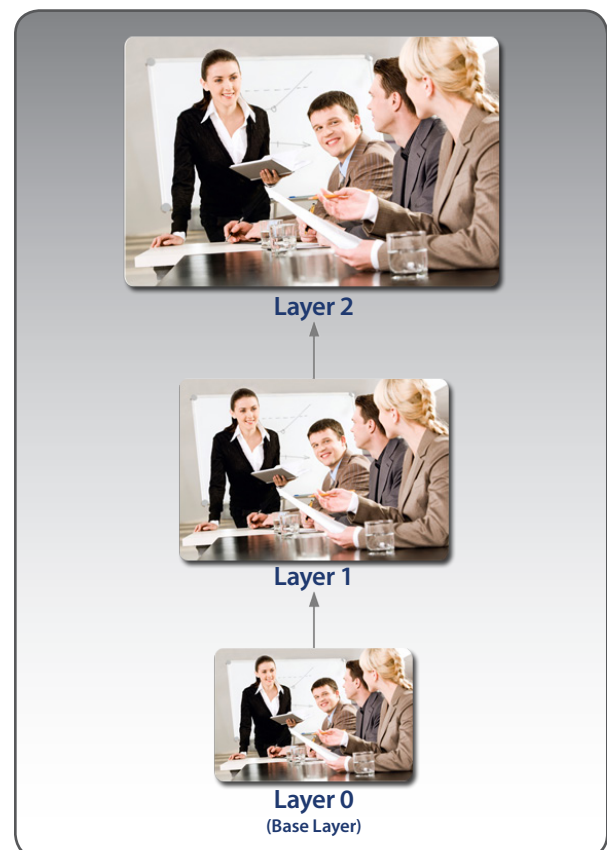


Figure 1: Diagram depicting interlayer dependencies of spatial scalability

SVC Uncovered

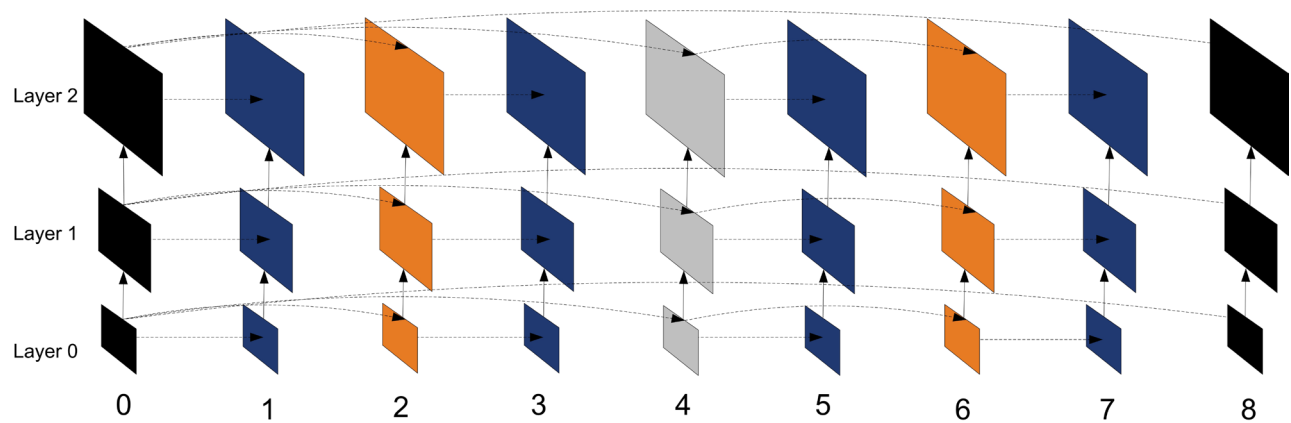


Illustration 2: Combination of temporal and spatial layer dependencies

2.3 Combining Temporal and Spatial Scalability

Temporal and spatial scalability are not mutually exclusive and in fact would most likely be used in conjunction with each other, which creates a single stream that contains multiple resolutions at multiple frame rates. *Illustration 2* depicts a diagram with both temporal and spatial layer dependencies.

3 Why SVC?

Having multiple resolutions at multiple frame rates can have its benefits even if only one resolution or one frame rate will be viewed at any given time. There are many applications for SVC, both outside and within the video communications space.

3.1 Video Conferencing Bridges

Video conferencing bridges provide the ability for multiple people to communicate on the same video call. They are essentially the same as teleconferencing bridges but with the addition of video. Just as you communicate with a teleconferencing bridge using a telephone, an endpoint is used to talk through a video bridge. Unlike telephones, though, endpoints vary widely in their capabilities. Some endpoints are run on a smartphone and send and receive low resolutions, and some are very powerful and capable of sending and receiving very high resolutions up to full HD at 1920 x 1080 resolution. The objective of a bridge is to provide the best possible experience for each individual in the call, which means sending low resolutions to less capable endpoints and high resolutions to more capable endpoints. This requires the bridge to decode the entire incoming video, merge everyone together, and re-encode a custom stream for each endpoint that is ideally suited to its capabilities.

W H I T E P A P E R

SVC Uncovered

That means that for every participant in the call, the bridge must have one decoder and one encoder. The inherent problem with all this decoding and re-encoding is that the bridge introduces additional delay, making it harder to communicate and expensive because specialized and costly hardware will most likely be needed.

SVC is able to address this deficiency by allowing a bridge to forward video frames from one participant to everyone else without the need to actually decode or re-encode any video. Everyone receives a separate stream from everyone else on the call without incurring any additional delay. In order to provide the best possible experience regardless of an endpoint's capability, the bridge would send a small spatial layer to smartphone endpoints and a high spatial layer to the more powerful endpoints. And because the bridge does not need to encode or decode any video, it can use commodity hardware at a fraction of the cost of the specialized hardware used in today's bridges.

3.2 Error Resiliency

The Internet is imperfect; data gets lost all the time. There are some methods to ensure that data always arrives intact, but they introduce delay, which is undesirable in video communications. Therefore, video conferencing and most other types of communications that require low delay/latency do not utilize those methods, which then places the responsibility on the decoder to attempt to fill in for the missing data. This often works well depending on where the loss is, but other times this cover-up can be noticeable in the form of frozen and broken video.

SVC is able to solve this problem because of the different dependencies between layers, for example, in the case of temporal scalability.

WHITE PAPER

SVC Uncovered

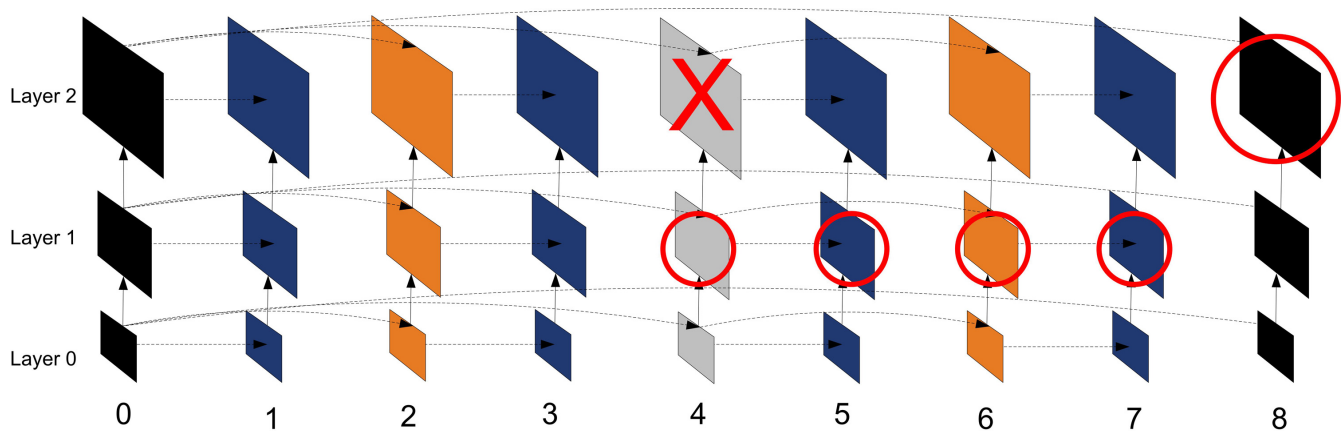


Illustration 3: Combination of temporal and spatial layer dependencies with frame loss

In *Illustration 1*, depicting temporal scalability, if a blue frame (frame 1, 3, 5, 7, 9, 11, 13, or 15) is intentionally skipped, decoding can continue because the blue frame is not a dependency for any other frame. The same logic applies if that frame is somehow lost in transmission. The side effect, of course, is that the video will pause, but a pause for a single frame is almost imperceptible to the naked eye. Now if, for example, an orange frame (frame 2, 6, 10, or 14) is lost, where there exists a dependency for other frames, then the pause would last two frames. And if a gray frame (frame 4 or 12) is lost, where there are more dependencies, then the pause would last four frames. Pauses inevitably become more noticeable as more frames are lost, but fortunately, spatial scalability solves this.

In *Illustration 3*, frame 4 from layer 2 is lost (marked with a red X). One option would be to skip it and pause the video for four frames, but a better alternative would be to drop down to layer 1 and display frames 4–7 (marked with red circles) at a lower resolution until we reach frame 8. Even though layer 1 is at a lower resolution, the average observer will find it more difficult to notice a slightly lower resolution image than to notice a long pause in the video.

4 Conclusion

SVC has great potential to lower the cost of entry into the world of video communications by reducing the cost of infrastructure products such as MCUs. The user experience can be improved with SVC supporting high quality video at any bandwidth, over any type of connection. Through the modalities of temporal and spatial scalability, SVC can efficiently handle multiple streams to ensure an optimized viewing experience, even in error-prone environments where packet loss may occur. Lastly, SVC can reduce latency in MCU environments by removing the need for each bridge to decode and re-encode video. Once the relevant standards organizations (like the Unified Communications Interoperability Forum) define how SVC-capable devices should talk to each other, then the potential for improvements to video communications can be fully realized.

Universal Video Collaboration

About LifeSize

LifeSize is a pioneer and world leader in high-definition video collaboration. Designed to make video conferencing truly universal, our full range of open standards-based systems offer enterprise-class, IT-friendly technologies that enable genuine human interaction over any distance. Founded in 2003 and acquired by Logitech in 2009, LifeSize, with its commitment to relentless innovation, continues to extend the highest-quality video conferencing capabilities to anyone, anywhere. As a founding member of the Unified Communications Interoperability Forum (UCIF), LifeSize partners with leading technology companies to enable interoperability of leading hardware and software solutions. The company is headquartered in Austin, Texas, with regional offices throughout Europe and Asia Pacific, and a network of over 1500 channel partners. Our systems are used by over 15,000 leading companies, in over 100 countries.

For more information, please visit www.lifesize.com



Corporate Headquarters:
1601 S. MoPac Expressway
Suite 100
Austin, Texas 78746 USA

Phone: +1 512 347 9300
Fax: +1 512 347 9301
Email: info@lifesize.com
www.lifesize.com

EMEA:
LifeSize Communications
Toll-Free Europe
008000 999 09 799

APAC:
LifeSize Communications
Hong Kong
Phone: +852 3189 7062