White Paper

Media and Entertainment
Video Streaming Providers, Broadcasters, and Studios

VisualOn Optimizer Reduces Encoding Bitrate Up to 40 Percent While Enhancing Visual Quality[1]

Easily integrating into encoding workflows using CPUs or GPUs and accelerating Intel® Quick Sync Video, Optimizer helps reduce delivery and storage costs for media delivery companies

Authors

Kevin A. Cone
Segment Manager, Intel

Gordon Kelly
Platform Solutions Architect, Intel

Surbhi Madan
Platform Solutions Architect, Intel

Savi Shi
Marketing Director, VisualOn

Hanyue Yang
Sr. Engineer Director, VisualOn

Roly Yu
Principal Engineer, VisualOn

Ivy Zhang
Staff Engineer, VisualOn

Dean Zhu
Staff Engineer, VisualOn

Key Takeaways:

- Lowers transmission and storage cost without impacting video quality
- Supports live real-time Content Adaptive Encoding on multiple streams without requiring a GPU
- Integrates with standard video CODECs: AVC, HEVC, and AV1

Executive Summary

Streaming high-resolution video comes with an inevitable trade-off between available bandwidth and quality of experience for the end user. Delivering uncompromised video quality typically results in high bitrates, which can result in slow starts, video buffering, and high content delivery network (CDN) and storage costs. Traditional solutions that attempt to minimize bandwidth without compromising quality are centered around the development of more intelligent video encoders.

Content Adaptive Encoding (CAE) is an innovative encoding process that is revolutionizing video streaming by reducing bandwidth usage while enhancing quality. Integrated into streaming media service providers' workflows, it enables more efficient usage of content delivery systems and reduces storage requirements, without sacrificing user experiences.

VisualOn Optimizer is an innovative CAE solution that dramatically reduces average bitrates while enhancing visual quality. It is encoder-agnostic and readily integrates with Intel® Quick Sync Video to deliver high-quality content with low bitrates.

This white paper describes VisualOn Optimizer, its impact on video encoding performance, compares it against other CAE solutions, and presents test results showing the benefit of combining the Optimizer with 4th Gen Intel® Xeon® Scalable processors and Intel® Quick Sync Video running on Intel® Data Center GPU Flex Series.

Content Adaptive Encoding

CAE was pioneered by Netflix from 2015 to 2018, with per-title, per-chunk, and per-shot encoding. Using their CAE technology, Netflix achieved over 30 percent bitrate reduction[2,3,4] without degrading visual quality, as measured by the Video Multimethod Assessment Fusion (VMAF) score.[5] The approach requires running hundreds of different encodings with different combinations of paraments to select the best results. Such an encoding regimen can be prohibitively expensive and difficult to scale for many companies, especially those with limited budgets or technical capabilities.

Building on the strides made by Netflix in encoding strategies, CAE takes video compression a step further by adapting the encoding process to the specific content of each video segment. Unlike traditional encoding methods that apply uniform settings across an entire video, CAE analyzes factors such as motion, texture, and complexity within the video to optimize encoding settings dynamically. This results in more efficient compression that preserves quality while reducing file size and bandwidth requirements. These are benefits that are

otherwise only achievable through the introduction of new, more complex compression standards that will entail high cost, long time to market, and compatibility issues within the whole ecosystem. Heuristic CAE solutions have since emerged that dramatically reduce the compute requirement while achieving close to optimal results. Jan Ozer has compared various options for H.264 encoding.[6]

Additionally innovative AI methods are now being applied to content transcoding to reduce bitrate and improve the visual experience. AI is used to dynamically analyze frames and scenes to tune encoder modes that impact both bitrate and video quality. Bitrate reduction helps lower cost of transmission and cost of storage for media providers.

In encoding, parameter choices are impacted by several factors, such as the type of delivery (live, Video on Demand (VoD), and file-to-file (F2F) transfer). Furthermore, media providers use various infrastructures in their data centers for encoding content. These technologies can include specialized software encoders running on CPUs, and GPUs with built-in transcoding accelerators.

Intel works with the ecosystem that supports these media providers to apply the power of Intel® technologies, such as the Intel® Video Processing Library (Intel® VPL), Intel Xeon Scalable processors, and Intel Data Center GPU Flex Series to their challenges. Intel's commitment to the ecosystem aims to help lower costs of operations while improving content delivery and thus the viewing customers' experiences.

VisualOn Optimizer Suite

VisualOn is a streaming solutions provider that offers a universal, encoder-agnostic, content-adaptive encoding technology. The VisualOn Optimizer suite enables streaming media companies to deliver compelling multimedia content with incredible user experience. VisualOn introduced the Optimizer suite at IBC 2023,[7] and it won the company the Product of the Year award at NAB 2024.[8]

Optimizer provides a single-pass transcoding solution that dynamically adjusts encoding settings on a per-frame basis. Supporting AVC, HEVC, and AV1, the Optimizer suite integrates with any encoder running on CPU- or GPU-based hardware. It supports a wide variety of use cases with product versions for the following deployment types:

- **Optimizer VOD.** For VoD workflows, using FFmpeg's filter-complex to transcode the entire ABR ladder in a single command.
- **Optimizer Live.** For streaming workflows with real-time transcoding. Its efficient implementation allows it to achieve zero additional latency with reducing both average and peak bitrates without compromising visual quality, ideal for large events.
- **Optimizer Fidelity.** For visually lossless file-to-file video transcoding to reduce the storage requirements of massive mezzanine video files.
- **Optimizer.** For general purpose file-to-file transcoding to reduce size of video files.

Optimizing Media Streaming with Advanced Dynamic Coding Technology

The VisualOn Optimizer uses dynamic encoding technology that combines parameter matching, scene recognition, and image quality enhancement. It leverages a machine learning-based, one-pass framework to predict and adjust encoder settings, optimizing each video segment based on its features. Based on VisualOn testing, the technology reduces bitrates up to 40 percent,[1] while maintaining or improving video quality, as measured by VMAF scores.

The technology also provides real-time feedback to dynamically adjust parameters based on scene analysis and quality metrics (PSNR, SSIM, and VMAF). It supports optional preprocessing steps, like sharpening and noise reduction, and has proven efficient and stable in live streaming tests with FFmpeg integration.

Optimizer readily integrates into any streaming workflow through a simple API call before or in parallel with the encoder (Figure 1).

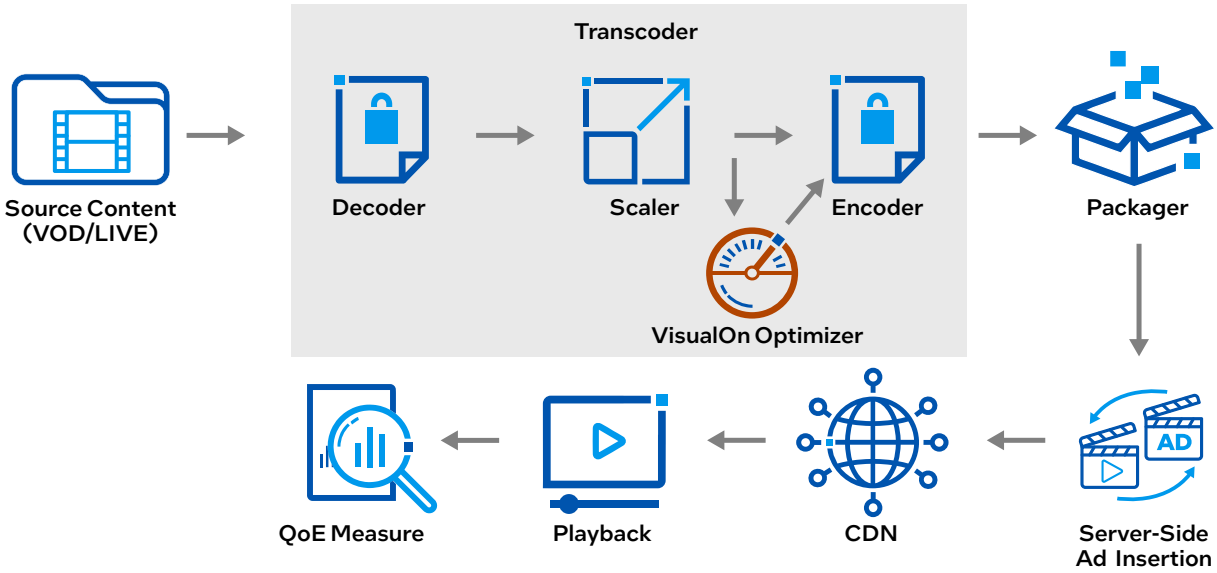

Figure 1. VisualOn Optimizer integrates into any streaming workflow.

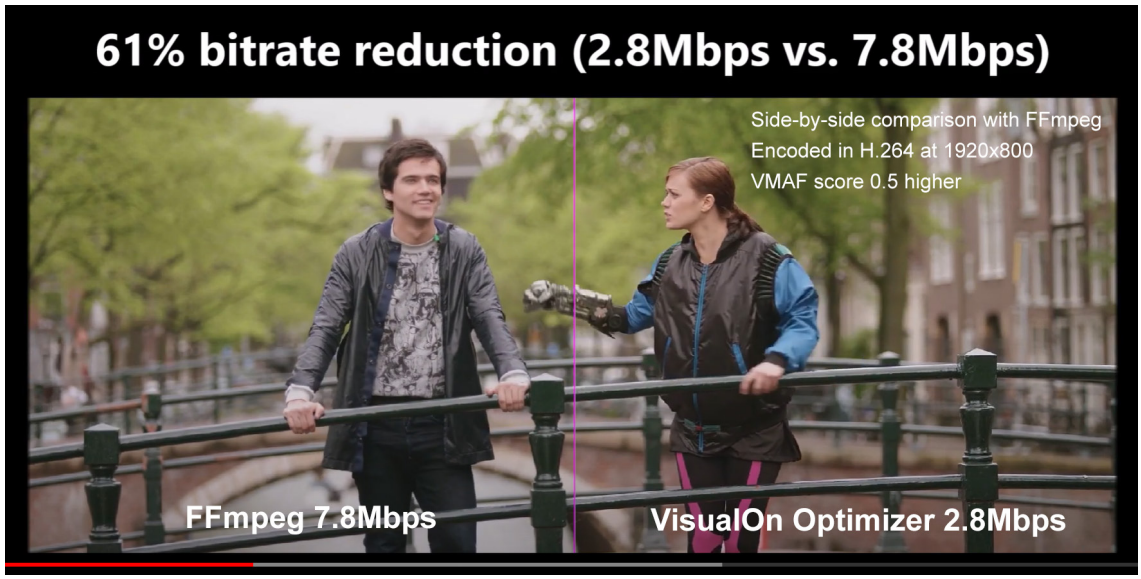

Figure 2. VisualOn Optimizer significantly reduces bit rate for more efficient content delivery and reduction in storage requirement.[9]

Figure 3. VisualOn Optimizer improves visual quality, compared to FFmpeg only (left) and VisualOn Optimizer (right).[10]

Enhancing Video Efficiency and Quality

VisualOn Optimizer is tightly embedded within the FFmpeg ecosystem. It can be readily integrated with a video encoder through FFmpeg's APIs, regardless of the video compression format or whether the encoder is CPU-based or GPU-based. Additionally, it has been integrated with Intel Quick Sync Video in Intel® Core™ processors and Intel Data Center GPU Flex Series for AVC, HEVC, and AV1 transcoding.

Reduced Bit Rates

Extensive benchmark results, as well as actual production deployments, reveal how Optimizer significantly reduces the average video bitrate. Figure 2 illustrates one such test compared to FFmpeg based on VisualOn's testing.[9]

Bit rate reduction leads to improved operational efficiency by reducing bandwidth requirements and delivery and storage costs, plus it helps lower energy consumption.

Better Visual Quality

The Optimizer drastically improves visual quality without increasing video bitrate, as illustrated in Figure 3 based on VisualOn's testing.10 Higher quality enhances user experience and improves Key Performance Indicators (KPIs).

Seamless Integration

Other CAE solutions are typically tightly bound to a particular encoder. The Optimizer is encoder-agnostic, seamlessly integrating into existing streaming workflows without disrupting operations. Since VisualOn Optimizer runs on CPUs and GPUs, often no additional hardware is required. This makes it suitable for wide deployment across multiple use cases.

Optimizer on Intel® Xeon® Scalable Processors and Intel Data Center GPU Flex Series with Intel Quick Sync Video

Using the Intel® Video Processing Library (Intel® VPL) together with FFmpeg, VisualOn integrated Optimizer to accelerate performance and improve efficiency on Intel technologies, including 4th Gen Intel Xeon Scalable processors and Intel Data Center GPU Flex Series with Intel Quick Sync Video built in.

Intel VPL

Intel VPL presents a unified programming interface for video processing and encoding across a variety of hardware—CPUs, GPUs, and other accelerators—simplifying development efforts and reducing time to market. Intel VPL helps accelerate building and running high-performance, portable media pipelines. It provides device discovery and selection in media centric and video analytics workloads, and API primitives for zero-copy buffer sharing. Intel VPL supports a wide range of video formats, resolutions, and features, making it versatile for different video processing needs.

4th Gen Intel Xeon Scalable Processors

For CPU-based encoding, Intel Xeon Scalable Processors deliver scalable performance with up to 56 cores per CPU and built-in Intel® AI Accelerator engines, such as Intel® Advanced Vector Extensions 512 (Intel® AVX-512). Intel AI Accelerator engines provide dedicated AI functions built into the silicon that enhance performance for specific AI and complex computing tasks.

Intel AVX-512 extends vector processing capabilities with wider vectors, a new extensible syntax, and richer functionality. With ultra-wide 512-bit vector operations, Intel AVX-512 can handle the most demanding computational tasks while reducing the energy needed to complete them. Applications can pack 32 double-precision and 64 single-precision floating point operations per clock cycle within the 512-bit vectors, as well as eight 64-bit and sixteen 32-bit integers.

Intel Xeon Scalable processors also provide improved power efficiency and TCO with platform enhancements like DDR5 memory and PCIe 5 support. Enhanced security features include Intel® Software Guard Extensions (Intel® SGX) and Intel® Total Memory Encryption to ensure comprehensive data protection. This family of high-performance CPUs is a solid foundation for encoding and other workloads in the data center, at the edge, or in the cloud.

Intel Data Center GPU Flex Series

Leveraging the Xᵉ-HPG microarchitecture, the Intel Data Center GPU Flex Series—with Intel Quick Sync Video—offers flexibility and efficient scaling to develop and run a range of powerful media processing and delivery, and media-based AI inference solutions. With two GPUs each with eight Xᵉ cores and four media engines per GPU, this series delivers outstanding compute density and energy efficiency for media and visual workloads. With support for multi-instance GPU (MIG) technology, the Intel Data Center GPU Flex Series allows multiple workloads to run simultaneously on a single GPU. The GPUs are built for power efficiency to help reduce overall energy footprint in data centers. They are backed by the Intel software and driver ecosystem, ensuring compatibility and performance optimization across various media applications and frameworks.

Intel Quick Sync Video

Intel Quick Sync Video is a hardware transcoder built into many generations of Intel Core processors and into Intel Data Center GPU Flex Series. VisualOn Optimizer easily integrates with Intel Quick Sync Video, dynamically adapting Intel Quick Sync Video transcoding parameters using AI methods to optimize the video data based on each frame. This integration makes VisualOn Optimizer an ideal choice for media encoding operations. Using Optimizer with Intel Quick Sync Video, transcoding for different content delivery methods can run quicker, use less bandwidth, and decrease storage requirements—while enhancing visual quality.

Benchmarks and Testing

Extensive testing by VisualOn shows how Optimizer performs compared to other CAE solutions and how it enhances Intel Quick Sync Video encoder running on Intel CPUs and Intel Data Center GPU Flex Series. All the following benchmarks were conducted by VisualOn.

Comparative Results: Other CAE Solutions

Compared to results from other CAE per-title encoding solutions from the Netflix Technology Blog,[5] VisualOn Optimizer delivers much lower average bitrate (Figure 4) and higher average VMAF (Figure 5).[11]

Comparative Results: 4th Gen Intel Xeon Scalable Processors and Intel Data Center GPU Flex Series with Intel Quick Sync Video

The following benchmarks show comparative results of CPU- and GPU-based encoding running without and with VisualOn Optimizer averaged across different content categories over all Adaptive Bit Rate (ABR) ladder rungs, and for the top rung of an ABR ladder. The benchmark was completed on a dual-socket server running 4th Gen Intel Xeon Scalable processors with the Intel Data Center GPU Flex 140 discrete accelerator.[12] The input test suite is the same as used in the Netflix Technolgy Blog,[5] using an average of results across all content types for the top-rung results.

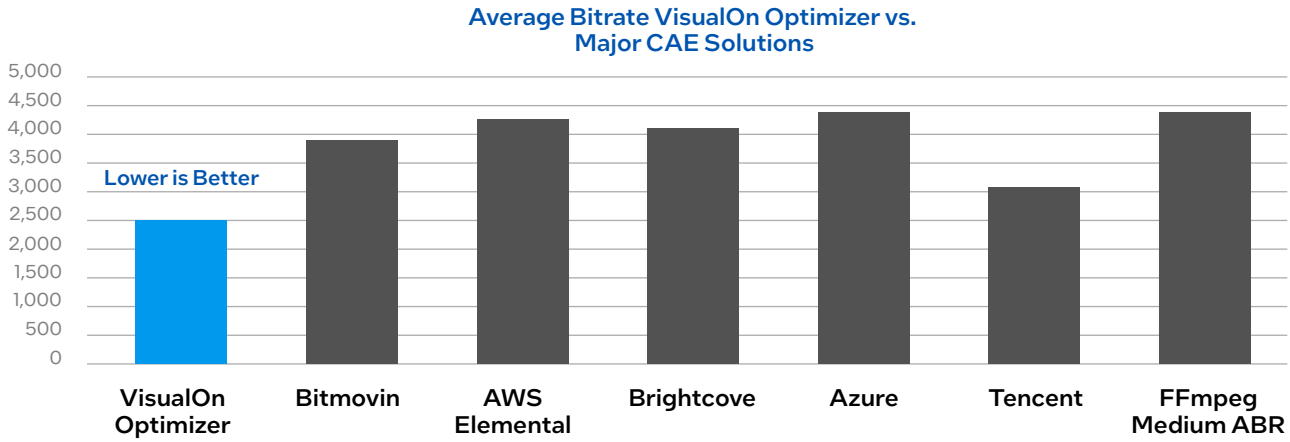

Figure 4. VisualOn Optimizer delivers lower average bitrate compared to other per-title CAE solutions.[11]

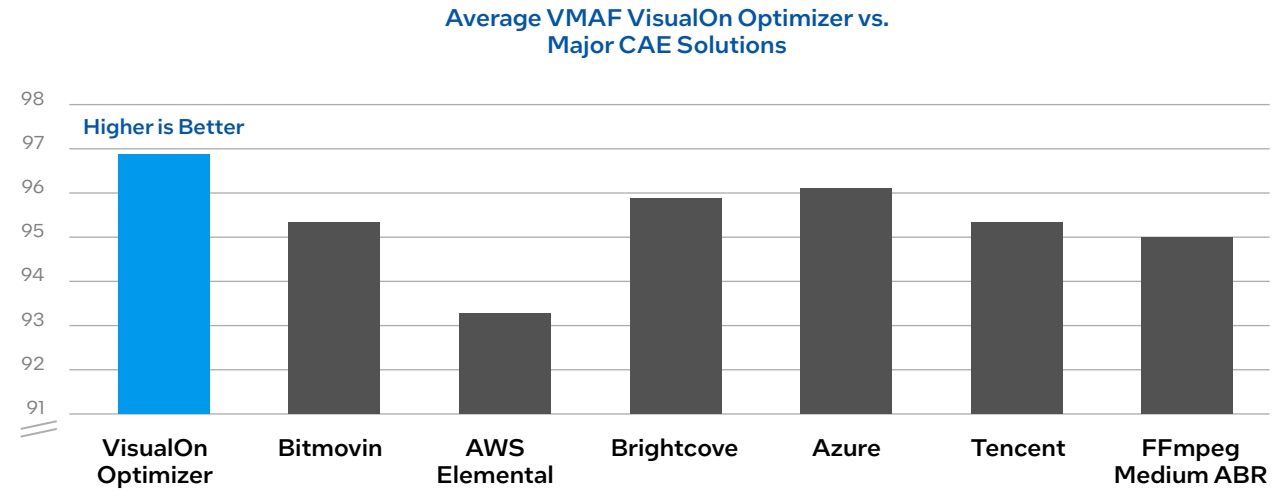

Figure 5. VisualOn Optimizer delivers higher average VMAF compared to other per-title CAE solutions.[11]

Data was collected for software (CPU) and Quick Sync Video hardware (GPU) versions of the AVC, HEVC, and AV1 encoders. The results (Table 1) show the bitrate and VMAF score for encoding using FFmpeg only and FFmpeg with Optimizer. The results reveal that bitrate is reduced by at least 15 percent and up to 40 percent, while VMAF scores generally improve or decrease by negligible amounts.[12]

Table 2 shows the results for the number of maximum parallel encoding sessions that can run in real-time on the "Tears of Steel" 1080p video clip on the Intel-based server without and with VisualOn Optimizer. Data is shown for the encoder running on the CPU alone (CPU), on the GPU alone (GPU), and for a fully loaded system with simultaneous encoding on the CPU and GPU. Optimizer's efficient implementation running on Intel hardware reduces bit rates, allowing jobs to scale easily without impacting the number of streams that can be processed.

Table 1. Encoding results without and with VisualOn Optimizer on 4th Gen Intel Xeon Scalable processors (CPU) and on the Intel Data Center GPU Flex 140 with Quick Sync Video (GPU).[12]

	Hardware	CODEC	FFmpeg		VisualOn Optimizer			
			Bitrate	VMAF	Bitrate	Delta	VMAF	Delta
Overall ABR Rungs	CPU	x264	1,713	68.19	962	-44%	73.59	+5.40
		x265	1,214	66.42	735	-39%	72.60	+6.17
		SVT AV1	774	65.80	654	-16%	64.39	-1.42
	GPU	H.264 QSV	1,651	67.43	1,169	-29%	73.53	+6.10
		H.265 QSV	1,171	65.84	712	-39%	70.18	+4.34
		AV1 QSV	863	63.93	704	-18%	70.49	+7.55
ABR Top Rung (1080p)	CPU	x264	4,432	94.95	2,562	-42%	96.87	+1.91
		x265	3,151	95.10	1,871	-41%	96.77	+1.67
		SVT AV1	2,004	94.43	1,723	-14%	93.07	-1.35
	GPU	H.264 QSV	4,349	95.08	3,023	-30%	96.31	+1.22
		H.265 QSV	3,051	94.95	1,782	-42%	94.79	-0.16
		AV1 QSV	2,320	93.28	1,842	-21%	95.75	+2.47

Table 2. Maximum number of real-time parallel encoding sessions on 1080p content without and with VisualOn Optimizer.[12]

Hardware	CODEC	FFmpeg		VisualOn Optimizer	
		Max Instances	Bitrate (kbps)	Max Instances	Bitrate (kbps)
CPU	x264	51	4,016	53	2,407
	x265	25	3,039	30	1,371
	SVT AV1	12	2,300	15	1,304
GPU	H.264 QSV	45	4,051	46	2,388
	H.265 QSV	48	3,251	48	1,605
	AV1 QSV	50	2,253	50	1,350
CPU + GPU	AVC	40 + 42	4,089 + 4,049	35 + 38	2,795 + 2,404
	HEVC	20 + 42	3,046 + 3,254	18 + 40	1,347 + 1,605
	AV1	10 + 45	2,282 + 2,260	10 + 45	1,278 + 1,228

Summary

Video encoding technologies and solutions have evolved rapidly over the last few years with CAE paving the way for more efficient content delivery. CAE uses AI to analyze content on a frame-by-frame and scene-by-scene basis and adapt encoder parameters, resulting in dramatically reduced bitrates while enhancing visual quality.

Intel is working with the ecosystem that supports media delivery companies to address their challenges, using advanced Intel technologies, processors, and GPUs. VisualOn provides an innovative, AI-powered solution that reduces bitrates up to 40 percent on average, while improving overall viewer experience. VisualOn Optimizer suite is encoder-agnostic and seamlessly integrates into encoding workflows, irrespective of the encoder being used or the infrastructure it runs on—CPUs or GPUs. The Optimizer easily integrates with Intel Quick Sync Video, dramatically reducing bitrate and allowing more encoding sessions on the same platform.

With the variety of infrastructures used across media provider data centers, at the edge, and in the cloud, VisualOn Optimizer combined with Intel Xeon Scalable processors and Intel Data Center GPU Flex Series GPUs can help reduce cost of operations and storage of encoded content through lower bitrate, while enhancing visual quality.

For more information, visit
www.visualon.com/index.php/visualon-optimizer/

[1] As shown in Table 1.

[2] https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2, 12/14/2015 (per title)

[3] https://netflixtechblog.com/more-efficient-mobile-encodes-for-netflix-downloads-625d7b082909, 12/1/2016 (per chunk)

[4] https://netflixtechblog.com/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f, 3/5/2018 (per shot)

[5] https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652, 6/1/2016 (VMAF)

[6] Ozer, Jan. "Report: Cloud-based Per-Title H.264 Encoding Benchmark Report" https://streaminglearningcenter.com/encoding/slc-releases-cloud-based-hevc-and-h-264-per-title-quality-and-opex-reports.html 6/6/2022

[7] https://www.visualon.com/index.php/press/visualon-introduces-first-universal-content-adaptive-encoding-solution-for-video-streaming/

[8] https://www.visualon.com/index.php/press/visualon-wins-2024-nab-show-product-of-the-year-award/

[9] Testing performed by VisualOn on 7/25/2023. Source video: http://ftp.nluug.nl/pub/graphics/blender/demo/movies/ToS/ToS-4k-1920.mov. FFmpeg command line for x264: ffmpeg -i input -b:v 7.8M -maxrate 7.8M -bufsize 15.6M output.mp4, and for x264 with Optimizer: ffmpeg -vo_optimizer -i input -vo_vmaf 98 output.mp4. Full data available at https://www.visualon.com/wp-content/uploads/2024/09/side-by-side-comparison.pdf.

[10] Testing performed by VisualOn on 9/05/2023. Source video: http://ftp.nluug.nl/pub/graphics/blender/demo/movies/ToS/ToS-4k-1920.mov. FFmpeg command line for x264: ffmpeg -i input -b:v 2.3M -maxrate 2.3M -bufsize 4.6M output.mp4, and for x264 with Optimizer: ffmpeg -vo_optimizer -i input output.mp4. Full data available at https://www.visualon.com/wp-content/uploads/2024/09/side-by-side-comparison.pdf.

[11] Based on testing performance by VisualOn on 7/30/2024. System configuration details: Dual socket Intel Xeon Scalable 8480+ (56 cores per socket), 192 GB system memory (16x32GB DDR5 4800), 500GB and 750 GB SSDs (Intel SSDSC2KB48 and Intel SSDSC2BB80), CPU microcode ver. 0x2b0005c0, Intel Turbo Boost Enabled (up to 3.80 GHz.), Hyperthreading enabled, Ubuntu ver 22.04.1 LTS with 5.17.0-107-generic kernel patches. Configuration details and full data available at https://www.visualon.com/wp-content/uploads/2024/09/Optimizer-benchmarks.pdf, configuration in the "H.264 Test Setup", results in the "H.264 top rung comparison" tab, with Optimizer's results extracted from the "x264 results" tab.Results from other encoders are from the report referenced in footnote 6.

[12] Tested by VisualOn on 8/2/2024. System configuration details: Dual socket Intel Xeon Scalable 8480+ (56 cores per socket) with Intel Datacenter Flex 140 GPU, 192 GB system memory (16x32GB DDR5 4800), 500GB and 750 GB SSDs (Intel SSDSC2KB48 and Intel SSDSC2BB80), CPU microcode ver. 0x2b0005c0, Intel Turbo Boost Enabled (up to 3.80 GHz.), Hyperthreading enabled, Intel Data Center GPU Flex 140, Ubuntu ver 22.04.1 LTS with 5.17.0-107-generic kernel patches. For full data details, see https://www.visualon.com/wp-content/uploads/2024/09/Performance-on-Intel-SDP.pdf.

0924/KAC/HB/PDF 361930-001US