AMD

Optimize Live Video Streaming with Accelerated Transcoding

Reduce Latency and Cost with Hardware Accelerated Transcoding in the Cloud

Executive Summary

Video streaming is a fast-growing market. While on-demand video can be effectively served by a variety of technologies, live video has stringent latency requirements. Platforms optimized with hardware accelerated transcoding can meet the computational needs of live streaming without compromising quality. Now with these acceleration technologies available in the cloud, broadcasters and content providers can scale their services seamlessly while reducing total cost of ownership.

Table of Contents

CHAPTER 1

Markets, Trends, and Costs

Video streaming can be broadly broken into two categories – on-demand and live streaming.

On-demand streaming includes TV shows, movies, and social media. This is provided by many companies, including HBO, Spotify, and Netflix.

Live streaming includes video calls, live sports gaming, and esports from companies like Zoom, Disney, and Twitch. The COVID-19 pandemic accelerated demand for live streaming services, peer-to-peer and business-to-business communications–increasing the load on service and content providers as families and workers were locked down at home. COVID-19 moved the growth curve forward, accelerating adoption by several years. Many people experienced an increase in live streaming across a variety of applications including video conferencing, eSports, telemedicine, eCommerce, and distance learning.

In fact, Bluewave Consulting pointed out that leading streaming platforms, including YouTube Live, Facebook Live, Twitch, and others, experienced an upsurge during the outbreak. For instance, Twitch's viewership escalated by 31% in March 2020, as more individuals connected with the platform.

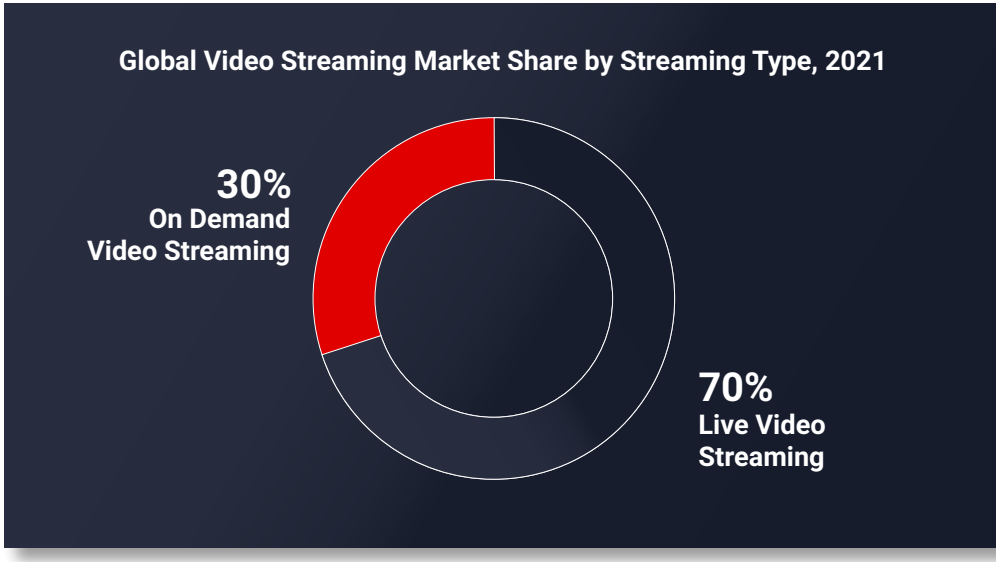


Figure 1: Many live streaming platforms gained surging popularity during the outbreak (Source: Bluewave Consulting, April 2022)

CHAPTER 2

Bandwidth Dominates Operating Cost

While increased demand will drive revenues, live streaming providers face a major challenge: managing higher bandwidth cost. This increases with the volume of streams, the number of users, and the bitrate or quality of the video delivered.

Higher video quality is key to improving quality of experience and viewer retention. However, the higher quality video and growth in users rapidly increases operational costs and can quickly erode profitability.

When it comes to video streaming, reducing the bits used in a video stream leads to direct operational savings. The more compression that can be achieved, the lower the cost to stream them. As shown in the example below, reducing the overall streaming bitrate by 30% can save a large live streaming provider millions of dollars each year. The challenge is how to reduce bitrate without compromising quality.

Encoded Bitrate	Data Per Mth. (TB) Per Stream	Cost Per Mth. ($0.05 per GB)	Monthly Cost @ 100K Streams	Annual Cost (100K Streams)
4Mbps	1.21	$60.48	$6,048,000	**$72,576,000**
2.8Mbps	0.85	$42.34	$4,234,000	**$50,808,000**
			Annual Savings	**$21,768,000**

Figure 2: Reducing streaming bitrate can save a provider millions of dollars each year.

To reduce operating costs, operators must focus on reducing the bandwidth used by the channels. As bitrate savings are multiplied across many channels and viewers, it can lead to significant operational savings.

Reducing bandwidth without compromising quality requires the use of highly advanced video codecs. For example, HEVC provides the same quality as H.264 but at a 35% lower bitrate, or more, depending on the quality and features of the implementation. The downside of HEVC is that it's significantly more computationally intensive than H.264, with the highest compression implementations needing the highest compute resources.

Operators must strike a balance between computational cost vs. bandwidth cost. Even when the choice for lower bandwidth is made, existing unaccelerated servers struggle to provide an efficient platform for advanced codecs such as HEVC and

the newer AV1 codec standard. Additionally, some servers are not capable of performing real-time HEVC encoding, especially at higher resolutions such as ultra-high-definition. This challenge increases further with newer codecs like AV1 and VVC which are four to five times more complex than VP9 or HEVC.

CHAPTER 3

Accelerated Transcoding in the Cloud

Dedicated Media Accelerators

Hardware acceleration is being used more and more in cloud data centers. For many cloud workloads, general-purpose CPUs can be inefficient. Because video can be about 80% of traffic in the cloud, transcoding is a natural target for specialized processing. Hardened video CODECs are a logical approach to achieving performance, power efficiency, and cost effectiveness in video streaming. Dedicated accelerators designed for multimedia applications perform significantly better than general purpose CPUs in terms of scaling channel density to deliver more streams in parallel, reduced power-per-channel to lower operating expenses, and improved latency for live streaming and interactivity.

Essential to a modern CODEC is support for HEVC and AVC transcoding up to 4Kp60 resolution. In addition, a hardened media accelerator should support a 'multi-stream mode' to allow applications to encode and decode multiple streams simultaneously. This parallelism is key to scaling the number of video streams while controlling costs. Instead of needing an array of CPUs to support an increasing number of streams, providers can achieve significantly greater channel density and efficiency via optimized hardware acceleration of data-intensive software functions.

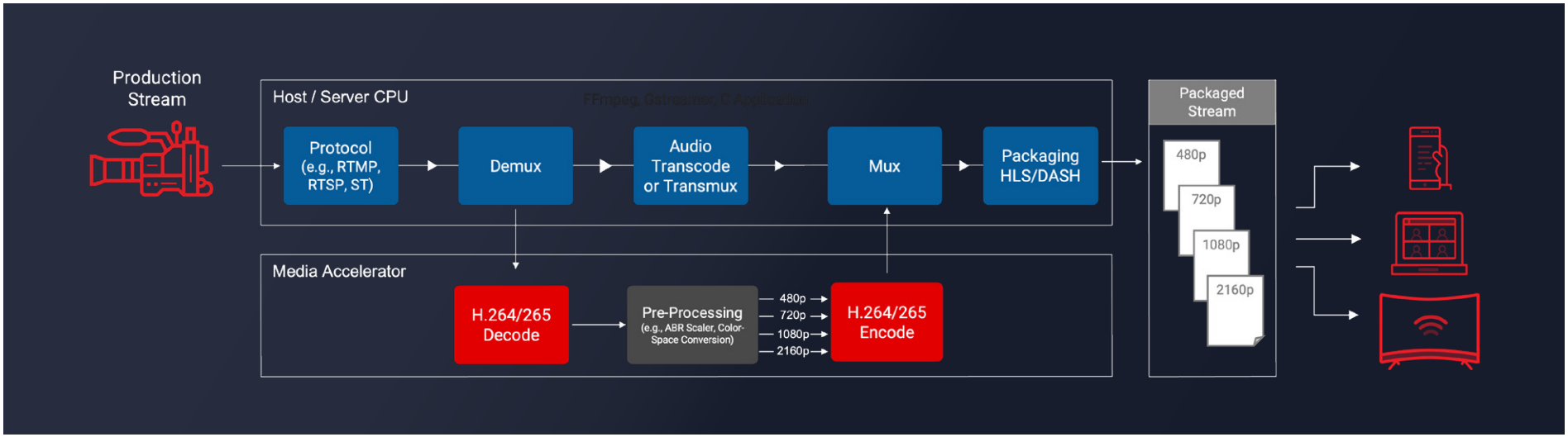


Figure 3: Media Accelerators Perform Transcoding and Additional Functions Across the Video Pipeline

Accelerating Pre- and Post-Processing Functions

To match the performance gains of the codec itself, the solution must also accelerate pre- and post-processing functions across the video pipeline. These functions include adaptive bitrate (ABR) scaling, up-and-down sampling, color space conversion, and more. This allows providers to maintain stream density. The increased compute efficiency, lower latency, and lower power of a dedicated media accelerator can result in greater encoding speeds compared to CPUs at the same video quality.

Media Accelerators vs. GPUs

IT operators, production engineers, and video specialists may consider GPUs as a viable option—given that they also offer a form of hardware-accelerated transcoding. While this approach may be practical for specific use cases, general-purpose GPUs are typically designed for graphics intensive or machine learning workloads,

reserving only a small percentage of the processor to video transcoding. As a result, GPUs are typically less cost-effective, less efficient for video workloads, and draw more power than purpose-built media accelerators, resulting in greater encoding speeds compared to CPUs at the same video quality.

A Familiar Development Environment

Accelerated transcoding must cover more than just the hardware, however. It also needs to encompass a comprehensive set of design and runtime software. The combined hardware and software solution should deliver a platform from which highly flexible, yet efficient systems can be built with design tools familiar to software and application developers. In the case of video application design, common software development kits (SDKs) would need to include FFmpeg and GStreamer—two of the most popular multimedia tool kits. Under the hood, the hardware platform allows the hardware to be purpose-built to the video application.

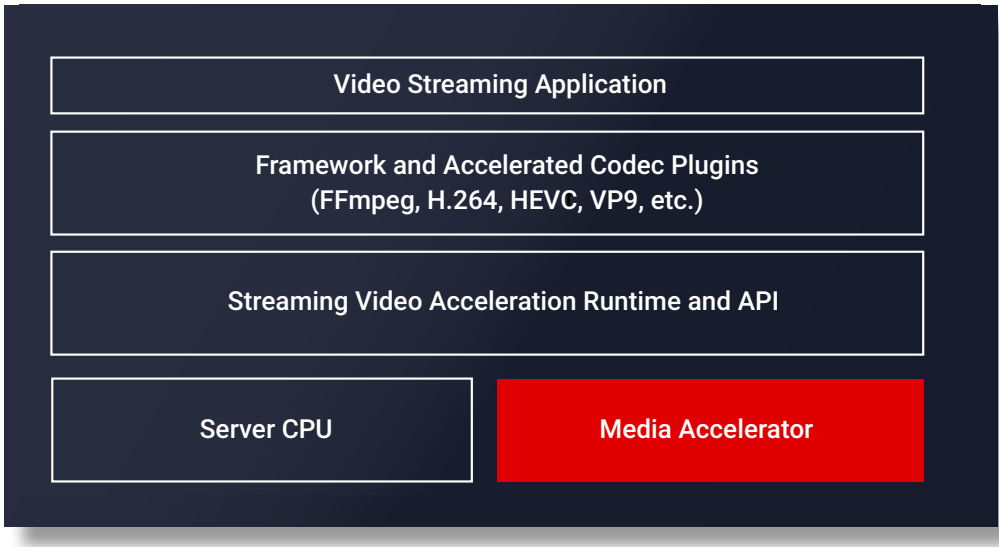
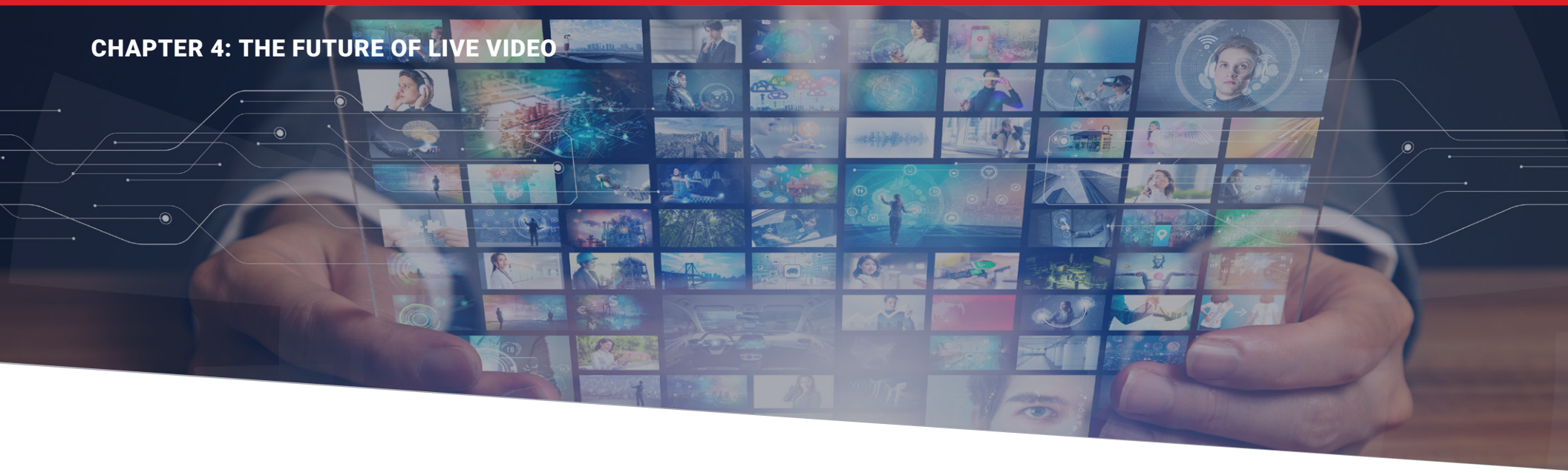


Figure 4: Example Media Accelerator Development Stack (Source: AMD)

Today, accelerated transcoding solutions for video processing are available through cloud services, so broadcasters and content providers can deliver real-time content at scale and achieve the benefits of the underlying technology without needing to know how it works, or program it directly.

CHAPTER 4

The Future of Live Video



Personalization

By its nature, streaming video is free of many of the limitations of traditional broadcast video. With broadcast video (one to many), such as a cable TV channel, there is a single video stream and all viewers see the same content at the same time, including the same advertising. Viewers who join in five minutes after the hour have missed those five minutes.

Streaming video (unicast or one-to-one) allows viewers to personalize their experience by selecting the content they watch at the time they want to watch it, and on the screen or device they want to watch it on. But personalization of video has much more room to grow.

Accelerated transcoding with flexible configurations enable content providers to efficiently adjust encoding of video streams to exactly match demand.

Video is also beginning to replace many in-person experiences. Connected fitness brings a personal trainer into your home gym. Trainers will not only guide clients as they workout but will also be able to control their equipment settings such as speed and incline.

The possibilities for personalization are endless. Media accelerators enable content providers to efficiently adjust encoding of video streams to exactly match demand.

Interactive Social Video

Traditionally, video has been a passive activity: viewers just watch. However, streaming video has the capability of introducing interactivity, engaging viewers in new and innovative ways. Some examples of interactive social video include:

Watch Parties: Participants watch their favorite sports game or movie live with friends, each from their own home.

Gamification: Participants answer real-time trivia questions or cast votes for their favorite contestant in a reality TV show.

Real-Time Betting: Users make live, online micro-bets during a game with the press of a button.

Follower Engagement: Fans interact with one another and the streamer through low-latency chat, often supporting their favorite streamers with recurring subscriptions or one-time donations.

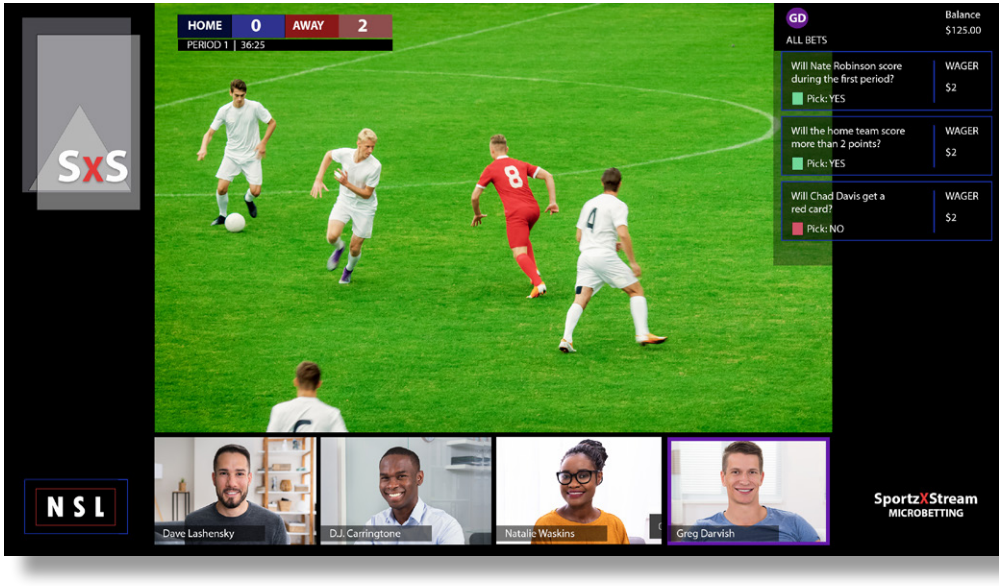


Figure 5: Interactive social video is a growing opportunity for the future of video streaming

Companies like Red5Pro, Skreens, and THEOlive offer live-streaming software and media frameworks that leverage cloud-based accelerated transcoding technology to deliver immersive viewing experiences. To make interactivity viable, the content provider must be able to deliver real-time video with ultra-low latency. In addition, the streaming platform needs to be flexible enough to support both low-cost, high-density video streams (i.e., content) alongside

multiple low-density streams (e.g., viewer streams in a watch party). The platform must be scalable as well, and able to adjust with shifting demand, particularly during a live major sporting event. If latency is too high and the experience is not in line with consumer expectations, it just doesn't work.

Smart Streaming

Today, we are just at the start of benefiting from the efficiencies that accelerated transcoding delivers. For example, the use of artificial intelligence (AI) makes video processing and encoding even more efficient. Consider a typical live gaming stream. Often there are three types of videos in the same frame: synthetic content (the game), natural content (a person talking), and text-based content (the leaderboard). A traditional application will encode all three types of video content in the exact same way, although they are inherently very different.

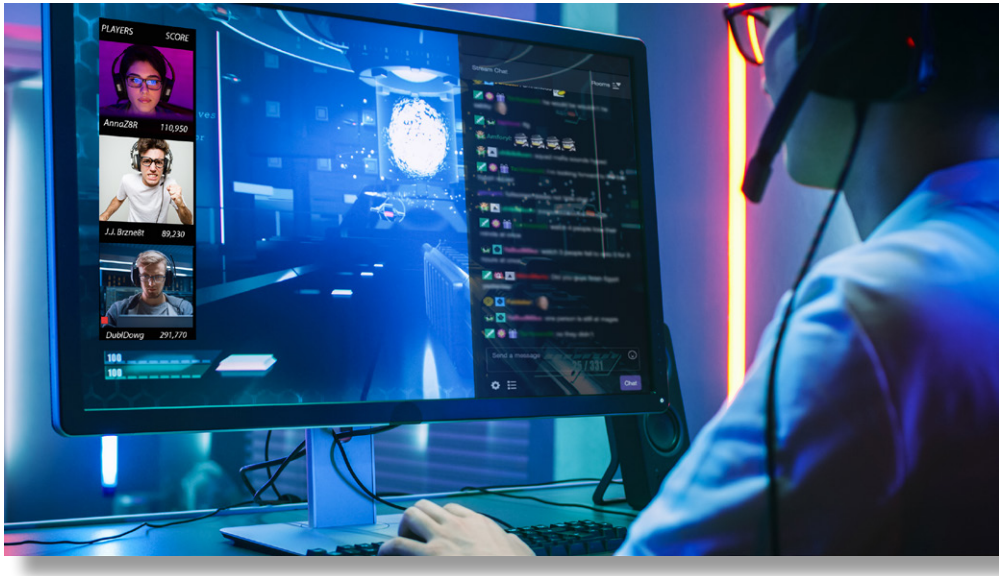
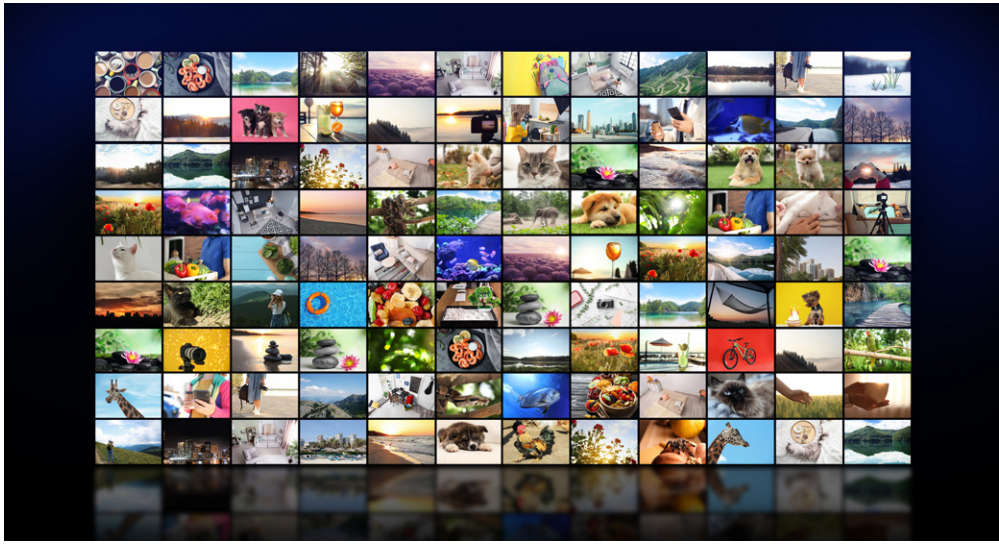


Figure 6: Smart streaming allows the system to identify three different types of video content and dynamically adjust the compute platform to encode each area

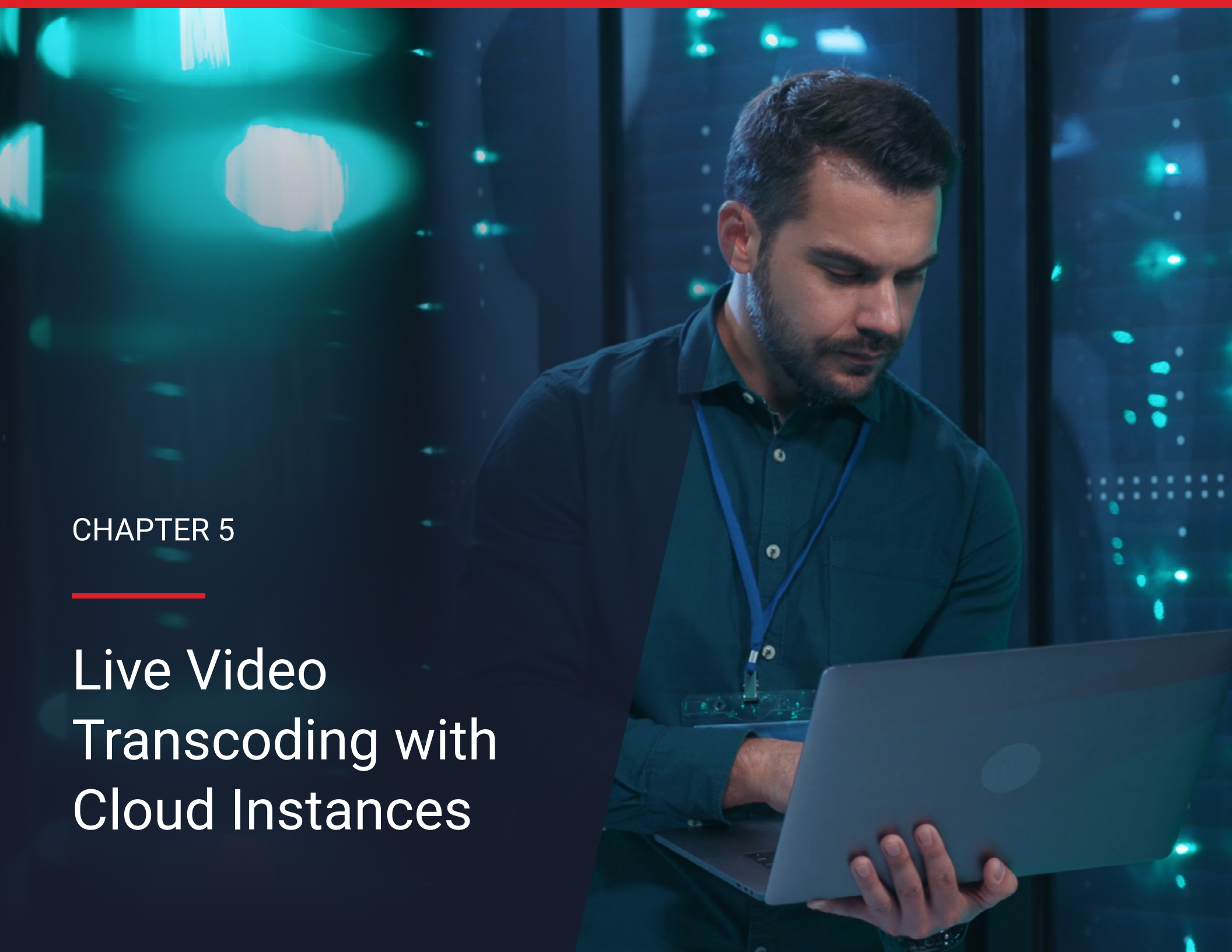
If the application can detect the three different types of video content and encode them appropriately based on their individual characteristics, a significant reduction in bandwidth can be achieved while not negatively impacting visual fidelity. For example, the human face and text leaderboard need to be very high quality and should have minimal quality loss.

A traditional application cannot identify the different needs of the three types of video streams, and therefore can only apply one level of quality and compression across the entire frame. If you want to clearly see the other person and read the scoreboard, you will need a lot of bandwidth.

With smart streaming, the application uses AI to automatically identify the different types of video content. Using AI it can establish “Regions of Interest” (ROI). The application can then use the ROI information to optimize the encoding for each area. The downside of AI inference models being incorporated (and making applications smart) is the added compute demands that are placed on already over-burdened CPUs. Platforms purpose-built for media acceleration, on the other hand, can be used to implement this type of smart application with dedicated processing engines for AI-based ROI detection.



Using media accelerators, video service providers can precisely configure how they want the available bits to be allocated to ensure video quality is focused on the key areas of the video frame. If the characteristics of the content or frame change, then the AI-based system will be able to identify the changes and automatically adapt the encoding parameters. As AI becomes more prevalent throughout the video streaming industry, there will be more innovative approaches to optimize video quality without impacting bandwidth or latency.



CHAPTER 5

Live Video Transcoding with Cloud Instances

Cloud services for accelerated transcoding are now an option for broadcasters and content providers. Amazon EC2 VT1 instances leverage media acceleration technology to deliver low-cost live video streams. Specifically, VT1 instances are powered by up to eight Alveo™ U30 media accelerator cards from AMD and support up to 96 vCPUs.

VT1 instances are optimized for workloads such as live broadcast, video conferencing, and just-in-time transcoding. The smallest vt1.3xlarge instance can deliver up to two 4K UHD streams at 60 frames per second (fps). The largest instance, vt1.24xlarge, can transcode up to 64 simultaneous 1080p60 streams in real time or power faster than real-time media asset transcoding. With a turnkey cloud solution, providers can effectively scale their video content as demand fluctuates.

Lower Cost for Video Transcoding

VT1 instances can deliver up to 30% lower cost per stream compared to Amazon EC2 G4dn GPU-based instances and up to 60% lower cost per stream compared to Amazon EC2 C5 CPU-based instances for live video encoding[1]. Content providers can also reduce their costs for just-in-time transcoding of file-based content for use cases where real-time experience is important.

Low Video Transcoding Latencies

Because they're powered by media accelerators, VT1 instances support ultra-low latencies, reducing the time it takes to transcode the first frame and helping support natural human interactions and experience. In the case of video conferencing, for example, this allows the support for video stream compositing, background blurring, and noise reduction.

> ***"We are adopting Amazon EC2 VT1 instances to cost-efficiently transcode millions of live streams to deliver high streaming quality for viewers all over the world. We chose EC2 VT1 instances because they deliver the stream density and low latency we need without compromising on video compression or visual quality."***
>
> ***- Martin Hess, GM, Interactive Video Service, Amazon, and VP, Video Platform - Twitch Interactive***

AWS Services for an End-to-End Solution

VT1 instances can be used with a range of AWS services to manage, scale, package, and prepare transcoding workloads. Providers can manage and scale their transcoding workloads via Amazon Elastic Container Service (ECS) and Amazon Elastic Kubernetes Service (EKS), store media assets with Amazon S3, and deliver the final content globally using Amazon CloudFront. VT1 instances can also be complemented with a variety of pre-built media services that serve content creation, distribution, and monetization that help broadcasters securely deliver video at a global scale.

[1] https://aws.amazon.com/ec2/instance-types/vt1/

CHAPTER 6

Summary

CHAPTER 6: SUMMARY

The key to reducing operational expenses for live video streaming relies on reducing bandwidth without compromising quality. However, the increased processing needs of modern codecs are pushing the capabilities of unaccelerated solutions.

Media acceleration platforms accessible through the cloud provide the performance, efficiency, and cost-effectiveness per stream to meet the demands of live video workloads at scale. They're optimized for low latency and can be programmed from standard software APIs and frameworks, such as FFmpeg and Gstreamer.

Solutions like the Amazon EC2 VT1 instance make it possible to efficiently process high-density, live streaming workloads while reducing operating costs. A comprehensive, software development kit is also available. The Alveo™ U30 Video SDK allows users to tap into the features of the underlying acceleration hardware.

Figure 7: The Alveo U30 Media Accelerator Card from AMD

The VT1 instance is powered by the AMD Alveo U30 card, offering cost-efficient, multi-stream video transcoding, with support for video streams up to 4K and UHD resolution at 60 fps.

For more information, visit:

www.xilinx.com/livestreaming

Learn More:

AMD Solution for Video Transcoding

Amazon AWS EC2 VT1

Alveo U30 Media Accelerator

Get in touch with us!

media-acceleration@amd.com

© Copyright 2023 Advanced Micro Devices, Inc. All rights reserved. Xilinx, the Xilinx logo, AMD, the AMD Arrow logo, Alveo, Artix, Kintex, Kria, Spartan, Versal, Vitis, Virtex, Vivado, Zynq, and other designated brands included herein are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies. AMBA, AMBA Designer, ARM, ARM1176JZ-S, CoreSight, Cortex, and PrimeCell are trademarks of ARM in the EU and other countries. PCIe, and PCI Express are trademarks of PCI-SIG and used under license.

Printed in the U.S.A. WF#1926700

AMD