

THE AWS SANITY CHECK

A BRUTAL
TCO ANALYSIS
BLEEDING OPEX ON
VIDEO ENCODING

By Mark Donnigan
for NETINT



EXECUTIVE SUMMARY

Page 2

THE RISE OF SPECIALIZED VIDEO INFRASTRUCTURE

Page 4

ECONOMIC MODEL ANALYSIS

Page 5

BEYOND COST SAVINGS TO FUTURE ARCHITECTURE

Page 11

LOW RISK MIGRATION PATH TO VPUs

Page 13

THE STRATEGIC IMPERATIVE

Page 14

REFERENCES

Page 15



EXECUTIVE SUMMARY

The video streaming industry is at a critical inflection point. For years, the default strategy for scaling video processing has been to rely on general-purpose public cloud infrastructure, with Amazon Web Services (AWS) being the dominant choice. Today, that strategy is increasingly obsolete both strategically and financially, as this paper reveals. The computational demands of high-quality video now outpace the efficiencies of generic CPU-based cloud hardware, leading to untenable costs. Precisely why the world's largest streaming platforms (Google's YouTube and Meta's Facebook/Instagram) have already moved to custom video processing silicon, achieving order-of-magnitude improvements in performance and efficiency.

This analysis finds that continuing to run video transcoding on AWS's standard infrastructure is an active choice to accept a fundamentally broken cost model that drains operating margins and stifles the ability to remain competitive.

Video service providers must re-evaluate their cloud strategy with the following considerations:

1. HYPERSCALERS PROVE THE NEW MODEL

Google and Meta have each developed specialized video encoding chips (ASICs), abandoning CPUs/GPUs after finding general-purpose servers economically unviable at scale. Google's custom Argos video (trans)coding unit (VCU) delivered a 20–33× increase in encoding throughput, allowing YouTube to avoid what would have been a requirement for millions of CPU cores. Meta's in-house MSVP (Meta Scalable Video Processor) achieves a 9×–50× leap in encoding efficiency for H.264 and VP9 streams, using a fraction of the power required by a CPU-based approach.

2. AWS ECONOMICS ARE FUNDAMENTALLY UNCOMPETITIVE FOR VIDEO

A detailed Total Cost of Ownership (TCO) analysis reveals that running transcoding on AWS can cost two to four times more per stream than using specialized alternatives. For example, an AWS VT1 instance (Xilinx FPGA accelerator) costs about \$62.72 per 1080p stream per month, whereas a comparable plan using dedicated VPUs (available via Akamai's cloud) costs only \$15.55 per stream – roughly one-quarter the cost. Using standard CPU-based AWS instances, the costs are nearly three times more per stream than the VPU-accelerated solution (see Table 2, page 9). Moreover, AWS's data egress fees (at \$0.09/GB) are eighteen times (18x) higher than those of specialized VPU cloud offerings (around \$0.005/GB), a disparity that can translate to millions of dollars in excess annual costs (see Table 1, page 8).

“

There are two types of companies in the video business. Those that are using video processing ASICs...

and those that will.”

David Ronca
formerly of Netflix and Meta

[Read post](#)

3. THE THREE-PRONGED AWS TAX

- A. AWS's pricing model complexity for media services (e.g., MediaConvert's "normalized minutes") makes cost unpredictable and often inflated. This is effectively a tax on transparency.
- B. AWS's own accelerated transcoding instances (VT1) offer only limited relief but at a premium price. Yet another tax on customer inertia that capitalizes on those hesitant to leave the AWS ecosystem.
- C. AWS's high data transfer charges act as a tax on success. The more popular your content, the more punitive the fees will be. These "AWS taxes" mean that sticking with AWS for video transcoding is far costlier than it appears, once all factors are accounted for.

4. SPECIALIZED HARDWARE TRANSFORMS SAVINGS AND CAPABILITIES

By migrating from general-purpose servers to Video Processing Units (VPUs), video platforms can reduce transcoding infrastructure costs by 70% or more per stream and decrease content delivery costs (e.g., egress) by up to 95%. These savings are not theoretical; one platform reported an \$8.6 million annual OPEX reduction after replacing CPU/GPU-based encoders with VPU-powered servers. Now, intensive codecs, such as HEVC and AV1, have become practical to deploy at scale with VPUs. Likewise, advanced features such as AI-driven encoding enhancements and multi-stream viewing (e.g., interactive multiview sports streams) can be delivered without an exponential spike in compute costs with VPU-based architectures.

5. A LOW-RISK MIGRATION PATH FROM AWS TO VPU

Start with a pilot program on a specialized VPU-powered cloud service (for example, using Akamai's new VPU Accelerated instances) to validate performance, cost savings, and integration approach. By running a side-by-side pilot for 30 days and comparing costs and output quality, organizations can quantify the impact on their specific content. This controlled approach builds internal confidence and data to support a broader transition.

The report urges CTOs and infrastructure leaders to rethink their reliance on general-purpose cloud instances for transcoding. The evidence suggests that maintaining the status quo means inadvertently losing money and falling behind those who adopt more efficient technology.



Ultimately, the analysis concludes that migrating transcoding workloads from legacy AWS infrastructure to specialized platforms supporting VPUs is imperative for any streaming provider seeking to reclaim margins and stay competitive.

THE RISE OF SPECIALIZED VIDEO INFRASTRUCTURE

General-purpose cloud computing is becoming a liability for scaling video streaming. The traditional approach of renting CPU-based virtual machines in the public cloud for video encoding made sense in an earlier era, but today it has become a fundamental mismatch for the workload.

Video transcoding is a task of highly repetitive, mathematically intensive, and massively parallel operations, essentially the opposite of what general-purpose CPUs are designed for. Even with advanced codecs and optimized software, a CPU is often 99% busy with tasks that are mostly irrelevant to the needs of video encoding, given its architecture. Graphics Processing Units (GPUs) offer more parallelism, but they, too, are a compromise, being designed for 3D graphics and AI, rather than video.

Facing this reality, the industry's largest players charted a different path. Over the past few years, hyperscalers such as Google and Meta have invested billions in developing custom video processing ASICs, purpose-built for transcoding. Their motivations were straightforward: at YouTube's scale, relying on CPUs became economically unsustainable. Google disclosed that encoding its 4Kp60 YouTube videos with the H.264 codec on CPUs would require an estimated 7.2 million Intel Skylake cores – a number that soars to over 33 million cores if using the more complex VP9 codec¹. Instead of attempting the impossible (or bankrupting itself on servers), Google's only viable choice was to engineer its own chip. The result was the **"Argos" Video Coding Unit (VCU), a custom encoding accelerator that delivers a 20–33× throughput boost compared to Google's prior optimized CPU-based system.**

Meta is confronting a firehose of video across Facebook, Instagram, and its other services, which is exactly why they built the **Meta Scalable Video Processor (MSVP)**⁴. The MSVP's **performance gains are staggering, with a nine times (9x) increase in throughput for H.264 and a fifty times (50x) increase for VP9 encoding**, all while consuming 10 watts per module.⁴

Even with their virtually unlimited resources and bulk pricing power for CPUs, hyperscalers have determined that general-purpose hardware could not meet their cost-performance requirements. As David Ronca (a key architect of Meta's video infrastructure) bluntly put it, companies who haven't adopted video ASICs are simply "those that will."

Across the industry, forward-thinking streaming services are exploring specialized hardware, while those clinging to the old all-purpose cloud model are facing escalating costs and diminished margins. But, specialized video processing is no longer exclusive domain of Google or Meta. Advances in Video Processing Units (VPUs), led by NETINT, have democratized hyperscaler-level efficiency.

Akamai has partnered with NETINT to offer IaaS (Infrastructure as a Service) VPU-accelerated cloud instances on their network through Linode. This means any streaming platform can access hardware that delivers an order of magnitude more throughput per dollar than CPU-based instances.

In one deployment, a media company that migrated its encoding from CPU/GPU instances to NETINT VPUs achieved an 80% reduction in server count and energy usage, resulting in an \$8.6 million annual

OPEX savings in transcoding costs. The efficiency gap between general-purpose and specialized solutions is simply too large to ignore, and it is rapidly redefining the economic baseline for video infrastructure.

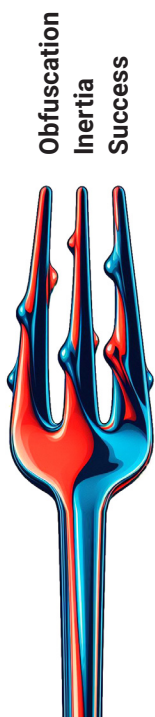
ECONOMIC MODEL ANALYSIS: AWS vs. SPECIALIZED TRANSCODING

For video engineering and operations executives, **Total Cost of Ownership (TCO)** is the ultimate metric for infrastructure decisions. When it comes to running video transcoding workloads on AWS, a superficial examination of cloud pricing can be misleading. The true cost of using AWS is a multifaceted combination of charges. Many of them are hidden or obscured, but together they make AWS far more expensive for this use case than it might appear from a simple on-demand instance price perspective. The AWS tax has three components:

1. **Tax on Obfuscation** (Complex Pricing). AWS's primary transcoding service employs a complex pricing scheme that obscures the actual cost of processing and encoding your content. By billing in abstract "normalized minutes" rather than actual resource usage, AWS makes it nearly impossible to predict or control transcoding costs. This lack of transparency levies additional cost for any customer trying to optimize their budget.
2. **Tax on Inertia** (Overpriced Acceleration). Acknowledging the inefficiency of CPUs for video, AWS offers "media accelerator" instances (the VT1 FPGA-based instances). These specialized instances are priced at a premium for subpar performance. Here, AWS capitalizes on customer reluctance to explore solutions outside its ecosystem.
3. **Tax on Success** (Data Egress Fees). AWS charges notoriously high rates for data transfer out of its cloud. For video providers, egress is one of the largest ongoing expenses, directly scaling with audience size and the number of viewing hours. AWS's egress pricing (often called the "egress tax") means that the more successful your service is (i.e., the more video you deliver), the more you are punished with exponential bandwidth bills.

Understanding these three cost drivers is essential to grasping why AWS's ostensibly convenient solution can bleed a video business dry. Let's examine each "tax" in detail and then present a direct head-to-head cost comparison between AWS and a specialized VPU-based solution.

3-Pronged AWS Tax:



AWS MEDIA CONVERT'S OPAQUE PRICING: THE OBFUSCATION TAX

The true cost of using AWS is a multifaceted combination of charges. Many of them are hidden or obscured, but together they make AWS far more expensive for this use case than it might appear from a simple on-demand instance price perspective.

AWS's flagship file-based transcoding service, **AWS Elemental MediaConvert**, advertises a pay-as-you-go model that appears straightforward – you are charged based on the minutes of video processed. This service bills in units of “Normalized Minutes”, an abstraction that decouples the billed amount from the actual video duration. Instead of charging purely for output duration, MediaConvert applies a series of multipliers based on the content's resolution, frame rate, codec, and advanced settings. Each multiplier inflates the number of “normalized” minutes you are billed for.¹⁴ The effect is that the same 60-second video can accrue many times more than 60 seconds of billable time, depending on how it's encoded.

For example, consider encoding a 60-second 1080p video at 30 fps using one of MediaConvert's higher quality settings (the Professional Tier, Multi-pass HQ mode – a common choice for premium VOD content). Under AWS's formulas, this single 60-second clip would be assessed a 3.5× multiplier (for resolution and multi-pass complexity).¹⁴ AWS would bill it as though it were 210 seconds (3.5 minutes) of content, despite it being only 60 seconds long.

At the entry-level Professional Tier rate of roughly \$0.012 per normalized minute, that one-minute clip costs about \$0.042 to process.¹⁴ Now consider a typical adaptive bitrate (ABR) encoding ladder for streaming with five output renditions (1080p,

720p, 480p, 360p, 240p). Each rendition could have its own multiplier (lower resolutions have smaller multipliers, higher resolutions have larger). In aggregate, those five renditions could easily accumulate 10, 15, or even 20 normalized minutes of processing for just one minute of video input to the transcoder.

This opaque system renders capacity planning uncertain, where even minor adjustments to encoding settings (such as using a higher frame rate or a multi-pass encode) could result in significant, unexpected spikes in the AWS bill. This complexity benefits AWS, not the customer. By obscuring the true relationship between usage and cost, AWS's normalized minute model makes it extremely difficult for streaming companies to optimize encoding workflows for cost.

In stark contrast, specialized transcoding platforms offer simpler and more predictable pricing. For instance, a dedicated VPU-based instance (such as an Akamai cloud instance with NETINT VPUs) has a fixed hourly rate that allows it to process a set number of streams concurrently, regardless of resolution, bitrate settings, or even codec. You pay for the machine, not an ever-shifting metric – meaning there will be no surprises. The savings in predictability and transparency allow engineering to focus on quality and efficiency rather than fighting the billing system and business model.

AWS VT1 ACCELERATORS: THE INERTIA TAX

AWS implicitly acknowledges the limitations of CPUs for video by offering its own line of hardware-accelerated instances, the VT1 family powered by Xilinx Alveo U30 video encoding FPGA cards.⁷ On paper, the VT1 instances are AWS's answer for more efficient transcoding. The existence of VT1 proves that specialized hardware is necessary for modern video workloads. However, a close look at VT1 pricing and performance reveals that AWS is charging a steep premium for marginal gains.

Let's consider AWS's MediaConvert managed service plans versus a VPU-based solution. A direct comparison (for 1080p live transcoding) shows that a solution built on AWS's VT1 instances ends up costing about \$62.72 per stream per month, whereas an equivalent plan using

NETINT Quadra VPUs in the Akamai cloud costs only \$15.55 per stream per month. **In other words, AWS's option is four times more expensive for the same output. This enormous gap means that even if a company is willing to adopt hardware acceleration, doing so within AWS will result in losing up to 75% the cost savings.**

Examining an AWS vt1.3x large instance, which includes one U30 FPGA card, and is priced around \$0.65 per hour on demand (roughly \$475 per month) reveals the non-competitiveness of the offer.¹⁵ A similarly equipped instance in Akamai's cloud (with a more powerful NETINT Quadra T1U VPU card) costs \$0.53 per hour,⁷ meaning AWS is charging a 22% premium for a hardware instance that delivers lower encoding throughput per dollar.

DATA EGRESS FEES: THE SUCCESS TAX

In the media streaming business, bandwidth is not a side consideration; it's one of the primary line items in the operations budget. Every video stream delivered to a viewer incurs data transfer from the cloud to the user. For AWS regions in the US/EU, data transfer out to the internet costs approximately \$0.08–\$0.12 per GB, with a common rate of around \$0.09/GB. This is many times higher than the bandwidth cost on specialized video delivery platforms. By comparison, Akamai's cloud, leveraging its CDN backbone, offers \$0.005/GB for video egress when using its Accelerated VPU instances.⁷

This disparity – roughly a 1,800% difference in per-GB pricing – means that for any sizable streaming operation, egress fees on AWS can dwarf the compute costs. Egress functions as a tax on your success: the more your audience and viewing time grow, the more exorbitant the AWS bill becomes.

The numbers speak for themselves. A moderately successful streaming service delivering approximately 1,000 TB of video per month (approximately 1 PB) would incur over \$1 million per year in egress fees on AWS compared to about \$60,000 per year using the alternative Akamai solution, resulting in a difference of over \$1.02 million annually.

At larger scales, the gap grows into millions of dollars every year. These aren't rounding errors or minor inefficiencies; they are massive drains on profitability. Every dollar spent on excessive egress fees is a dollar not spent on content acquisition, customer experience improvements, or marketing to grow the business. This is a direct hit to EBITDA and to the company's valuation.

Eliminating this tax can instantly save millions for any video service at scale, making a compelling case for migration all on its own.

Table 1
Annual Egress Costs

Various scales of data delivery, comparing AWS vs. the Akamai Accelerated VPU service.

Note: Calculations assume standard AWS data transfer pricing of \$0.09/GB, versus \$0.005/GB for an Akamai Accelerated VPU instance.

			
Monthly Data Delivered	AWS Annual Egress @ \$0.09/GB	AKAMAI Annual Egress @ \$0.005/GB	Annual OPEX Difference
100 TB	\$108,000	\$6,000	\$102,000
500 TB	\$540,000	\$30,000	\$510,000
1 PB (1,000 TB)	\$1,080,000	\$60,000	\$1,020,000
5 PB (5,000 TB)	\$5,400,000	\$300,000	\$5,100,000

TCO SHOWDOWN: SPECIALIZED VPU CLOUD vs. AWS

When all three facets of cost are considered together, the TCO advantage of specialized infrastructure becomes even more striking. Table 2 compares the cost to operate one 1080p video channel (stream) using three different approaches: (a) a VPU-accelerated instance in the Akamai cloud, (b) an AWS VT1 accelerated instance, and (c) a generic CPU-based cloud instance.

This analysis combines instance pricing, the expected transcoding density (i.e., the number of channels each instance can handle), and related costs to yield a normalized cost-per-stream-per-month for each scenario.

Table 2: **Monthly cost-per-stream**

Live 1080p transcoding workload on three infrastructures.

	Plan / Platform	Instance configuration	Max channels per instance	Monthly cost per instance	Cost per stream per month
	Akamai VPU-Accelerated Plan (x2 Small)	2 × NETINT Quadra T1U VPU cards	18 streams	\$488	\$27.11
	AWS VT1 Accelerator Plan	1 × Xilinx U30 FPGA (vt1.3x large)	8 streams	\$475	\$59.38
	Generic CPU Plan	8-core Dedicated CPU instance	2 streams	\$144	\$72.00

As shown in Table 2, the VPU-powered solution is dramatically more cost-efficient per-stream with AWS VT1 setup costs over twice as much as the VPU plan, and a standard CPU-only approach (used by smaller services) costs nearly three times as much.

Running this workload on AWS means choosing to spend 200–300% more than necessary for the same output. Over time, such a cost differential becomes increasingly difficult to justify in a competitive marketplace.

Importantly, this analysis uses conservative assumptions favoring AWS: the VPU instance's cost per stream is based on published "per stream" pricing of \$15.55 in a pay-as-you-go scenario, adjusted here to the flat instance cost (yielding \$27.11) to ensure we're not undercounting the VPU cost⁷.

Cost per stream is a critical metric, but it's equally important to consider value for money – how much work you get done for each dollar spent. Table 3 addresses this by comparing the transcoding throughput per dollar for a VPU instance versus a CPU instance, holding the cloud provider constant.

In this case, we utilize data from tests conducted by Akamai and system integrator G&L Systemhaus, which measured the number of simultaneous 1080p HEVC streams that each instance type can process at a given price point. Both instances cost roughly the same (~\$0.53 vs \$0.52 per hour), so the question is: how many streams (outputs) can each handle?

Table 3:
Performance Per Unit

Number of 1080p HEVC encoding "ladders" per unit cost for a VPU-based instance vs. a CPU-based instance. (An ABR ladder represents one full set of output resolutions for a channel.)

Platform Type 	Hourly Cost	Transcoding Throughput (HEVC 1080p30)	Calculated Output Ladders per \$1.00/Hour
Akamai Cloud – VPU Instance Quadra T1U, x1 Medium	\$0.53	6 × ABR ladders (4 renditions)	11.32
Akamai Cloud – Premium CPU 32 GB v CPU instance	\$0.52	1 × ABR ladder (4 renditions)	1.92

For the same hourly cost, the VPU-powered instance delivers six (6) times the output of the CPU-based instance.

In this real-world test, the VPU system handled six (6) simultaneous 1080p HEVC encoding ladders (each ladder comprising multiple renditions for adaptive streaming). The CPU system could only handle one ladder under the same settings, resulting in 11.3 encoding ladders per \$1.00/hour on the VPU instance, versus 1.9 ladders per \$1.00/hour on the CPU. It's important to note that the economics skew more significantly in favor of VPUs when comparing an advanced codec like AV1 running on cloud-based CPU instances.

This underscores a crucial point: simply comparing cloud instances by their hourly price is misleading unless the work being done is also taken into account. A cheaper instance that accomplishes very little ends up costing more per stream delivered. The VPU solution may have a similar upfront cost to a high-end CPU VM, but it produces significantly more output, resulting in a far lower cost per stream. In short, specialized video hardware doesn't just marginally improve efficiency; it delivers order-of-magnitude leaps that recalibrate the cost structure of streaming operations.

BEYOND COST SAVINGS TO FUTURE ARCHITECTURE

While the immediate incentive to shift away from AWS general-purpose infrastructure is the dramatic OPEX reduction, the strategic implications are equally significant. Adopting a VPU-based transcoding architecture is not merely a way to save money but an investment in the future capabilities and competitiveness of a video platform. By freeing up resources and enabling new technical feats, VPUs lay the groundwork for a more agile and innovative operational architecture.

One major advantage is the ability to deploy advanced video codecs and formats at scale. Historically, formats like HEVC (H.265) or AV1 have promised better compression (and thus lower bandwidth per stream) but at the cost of a much higher computational load.

On a CPU-based transcoding farm, switching from H.264 to AV1 can be prohibitive since the encoding complexity of AV1 is five to ten times higher. This is why many streaming services have been slow to adopt these next-gen codecs despite the bandwidth savings they deliver. A VPU-based approach fundamentally changes this equation.

Specialized video ASICs are extremely efficient at encoding heavy codecs, making it economically feasible to use them widely. If a VPU can encode AV1 ten times more efficiently than a CPU, a service can roll out AV1 streams to improve user experience (higher quality at lower bitrate) and simultaneously reduce egress costs (since each stream uses fewer bits) without a net OPEX increase.

This creates a virtuous cycle where more efficient codecs reduce bandwidth usage, which in turn further reduces the already lower egress bills on the new platform, allowing viewers to enjoy better quality. The old paradigm forced a trade-off between cost and quality; the new paradigm aligns them, allowing services to deliver top-tier quality more affordably.

The capabilities of modern VPUs extend beyond basic transcoding. Many VPU solutions come with on-board AI engines or can be coupled with AI software to enable content-aware encoding. This means the encoding process can incorporate real-time analysis of the video content. For example, automatically detecting faces, text, or important motion regions and allocating additional bits to those areas while compressing less important background areas.⁷

Adopting specialized transcoding hardware future-proofs a streaming operation: not only does it reduce costs today, but it also unlocks headroom for tomorrow's upgrades in quality and interactivity.

Such techniques can further improve visual quality without increasing file size, which is computationally impractical to achieve on CPU pipelines. Likewise, specialized hardware can facilitate entirely new user experiences. A case in point is the emerging trend of multi-view streaming (such as watching multiple camera angles or multiple games simultaneously on one screen).

A feature like YouTube TV's multi-view for sports would require encoding and delivering several synchronized streams to each user. On a general-purpose cloud infrastructure, supporting that at scale would mean a multiplicative explosion of encoding workload and bandwidth – likely cost-prohibitive for most providers. However, the extreme density and low cost of VPU-based encoding make such features feasible for widespread deployment.

A single 1RU VPU server can handle multiple parallel encodes that would have saturated dozens of CPU servers, enabling innovative service offerings (such as personalized multi-camera angles and real-time highlights) without incurring significant costs. In essence, adopting specialized transcoding hardware future-proofs a streaming operation: not only does it reduce costs today, but it also unlocks headroom for tomorrow's upgrades in quality and interactivity.

Finally, there is a competitive strategic aspect to the adoption of VPUs. If your competitors adopt efficient transcoding and achieve a 70% or better cost advantage, they can afford to invest more in content, better technology, undercut pricing, or simply enjoy better margins than you. They can also pivot more aggressively to 4K or HDR content, knowing the infrastructure costs won't spiral out of control.

Moving to specialized video processing based on VPUs isn't just about saving money; it's about reclaiming flexibility and control. It allows engineering leaders to say "yes" to new formats, new features, and growth opportunities that would have been budgetarily off-limits under the old cost structure. Shifting away from AWS's general-purpose cloud is an investment in strategic agility and cost efficiency – positioning those who make the shift as innovation and profit leaders.

A LOW-RISK MIGRATION PATH TO VPUS

Migrating an entire video pipeline away from a legacy cloud environment is a significant undertaking, and prudent technology leaders will want proof of value before committing to a wholesale change. The good news is that validating a VPU-based approach can be done as short as one month and with minimal risk or disruption. The recommended path forward is to run a focused pilot program that puts the TCO claims to the test on your own content. By executing a side-by-side comparison, you can gather data specific to your workloads and build internal confidence for a larger transition:

1. Scope the Pilot

Identify and isolate a small, representative portion of your video workload to use as the test case. This could be a single 24/7 live channel, or a subset of your VOD library. The key is to select content that accurately reflects the range of resolutions, bitrates, and complexity within your broader library.

2. Deploy on a VPU-Powered Platform

Akamai's Cloud and Accelerated VPU instances support container orchestration and Infrastructure-as-Code tools like Kubernetes and Terraform, so you can integrate the new instance into your workflow with minimal effort. For a competent engineering team, this setup is not a bespoke project but a standard deployment task.

3. Run in Parallel

Operate the chosen workload in parallel on both environments for a defined period (e.g., 30 days). That is, continue running the workflow on your existing AWS setup and run the identical content through the new VPU-based instance. Ensure identical output quality and settings for apples-to-apples comparison. This parallel run will generate two sets of results under real-world conditions: one from AWS, and one from the specialized platform.

4. Measure and Compare

At the end of the pilot period, collect the metrics and tally the full costs. Measure the compute hours consumed and actual dollars spent for that workload, on AWS and the VPU side. Model out the data egress costs for the delivered video in both scenarios (even if you didn't actually deliver to end-users from the new platform, you can calculate what it would have cost at the new rates). Compare the all-in costs per stream, as well as any differences in output quality or other performance metrics. With pilot data in hand, you can confidently project the impact of a full migration to create a business case to support shifting your infrastructure strategy.

It is worth noting that migrating off AWS does not mean abandoning AWS entirely. Many organizations adopt a hybrid cloud approach where they continue to use AWS for general services and compute tasks where it makes sense, but offload the video transcoding workload to an optimized environment. By following the pilot-to-migration plan, video streaming companies can reclaim control of their cost structure in a methodical way and capture the upside of specialized hardware.

THE STRATEGIC IMPERATIVE FOR VIDEO LEADERS

Ultimately, the infrastructure decision comes down to leadership vision, innovative spirit, and a willingness to implement technology that is in the growth phase of the adoption cycle. The data has been presented, the cost models are clear, and the strategic advantages of specialized video processing are undeniable. Sticking with the status quo, continuing to run transcoding on AWS's general-purpose (and expensive) framework, is no longer a safe or neutral choice. It has become a high-cost strategy that drains resources and puts a company at a competitive disadvantage.

This reality leads to a final, pointed question that every video executive using AWS for transcoding should be prepared to answer: **Knowing that a technically superior, more scalable, and future-proof alternative exists, one that can cut per-stream compute costs by over 70%, can you confidently justify your current AWS transcoding expenditures to your board and shareholders?**



The answer to that question will define your company's financial health and technical trajectory for the next decade. In the face of such a stark contrast, the mandate for forward-looking leaders is clear: it's time to reclaim your margins, rearchitect your infrastructure, and realign your strategy with the future of video streaming.



REFERENCES

1. Google New Custom Silicon Replaces 10 Million Intel CPUs – Semianalysis (Dylan Patel blog), June 2021.
2. Google's new video chip for YouTube supports AV1 encoding – XDA Developers, Apr 2021.
3. Google-developed "Argos" VCU chip helps YouTube process videos more efficiently – 9to5Google, Apr 2021.
4. MSVP: Meta's first ASIC for video transcoding, Meta AI Blog, May 2023.
5. MSVP: Meta's Scalable Video Processor – At Scale Conference 2022 (technical presentation).
6. Meta unveils first in-house ASIC – a video processor – Converge Digest, May 2022.
7. Akamai+NETINT "Accelerating Video Processing" Webinar – Joint Webinar Presentation, 2025. (Source for performance and cost data on VPU vs. AWS).
8. "Google Argos" Video Transcoder – NETINT Tech Blog, 2021.
9. Video Transcoding with AWS Elemental MediaConvert – AWS Documentation (Service Overview), accessed June 24, 2025.
10. AWS Elemental MediaConvert Pricing – AWS Pricing Page, accessed June 24, 2025.
11. Akamai introduces cloud compute with NETINT VPUs – Investing.com News, Sept 2023.
12. VPU Use Cases (Case Studies) – NETINT Technologies, accessed 2025.
13. AWS announces VT1 instances for video transcoding – AWS News Blog, May 2021.
14. AWS data transfer pricing (Bandwidth Pricing) – Amazon Web Services, accessed 2025.
15. AWS EC2 VT1 instance on-demand pricing (vt1.3xlarge, 1xU30 FPGA ≈ \$0.65/hour) – Amazon EC2 Pricing, 2025.