

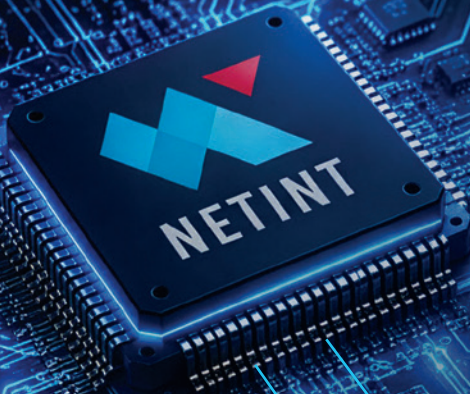
VP

ECO

SYS

TEM

IBC 26



+225,000

VPUs sold globally

1.1 trillion

minutes encoded

*Celebrating new
milestones, and we're
just getting started.*

VPU ECOSYSTEM MAGAZINE

Table of Contents

Introduction - Message from Mark Donnigan	5
Ecosystem Purpose - Partner Directory	6
The New Video Capacity Problem - NETINT	8
Three Ways to Add Video Capacity in a Constrained Market - NETINT	12
VPU-Powered Capacity as a Service - NETINT	16
CASE STUDY: Zapping - Transforming Live Streaming TV, One Rack at a Time	20
Get More From the Servers You Already Own - NETINT	26
Build Dedicated VPU-Based Video Infrastructure - NETINT	30
Field-Proven Methodology for Deploying NETINT Quadra VPUs - NETINT	36
The State of Video Encoding Industry Report Summary - NETINT	42
1 - <i>The GPU Monoculture is Cracking</i>	46
2 - <i>The Dual Mandate Replaced the Old Tradeoff</i>	47
3 - <i>AV1 Is Not a Future Codec</i>	48
4 - <i>AI Has Crossed from Experiment to Infrastructure</i>	49
5 - <i>The Binding Constraint Is Organizational, Not Technical</i>	50
6 - <i>The TCO Blind Spot Is the Industry's Biggest Vulnerability</i>	51
7 - <i>At Scale, Hybrid Wins. Hardware-Only Loses</i>	52
8 - <i>Edge is a Developing Market Waiting for a Reason</i>	53
9 - <i>Precision Engineers: Small in Number, Yet Large in Influence</i>	54
10 - <i>Four Buying Personalities</i>	55
When Economics Break: A Framework for Escaping the Video Compute Trap - Akamai	59
Maximizing Encoding Efficiency in Live Streaming Workflows - Wowza	65
VPU Capacity Without a Hardware Cycle - NetActuate	70
Reliable Live Video Meets Efficient Processing - Zixi	75
Scaling Performance: Critical Video Workloads - i3D.net	79
Beyond General-Purpose Compute - Advantech	83
The Next Layer of Video Efficiency - V-Nova	89
Test VPUs in the Cloud Today - Cloud Pilot Program	94

The market is not waiting
for a better encoder.

It's waiting for infrastructure
that removes friction and
never ending cost increases
without trading off quality.



A MESSAGE FROM MARK DONNIGAN

The VPU did not arrive as a product pitch.

It arrived as an answer to an economics problem.

I have been in this industry long enough to know the difference between a technology trend and a structural shift. Trends get talked about at conferences. Structural shifts get built into infrastructure budgets quietly at first, then all at once.

I have watched video infrastructure teams make the same accommodation for two decades: take whatever compute was available, add a software layer on top, and absorb the cost.

That worked when workloads were modest. It stopped working when video became the majority of internet traffic and the cost of the accommodation became the thing threatening the business model.

What follows is what happens when purpose-built silicon meets the organizations operating at the edge of what general-purpose compute can sustain.

Each partner in this publication arrived at the same conclusions from different directions. That convergence is not marketing. It is signal.

The organizations that recognize this shift earliest will not just reduce their current costs. They will build a structural advantage that compounds.

That is what this publication exists to accelerate.

Mark Donnigan

Chief Marketing Officer
NETINT Technologies

Why the VPU

IT EXISTS BECAUSE

No one deploys a chip on its own. A VPU takes over one stage of the video pipeline: encoding. Ingest, packaging and delivery run on other companies' platforms.

A deployment is only complete when the VPU's encoded output is connected to the ingest, packaging and delivery systems on either side of it.

IT IS MADE UP OF

19 companies in four groups: CDNs and platforms, servers and systems, video software, and integrators and services.

Every one of them runs VPUs in production or builds products on them. There is no formal program to join. Membership means the integration exists and runs in production. We promote the network because it works.

IT IS ESSENTIAL BECAUSE

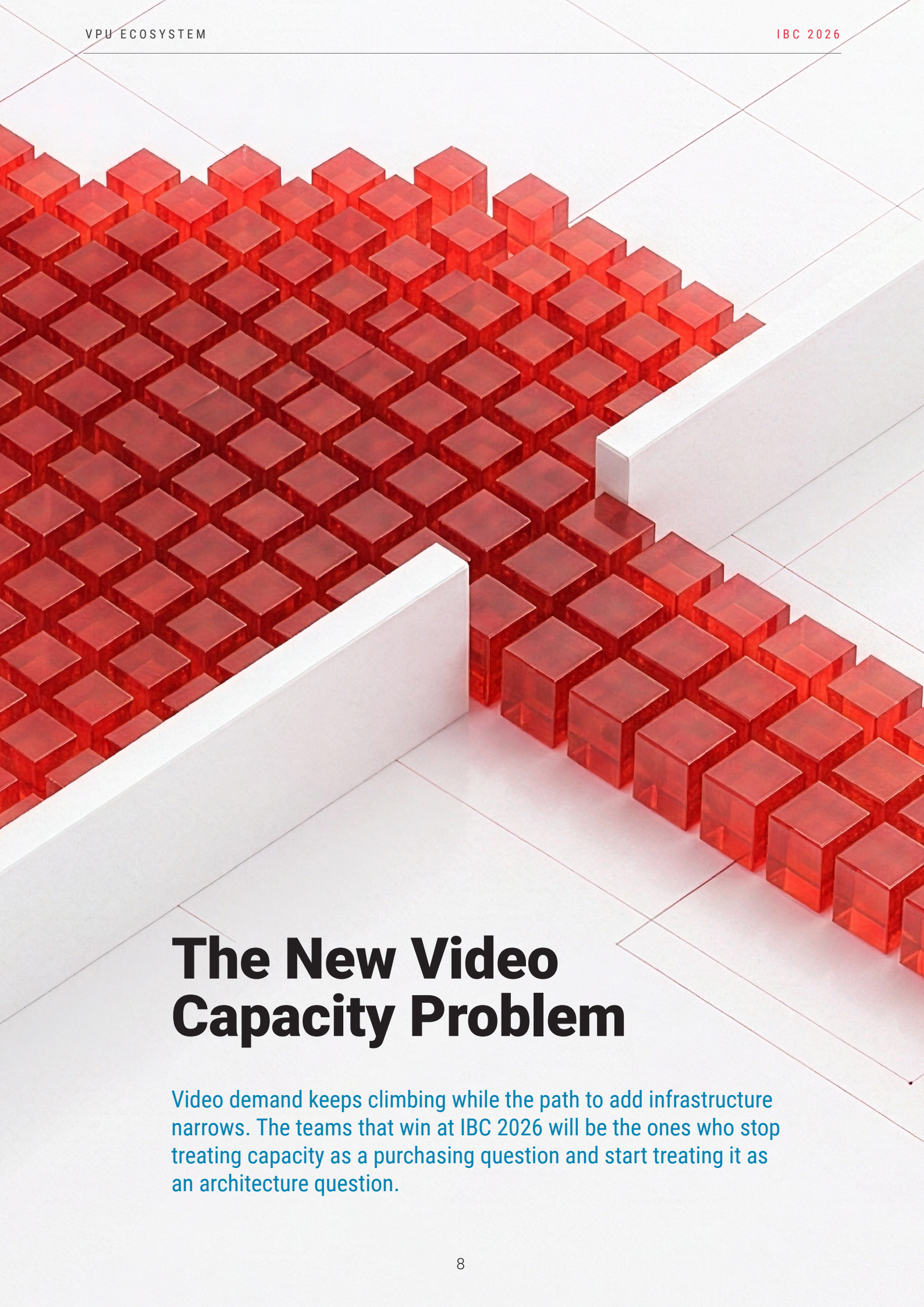
Every integration on this page is validated and already running, so you are not the first one through. When trouble occurs, the partner will contact us and we troubleshoot together as a team to find the root cause.

NETINT field application engineers often work with the partner and their end customer to root cause issues. We are more concerned with addressing these issues than passing them onto someone else.

Ecosystem exists.

				
	INGEST	VPU STAGE ENCODE	PACKAGE	DELIVER
CDN & PLATFORM				
Akamai Cloud				
CDN77				
i3D.net				
Jet-Stream				
NetActuate				
Zixi				
SERVERS AND SYSTEMS				
Advantech				
Dell Technologies				
VIDEO SOFTWARE				
Capella Systems				
Cires21				
EasyTools				
Net Insight				
Norsk				
Scalstrm				
V-Nova				
Wowza				
INTEGRATORS AND SERVICES				
Arcadian				
G&L Systemhaus				
Misao Network				

The Ecosystem exists to take the unknowns out of a VPU deployment before you start one.

An abstract graphic featuring a grid of red, translucent cubes arranged in a pattern that recedes into the distance. Two white, rectangular blocks or walls are positioned in the foreground, creating a sense of depth and perspective. The background is a light gray with faint red grid lines.

The New Video Capacity Problem

Video demand keeps climbing while the path to add infrastructure narrows. The teams that win at IBC 2026 will be the ones who stop treating capacity as a purchasing question and start treating it as an architecture question.

Two curves moving in opposite directions.

For most of the last decade, adding video capacity was a procurement exercise. Forecast the growth, buy the servers, rack them, and the economics took care of themselves. That model still exists, but it no longer fits the market every video team now operates in.

The pressure is not coming from one place. More live channels, more user-generated content, more short-form, more codecs, more renditions of every asset, and more cloud transcoding all push encoding

demand up at once. Live makes it harder still, because the processing has to happen in real time and the event calendar does not move.

At the same time, the supply side has tightened in ways that have little to do with video. The AI build-out has pulled servers, memory, power, and data center planning toward high-density GPU systems. Even teams that will never buy an AI server are now buying in the same constrained market.

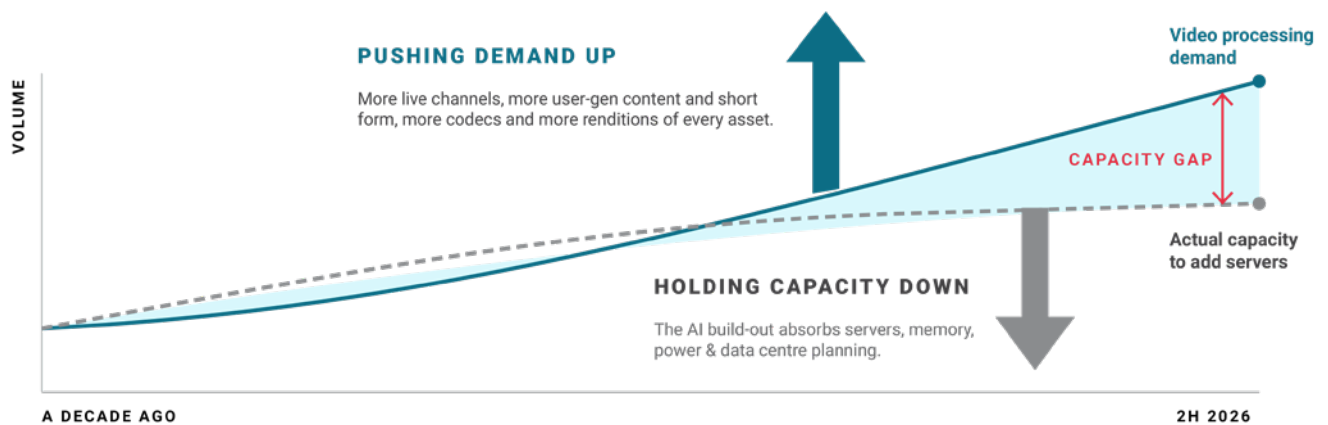
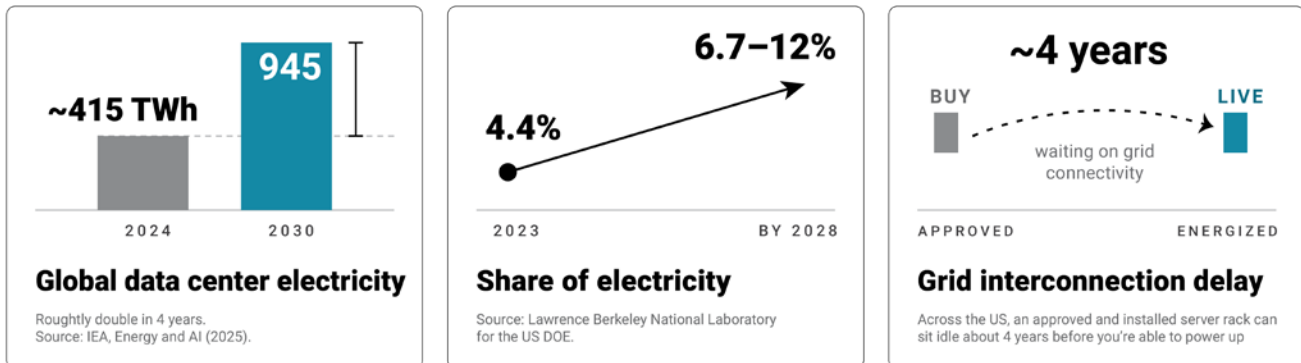


Figure 1. Video processing demand keeps rising while the practical capacity to add general-purpose servers flattens.

The constraint moved from compute to the grid.



The clearest signal is power. A U.S. Department of Energy study from Lawrence Berkeley National Laboratory found that data centers consumed about 4.4 percent of U.S. electricity in 2023, and projected that share to reach somewhere between 6.7 and 12 percent by 2028. The International Energy Agency reaches a similar conclusion globally, with data center consumption set to roughly double from about 415 terawatt-hours in 2024 to around 945 by 2030.

Figure 2. The grid is now the constraint. Data center electricity demand is set to roughly double this decade, and AI is the named driver in every major forecast.

Capacity that exists on paper is not the same as capacity you can energize. Industry analysts now cite grid interconnection waits of roughly four years across much of the United States, a figure corroborated by Berkeley Lab and the IEA. For a video team, that is not a distant macro statistic. It is the reason a new rack of servers can be approved and still sit idle.

The old model creates new bottlenecks.

Server-based expansion now runs into four walls. Procurement can be slow as component allocation and budget cycles stretch out. Even when hardware arrives, power, cooling, and rack space can limit where it can be deployed. GPU supply is its own constraint: by early 2026, analysts at SemiAnalysis described on-demand GPU rental as sold out across all types, with memory suppliers reporting HBM capacity booked through the year.

CPU-only encoding compounds the problem. General-purpose processors are flexible, but at scale they consume power and rack space quickly, and the newest codecs make this worse rather than better. A peer-reviewed IEEE study found that software encoding with AV1 or VVC can be tens to hundreds of times more compute-intensive than H.264. The codecs that save bandwidth cost far more to produce in software, which is exactly the gap purpose-built hardware closes.

So the better question is not what can encode video. It is what can encode video efficiently, predictably, and at the density the business actually requires.

Why purpose-built video processing matters now.

A video processing unit is purpose-built for video workloads. NETINT's Quadra VPUs use ASIC-based acceleration designed for high-density, low-power encoding and transcoding rather than relying on general-purpose compute. In benchmarking conducted by Akamai with Cires21, a NETINT Quadra T1U delivered roughly 4.7x greater energy efficiency than an NVIDIA RTX 4000 Ada GPU on an adaptive-bitrate workload, after normalizing for throughput. The VPU sustained six ABR jobs while drawing 13 watts; the GPU sustained eight while drawing 82 watts.

That efficiency is not only an electricity-bill story. It changes rack density, deployment timelines, capital planning, and how much capacity a team can stand up before the next infrastructure condition has to be solved. In a constrained market, efficiency is optionality.

The platforms that win will not be the ones that buy the most servers. They will be the ones that find smarter ways to process more video from where they already are.

The answer is not one path.

There is no single fix, because there is no single bottleneck. Some teams need capacity before a build can finish, and are better served by hosted VPU capacity through an infrastructure partner. Some already own servers that are underused or aging, and can evaluate them for VPU acceleration once power, thermals, and PCIe are validated. Others need owned infrastructure built specifically for video, and should design dedicated VPU systems from the start.

This is why the ecosystem matters. NETINT provides the VPU foundation. Partners provide the paths to access it, build with it, validate it, and connect it to real workflows. At IBC 2026, the NETINT VPU Ecosystem Pavilion is organized around exactly those three paths.

The most useful conversation at the show will not begin with how many servers a team can buy. It will begin with one specific constraint, and a practical question: what is the most efficient way to add video capacity from where we are today?

Sources & further reading

1. Lawrence Berkeley National Laboratory, 2024 U.S. Data Center Energy Usage Report (DOE-commissioned, Dec 2024). <https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf>
2. International Energy Agency, Energy and AI (2025). <https://www.iea.org/reports/energy-and-ai/executive-summary>
3. SemiAnalysis, The Great GPU Shortage (Apr 2, 2026). <https://newsletter.semianalysis.com/p/the-great-gpu-shortage-rental-capacity>
4. IEEE, Energy Consumption and Carbon Emissions of Modern Software Video Encoders (2023). <https://ieeexplore.ieee.org/document/10374545/>
5. Akamai, Benchmarking VPUs and GPUs for Media Workloads (Aug 19, 2025). <https://www.akamai.com/blog/developers/benchmarking-vpus-and-gpus-for-media-workloads>

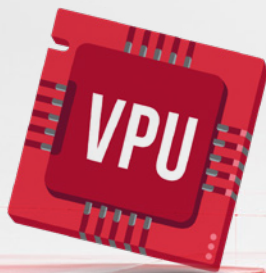
Three Ways to Add Video Capacity in a Constrained Market

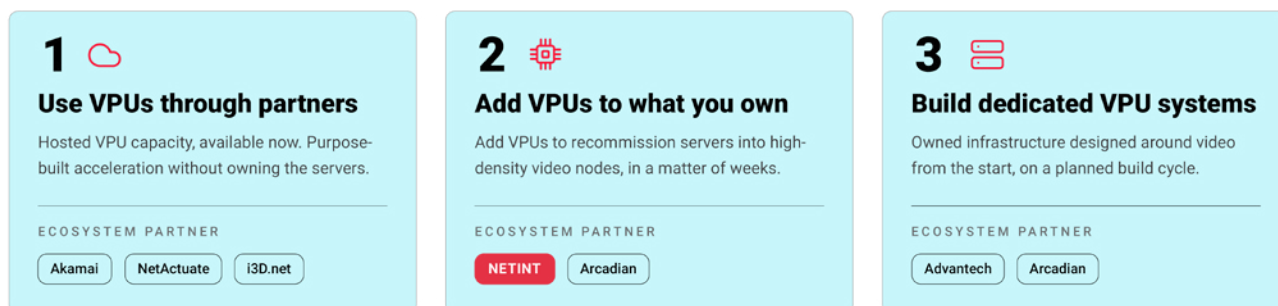
When new servers are slow, expensive, or simply unavailable, the right move depends entirely on the bottleneck you actually have. NETINT's VPU Ecosystem is organized around three practical paths, not one product pitch.

Video platforms need more processing capacity and the way to add it changed. The familiar pattern of forecast, buy, deploy still works, but it is no longer the single answer and certainly not the fastest one.

Power, cooling, rack density, procurement timing, and component allocation now shape infrastructure decisions as much as raw compute requirements. So the practical question is how to add capacity starting from where a team is today.

That reframing matters because it changes conversation from a single purchase to a choice among deployment models. At IBC 2026, the NETINT VPU Ecosystem Pavilion is built around a few of them.





1. Use VPU capacity through partners.

Some platforms need capacity before a traditional build can finish. Others do not want to own and operate every server their growth requires. For them, partner-hosted capacity is a practical option that still delivers purpose-built video acceleration without the full burden of ownership.

This path fits teams facing procurement delays, limited data center space, power or cooling limits, regional expansion, or unpredictable traffic peaks. In the pavilion it maps most directly to NetActuate, which represents distributed infrastructure, and i3D, which brings performance infrastructure for latency-sensitive workloads. The message is simple: you do not always need to own every server to access more video capacity.

It's especially relevant for cloud and CDN transcoding, live streaming, and regional video services. The goal is not to replace owned infrastructure in every case. It is to give video teams another way to scale when traditional expansion is slow, expensive or impractical, while still getting the benefit of purpose-built video acceleration.

2. Get more from the infrastructure you have.

Not every platform starts from zero. Many organizations hold servers that are underused, aging, or no longer ideal for their original purpose. In a constrained market, those assets deserve a second look before they are written off.

This is not a blanket endorsement of old hardware. A server has to be evaluated for PCIe layout and lanes, airflow, power headroom, thermals under sustained load, BIOS settings, operating system support, drivers, and workflow integration. When a system passes, a VPU can turn it into a high-density video node. This path is NETINT-led because it depends on VPU capability and technical validation, and it creates a natural role for a systems integration partner such as Arcadian.

3. Build dedicated VPU-based systems.

Some customers need owned infrastructure built specifically for video. For them, the right answer is neither hosted capacity nor retrofitting, but dedicated systems designed around predictable encoding capacity, high stream density, lower power per stream, and clear operational ownership.

Dedicated infrastructure is more than the accelerator card. Server design, air flow, I/O, PCIe layout, power delivery, form factor, networking, and software all shape the outcome. This is where a hardware partner like Advantech fits, and a software integration partner like Arcadian can help plan, validate, and roll out systems into production.

**More video capacity.
Three ways to get it.**

**Use it through partners,
add it to infrastructure
you already have, or
build dedicated VPU
systems.**

The right path depends on the bottleneck.

The three paths are not separate campaigns. They are different answers to the same question: how do we add video capacity in the most practical way? The first real step is naming the bottleneck honestly.

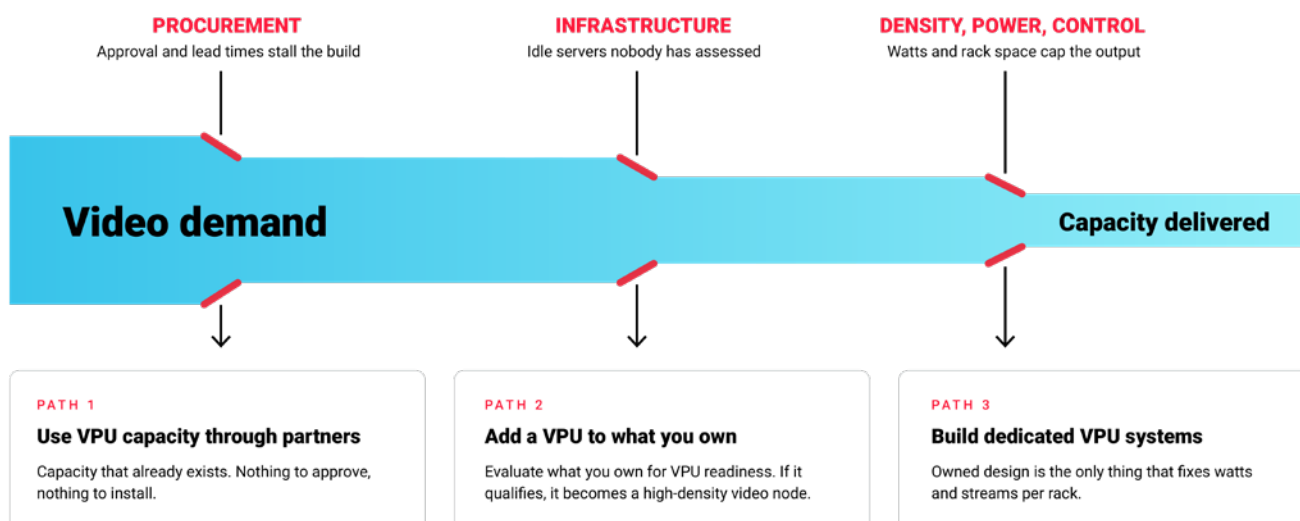
If the bottleneck is procurement, hosted capacity may be the fastest path. If it is underused infrastructure, evaluate existing servers for VPU readiness. If it is density, power, or control, dedicated systems are likely the strongest option. If it is workflow complexity, a transport, streaming, or integration partner becomes essential.

Naming the wrong bottleneck is expensive. A team that buys dedicated systems to solve a procurement deadline waits just as long as before. A team that rents hosted capacity to solve a control or data-residency requirement inherits a model it cannot fully govern. The framework is valuable precisely because it forces the diagnosis before the purchase.

That is also why the pavilion includes more than one kind of partner. NetActuate and i3D support capacity through partners. Advantech supports dedicated systems. Arcadian supports assessment, integration, and deployment. Zixi brings live transport reliability, Wowza brings streaming workflow context, and V-Nova brings the future compression-efficiency layer. NETINT provides the VPU foundation, and the ecosystem provides the paths to use it.

The most productive way to start is not with a broad infrastructure roadmap or a product comparison. It is with one specific problem: a maxed-out server fleet, an approval that will not move fast enough, idle hardware of uncertain value, or a live workflow outgrowing its power budget. Bring one bottleneck, and the right VPU path becomes far easier to find.

Figure 2. Bring one bottleneck, get one path. *The most useful IBC conversation starts with a single, concrete constraint.*



Sources & further reading

1. NETINT, VPU Ecosystem (partner framework). <https://netint.com/ecosystem/>
2. Akamai, Benchmarking VPUs and GPUs for Media Workloads (Aug 19, 2025). <https://www.akamai.com/blog/developers/benchmarking-vpus-and-gpus-for-media-workloads>
3. SemiAnalysis, The Great GPU Shortage (Apr 2, 2026). <https://newsletter.semianalysis.com/p/the-great-gpu-shortage-rental-capacity>

VPU-Powered Capacity as a Service



Add encoding capacity without owning every server. When demand has a fixed date, time-to-capacity becomes the metric that matters most.

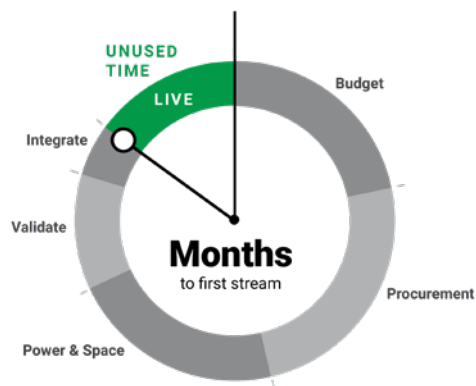
Video engineers usually know they need more encoding capacity well before they know how they will get it. The workload is growing, the live calendar is fixed, the transcoding queue is lengthening, and the cloud bill is climbing. The business still needs capacity.

For some platforms, the practical answer is not to buy, rack, power, and operate more servers right away. It is to access the efficiency of NETINT VPUs through an infrastructure partner, without owning every layer of the deployment.

The real problem: Capacity is needed before infrastructure is ready.

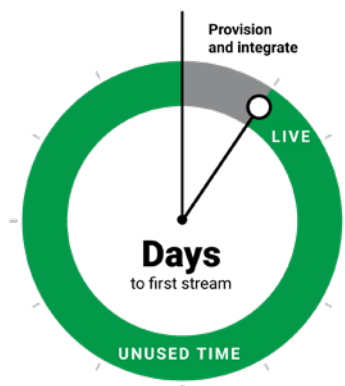
Traditional planning assumes time. Time to approve budget, procure servers, install and validate them, secure power and space, and integrate everything into the workflow. Video does not always wait for that sequence.

Live events happen on fixed dates. Channels launch, traffic spikes, regions change, and a CPU-based pipeline reaches its limit. When that happens, the question becomes direct: how do we add capacity now, without turning it into a full infrastructure project?



Own the build

Months of the year consumed executing five stages before a single stream runs.



Capacity as a service

The partner already owns the rack, solution available in a matter of days.

Figure 1. Two clocks.

An owned build runs through budget, procurement, power, and validation; a service model gets a team to a real pipeline in days.

What this model changes.

VPU-powered capacity as a service separates the need for video processing from the burden of owning and operating every server. The customer still needs encoding or transcoding, but the infrastructure can be delivered through a partner environment, whether hosted, regional, bare metal, or managed.

Convenience is a bonus. The real value is access to video-specific processing without first solving every internal infrastructure condition. For the customer, that means faster deployment, less upfront hardware commitment, and more flexibility in where capacity sits.

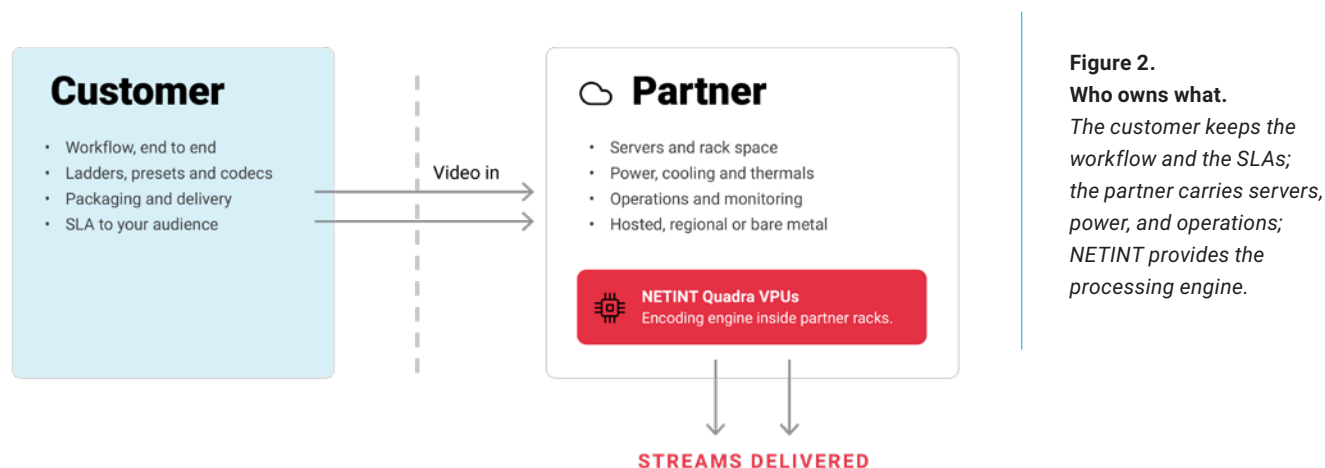
For the partner, VPUs improve the economics of offering video processing. More streams per server, lower power per stream, and higher density make hosted video infrastructure more practical to operate and price.

Why VPUs make the economics work.

A hosted model only works if the underlying economics make sense. If each stream consumes too much power, rack space, or general-purpose compute, the service becomes hard to scale for both provider and customer.

NETINT VPUs are designed specifically for video. They let infrastructure partners offer high-density encoding and transcoding at lower power than CPU-only approaches. Akamai, the first cloud provider to offer VPUs, runs NETINT Quadra T1U accelerators behind instances that start around 280 dollars per month, with each card handling roughly 32 live 1080p30 streams at peak power near 13 watts. That is what video-specific infrastructure looks like when it is designed around video from the beginning.

Treating video as a general-purpose compute workload is what makes hosted transcoding expensive. When the infrastructure is designed around video instead, predictable cost, density, and power efficiency follow, which is what lets a provider price the service and a customer trust the economics. The same purpose-built silicon that lowers a buyer's owned-infrastructure cost also improves the math for the partner offering it as a service.



Where this fits.

This model is for customers who need capacity now but cannot wait for a full purchase, deployment, and validation cycle. Good candidates include platforms expanding into new regions, live providers with peak needs, cloud and CDN transcoding workflows, interactive and gaming video, and teams trying to reduce dependence on expensive CPU-based processing.

The common thread is straightforward. The customer needs more video processing, and the fastest practical path runs through a partner that already operates the infrastructure. In the pavilion, that path connects most directly to NetActuate for distributed reach and i3D for high-performance, latency-sensitive workloads.

What customers should ask.

The first question is not whether capacity as a service beats owned infrastructure. It is whether it is the right model for the current bottleneck. Teams should ask where they need capacity, how quickly, whether demand is steady or peak, whether regional placement matters, and what they are really trying to reduce: cost, power, deployment time, or operational burden.

If the answers point toward speed, flexibility, regional reach, or reduced ownership, a partner-powered model is likely the right path. If they point toward strict control, custom integration, or steady high-volume demand, owned infrastructure may still win. The model is a tool for a situation, not a verdict on owned versus hosted in general.

VPU-powered capacity as a service is not about replacing every owned deployment. Some customers will still build their own systems, some will reuse existing servers, and some will use a partner model because it gets them to capacity faster. At IBC 2026 the central point is simply that the choice now exists, on purpose-built video silicon, wherever the workload needs to run.

You do not always need to own every server to access more video capacity.

The value is having the choice.

Sources & further reading

1. Akamai, Akamai First to Offer VPUs in the Cloud (Mar 27, 2025). <https://www.akamai.com/newsroom/press-release/akamai-first-to-offer-vpus-in-the-cloud>
2. Akamai TechDocs, Accelerated Compute Instances (NETINT Quadra T1U). <https://techdocs.akamai.com/cloud-computing/docs/accelerated-compute-instances>
3. NETINT, VPU Ecosystem. <https://netint.com/ecosystem/>

CASE STUDY

Transforming Live Streaming TV, One Rack at a Time

How Zapping scaled premium low latency streaming across Latin America using NETINT VPUs.



Meet Nacho.

Ignacio "Nacho" Opazo
Co-founder & CTO, Zapping



Zapping at a glance:

COMPANY	Zapping
FOUNDED / HQ	Santiago, Chile
CATEGORY	Live OTT streaming television
MARKETS	Chile, Brazil, Ecuador, Peru
SUBSCRIBERS	600,000+
CHANNELS	500+ across all countries
NETINT HARDWARE	NETINT T408 VPUs (2022 to 2023), Quadra T1U VPUs (2025 to present)
KEY OUTCOMES	Server footprint reduced from about 100 CPU servers down to 25 VPU servers. About 90% less encoding power; substantial capex and opex savings.

The Company.

Zapping set out to do what incumbents said could not be done profitably: deliver premium, cable quality live television, including the region's most watched soccer, over the internet, with lower latency than local broadcast channels.

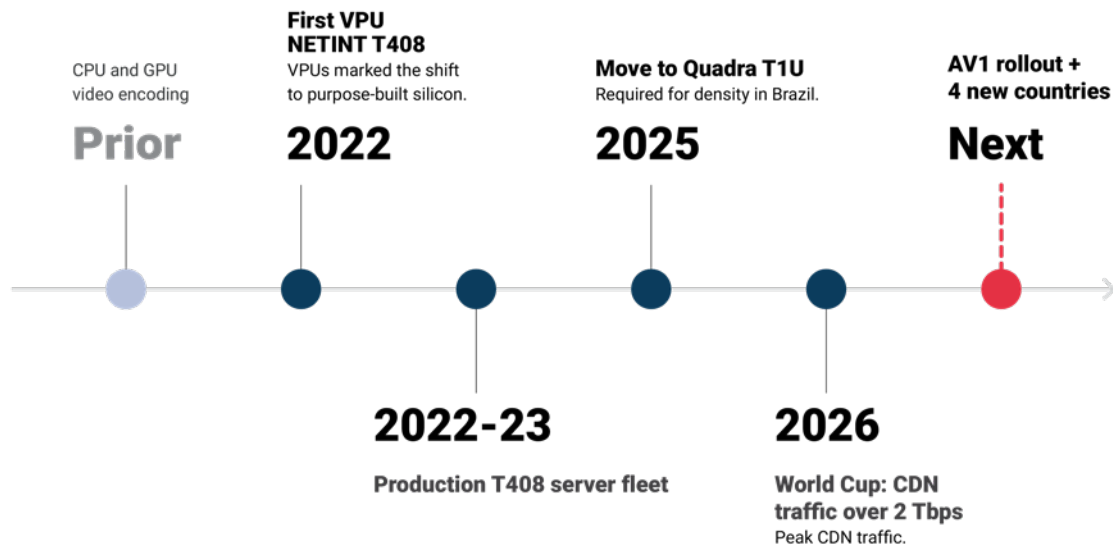
Founded in Chile and led by co-founder and CTO/CPO Ignacio "Nacho" Opazo, Zapping positions itself as much a technology company as a content company. It built its own encoders, low latency delivery architecture, custom apps, and proprietary CDN. As Nacho puts it:



"Zapping is the Netflix of live streaming here in Chile, in Latin America. We developed all our technology, the encoders, our low latency, and the apps on each platform. We developed our own CDN; I think it is bigger than Akamai and Fastly here in Chile and Ecuador. We are taking the same steps as Netflix: you make your platform, you make the UI, you make the encoding process, and then you deliver."

That vertically integrated approach turned Zapping into one of Latin America's fastest growing live television streaming platforms and, in the process, made encoding density, power, and cost a competitive advantage.

Zapping's Journey on NETINT:



CHALLENGE: Soccer, Latency and the Economics of Scale.

In Latin America, live sports, soccer above all, drive subscriptions. They also create demanding engineering problems, starting with latency. "Here in Chile, the soccer matches are premium. You can hear your neighbor screaming if they have a cable operator with lower latency," Nacho explains. "In Brazil it is more complicated, because some matches are free to air. So we must be fast enough to try and match free to air."

Soccer is punishing to compress, with wide shots of the field where players are small but jersey numbers must stay visible. With crowds, waving flags, and fast pans, preserving detail and color integrity without inflating bitrate is essential.



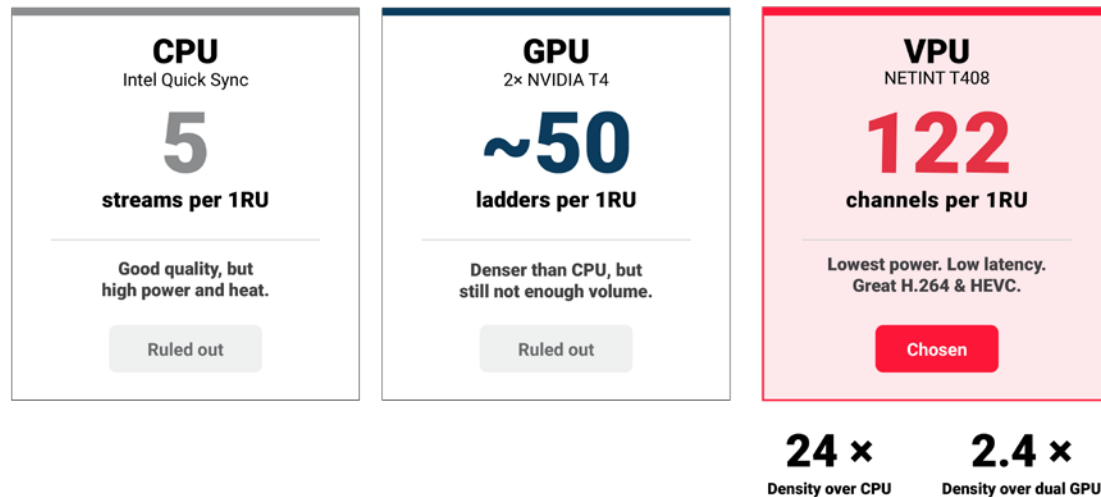
"We developed our own upscaling technology to improve sharpness and achieve superior quality compared to what broadcasters offer. We can upscale images to 4K and convert SDR to HDR."

"Video files are huge, and when you have a soccer match, thousands of people hit your servers, saturating the region's internet." In some markets, inter ISP trunks were as small as 10 Gbps, and Zapping would overload them. That prompted the company to build its own CDN and implement a server cache system within every possible ISP across Latin America.

Every one of those requirements had to be met affordably enough to install and operate data center capacity across South America, market by market. This made channel density per rack unit, power draw, and heat dissipation high priority problems to solve.

JOURNEY: From CPU to GPU to VPU.

Nacho tested every computing architecture, CPU and GPU, before settling on VPUs. What he ruled out is as instructive as what he chose.



How Zapping Chose:

CPU (Intel Quick Sync) delivered good quality, but “my problem was getting more density for the rack unit. Intel enabled just five streams per 1RU. Plus, the power and heat produced was really high.”

GPU (NVIDIA P2000 to T4) was an improvement, with roughly 50 complete encoding ladders per 1RU on a dual T4, 80 core configuration, “but still insufficient.”

VPU (NETINT Logan to Quadra series). Nacho discovered a NETINT article showing he could achieve 122 channels per rack unit using the first generation T408. “I found the power draw was really low, as was the latency, and the quality of both H.264 and HEVC is really good.” Integration, a common worry with dedicated silicon, was a non issue: “We control the encoders with FFmpeg, and converting over to the NETINT transcoders was really seamless. Really, really easy.”

Zapping moved into production on T408 based streaming servers in 2022 and 2023, then transitioned to the second generation Quadra T1U VPUs in 2025. This provided the density and codec headroom needed to keep pace with expansion into larger markets like Brazil. As Nacho framed it: “I need to get more dense. Brazil is a really big country. We need more power and more density in the rack.”

The density math is what changed the business. Using NETINT VPUs, Zapping reduced the server count needed to serve its ladders from roughly 100 CPU based servers to a fraction of that.



“We currently have 25 servers of varying density depending on their purpose, using T408 and Quadra VPUs. Each server is serving 50 to 60 channels with full ABR ladders consisting of 6 to 8 profiles using H.264 and HEVC.”

RESULTS:



“This new configuration allows us to reduce the number of rack units, energy costs, and capex, while giving us high availability and total control over the quality of our encoding, surpassing our competition in image quality and customer satisfaction.”

Density and footprint. Zapping went from about 100 software based servers to 25 VPU servers for the same encoding volume, roughly 75% fewer servers.

Power and sustainability. Previous infrastructure consumed about 600 watts per server, while the VPU based servers require just 250 watts and provide three to four times the channel density. Across the fleet, that works out to roughly 90% less power for the same encoding workload, which directly reduces both cost and Zapping’s footprint in every data center it deploys.

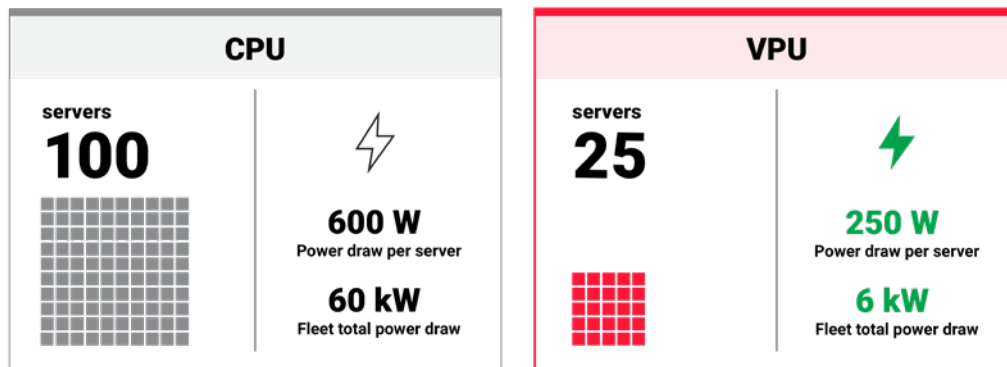


Figure 3.
The efficiency gain:
same workload on a
quarter of the servers

Cost. Reducing the server count from more than 100 to 25 for the same volume of encoding, together with the drop in power, drives substantial capex and opex savings. As Nacho puts it, NETINT has helped “accelerate our expansion while increasing our profitability.”

Latency. By placing a NETINT server inside each broadcaster’s CDN, Zapping originates its streams as close to the source signal as possible. In practice, Zapping lands within a few seconds of free to air, and in some cases ahead of it.

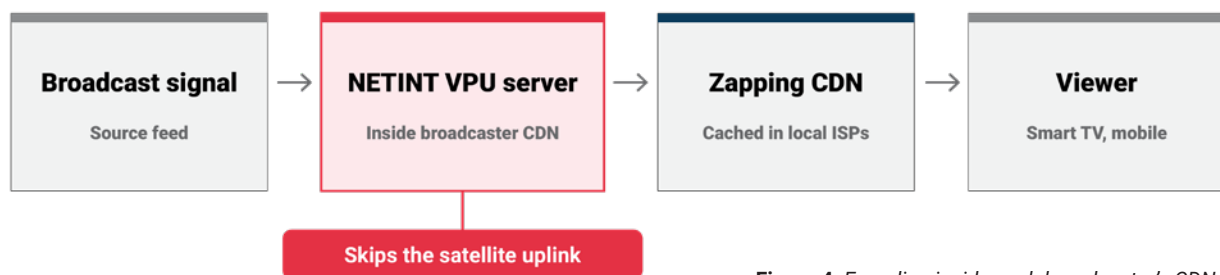


Figure 4. Encoding inside each broadcaster’s CDN keeps Zapping close to the source signal.

Bandwidth via HEVC and AV1. “HEVC is really important. Using these advanced codecs, we achieve a 40% lower bitrate than H.264 at the same quality; the user can see the difference.” Alongside Super Resolution (4K upscaling), Zapping is now beginning to work with AV1 and hopes to replace HEVC in the near future.



“During the 2026 World Cup, our CDN reached consolidated traffic over 2 Tbps.”

Why it matters for streaming video platforms.

Zapping is proof of a thesis at the heart of the VPU Ecosystem: that purpose built video silicon lets an independent operator compete with, and outperform, far larger incumbents. By collapsing the cost, power, and rack space of encoding, VPUs let Zapping put capital into content and its own CDN instead of ever larger data centers.



“NETINT servers are an affordable, functional, and high performance element of our success: unparalleled density, excellent low latency, and H.264 and HEVC quality with extremely low power consumption. NETINT has helped accelerate our expansion while increasing our profitability.” - Nacho Opazo, Co-founder and CTO, Zapping

Zapping continues to advance its software that clusters VPUs and distributes video encoding jobs across its data centers, while unifying these flows with its Stream Gateway. Its Upscaling and SDR to HDR (Super Resolution) technology demands ever more efficient encoding, so the team is now optimizing encoding using neural networks.



“With VPUs, we continue to open new countries in Latin America, and hope to open four new countries in the medium term.”

Get More From the Servers You Already Own

When more encoding capacity is needed, buying servers is the instinct. It should not be the only option. In a constrained market, the infrastructure already in the rack deserves a second look.

Many organizations are sitting on servers that are underused, aging, or no longer assigned to business-critical work. Some are waiting to be decommissioned. For general-purpose compute they may be past their prime, but for video processing that assumption can be wrong.

Encoding and transcoding do not have to depend on the CPU. When a server can host dedicated video acceleration, its usefulness then changes. A system that struggles with CPU-heavy workloads can still support high-density video processing once it's paired with a NETINT VPU.



The server still matters, but the workload changes.

A VPU does not make server design irrelevant. The host still has to provide the right physical, electrical, thermal, and software environment. What changes is the performance equation, because video processing moves off the CPU and onto a dedicated accelerator.

In a CPU-only model, older servers often cannot deliver the density or power efficiency modern video demands. With a VPU handling encode and transcode, the CPU, memory, storage, and network still matter, but they no longer carry the full video burden. That is what can make selected existing servers useful again, especially when a team has rack space and growing demand but does not want to wait for a full refresh.

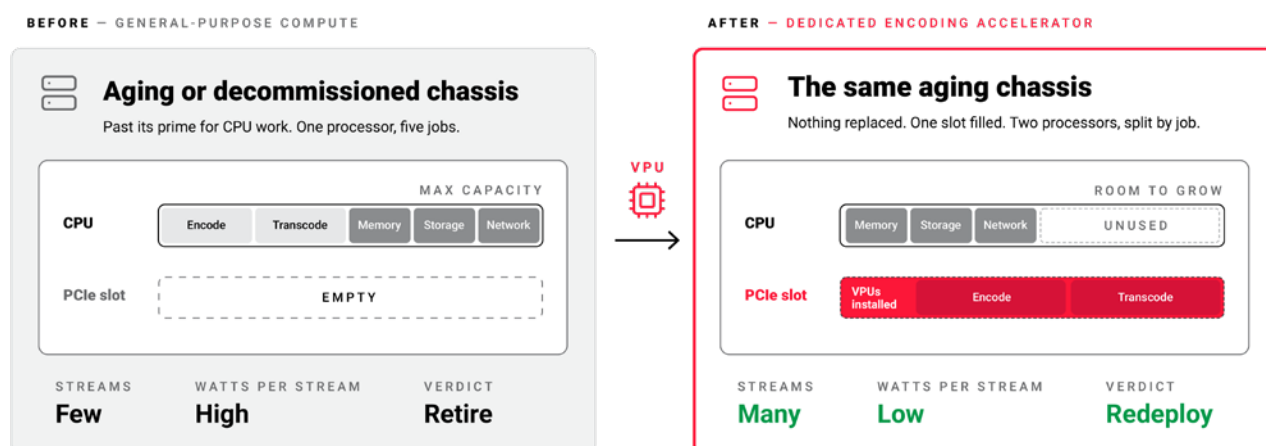


Figure 1. A second life. An idle or aging host plus a NETINT VPU becomes a dedicated, high-density encoding node.

Reuse requires validation.

This has to be handled carefully. An open PCIe slot does not make a server production-ready. Live video workloads run for long periods, generate heat, and demand predictable behavior, so the host has to be tested for the conditions it will actually face.

Before committing, validate the basics: PCIe slot and lane availability, airflow over the card, power headroom, thermal behavior under sustained load, BIOS settings, operating system support, driver compatibility, monitoring, and workflow integration. If the server passes, it joins the capacity plan. If it does not, the evaluation still pays off by pointing toward a different server, a dedicated system, or a hosted model.

The goal is not to force reuse. It is to avoid writing off infrastructure before understanding what it can do. A VPU-based approach lets a team create a dedicated encoding pool, offload work from CPU-based systems, or test a new processing architecture using equipment already in the environment. The same evaluation that qualifies a server also reveals its real limits, which is information worth having either way.

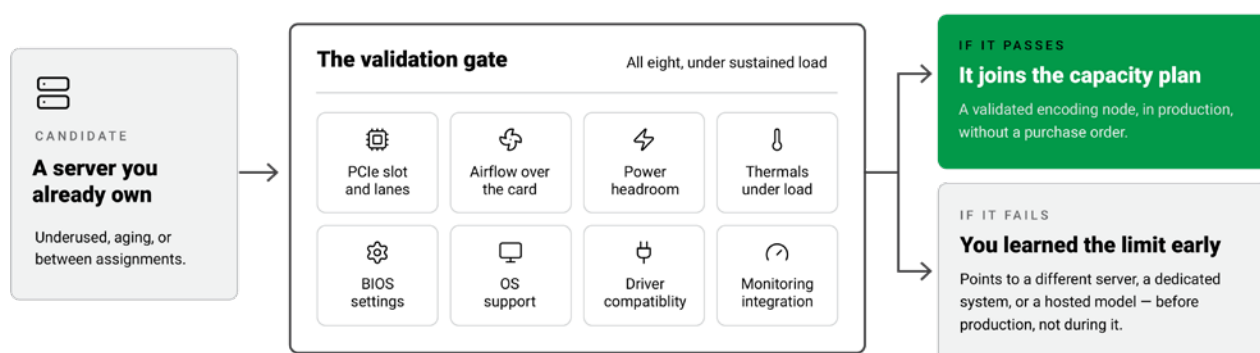


Figure 2. Reuse only after the host clears the gate. A pass yields a validated capacity unit; a fail is a lesson learned before production, not during it.

Why this matters?

New server deployments take time. Budget approval, procurement, installation, validation, and rollout all slow capacity expansion, and that delay hurts when the workload is already growing. Existing infrastructure can sometimes provide a faster starting point.

It will not solve every capacity problem, but it can prove a VPU workflow, support a specific workload, relieve CPU pressure, or bridge a gap while larger plans move forward. It also keeps the discussion grounded in real infrastructure rather than abstract capacity planning. Instead of starting with a large new deployment, a team can start with a focused assessment of which servers are available and technically suitable.

There is also a cost dimension that is easy to overlook. CPU-only encoding consumes power and rack space disproportionate to the streams it produces, so shifting that work to a VPU on an existing host can lower the cost and power impact of the same output. The point is not to glorify old hardware. It is to make a clear-eyed decision about what each system can still contribute before signing a purchase order.

Where a deployment partner helps.

Many teams will not want to evaluate this alone. Reuse sounds simple until the details begin: PCIe behavior, firmware, drivers, orchestration, and production monitoring. This is where a systems integration partner such as Arcadian can help assess whether existing servers are suitable, identify technical blockers, and define a path from evaluation to production. For teams with available infrastructure but limited time to validate every detail internally, that role can be the difference between a clean rollout and a stalled one.

There is an old idea behind this approach. You did not throw something away because it was no longer perfect. You looked at it again, repaired what could be repaired, and saved the new purchase for when it was truly needed. Video infrastructure can follow the same logic.

An existing server may no longer suit its original purpose, but that does not mean it has no value. With the right validation and the right video acceleration, selected systems can earn a second life as encoding or transcoding capacity. Before buying more infrastructure, it is worth asking a simple question: what do we already own that can still work harder for us?

**Use what you already have
before buying something new.**

**That is not recycling as a slogan.
It is practical engineering.**

Sources & further reading

1. NETINT, VPU Ecosystem and deployment paths. <https://netint.com/ecosystem/>
2. Arcadian, NETINT VPU integration to cut encoding OPEX (Aug 7, 2025). <https://www.arcadian.la/post/arcadian-announces-netint-vpu-integration-to-cut-video-encoding-opex-50>
3. Akamai, Benchmarking VPUs and GPUs for Media Workloads (Aug 19, 2025). <https://www.akamai.com/blog/developers/benchmarking-vpus-and-gpus-for-media-workloads>

Build Dedicated VPU-Based Video Infrastructure

Some infrastructure decisions are temporary. Others become the foundation. When video processing is central to the business, the strongest move is to build the system around video from the start.

Hosted capacity helps when deployment speed is the immediate issue. Existing servers can sometimes be reused when the environment is suitable. Dedicated VPU infrastructure is for teams that want control: to own the hardware, standardize the deployment, and build capacity they expect to operate for years. That means choosing infrastructure based on the actual workload. Codecs, resolutions, latency targets, density requirements, power limits, and the operating model all point to the right design. The direct NETINT path is to choose the right VPU, place it in the right system, control it with the right software, and scale from a known architecture.



Start with the workload.

A dedicated system should be designed around the video it will run every day. A live platform cares most about real-time performance, predictable latency, and stream stability. A user-generated or short-form platform cares about high-volume transcoding, cost per output, and density. A broadcaster or event team may need compact systems for onsite, edge, or mobile use.

Those workloads call for different choices. Codec support, output ladder, resolution mix, input format, latency target, throughput, location, and operating model all shape the right product path. The decision should start with the video requirement, not the part number.

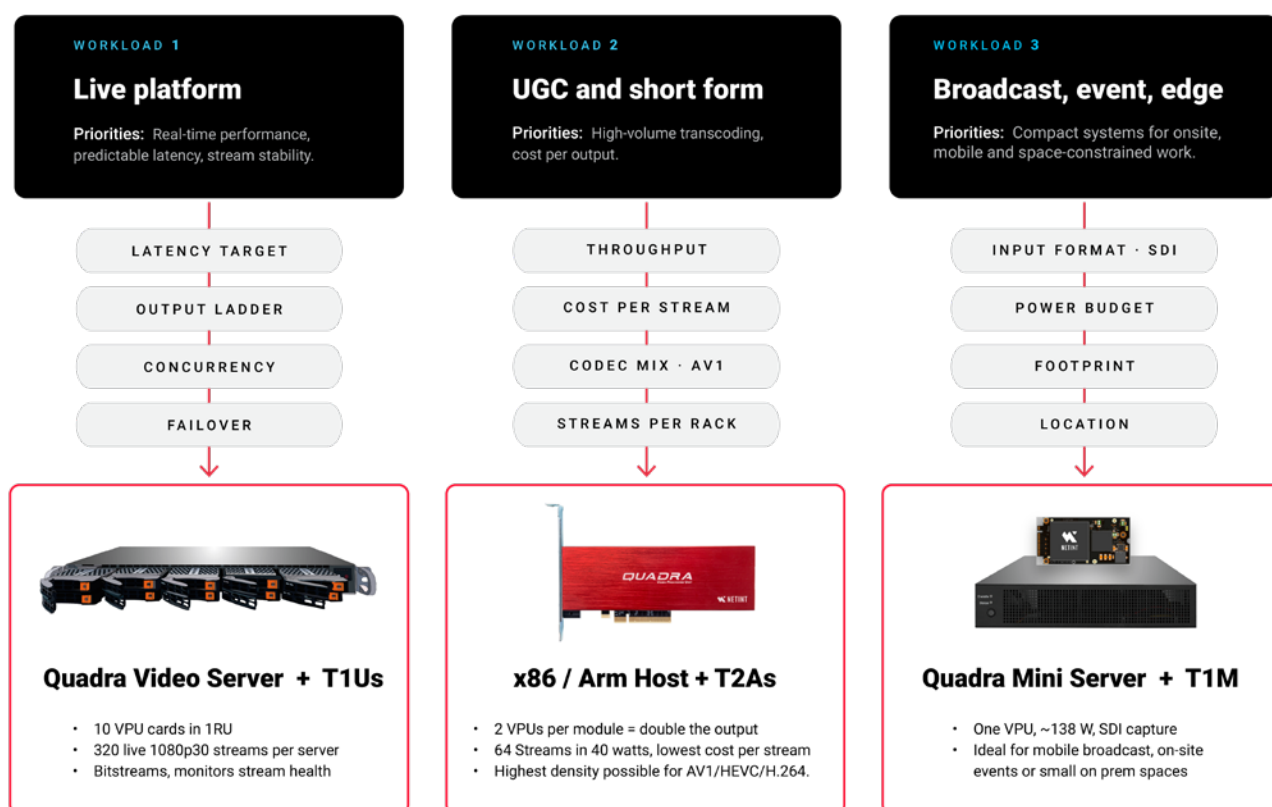


Figure 1. The video requirement points to the product path. Start with the workload, then choose the silicon, system, and software that match it.

Choose the silicon: Quadra VPUs.

The Quadra family is the core of the dedicated story, with four form factors built on NETINT's Codensity G5 ASIC and a shared software stack across FFmpeg, GStreamer, and the SDK. The T1M is the compact M.2 part for edge, embedded, and space-constrained designs at roughly 8 to 10 watts. The T1U is the U.2 scale-out option at about 17 watts, handling on the order of 32 live 1080p30 streams per card.

The T1A is the PCIe add-in card for server-based deployments where power, thermals, and tuning matter most. The T2A is the density path, a dual-ASIC card that maximizes output per server and lowers cost per stream in high-volume environments. The decision should start with the system design: form factor, how many VPUs must fit, the realistic power budget, available airflow, and the codecs and resolutions required.

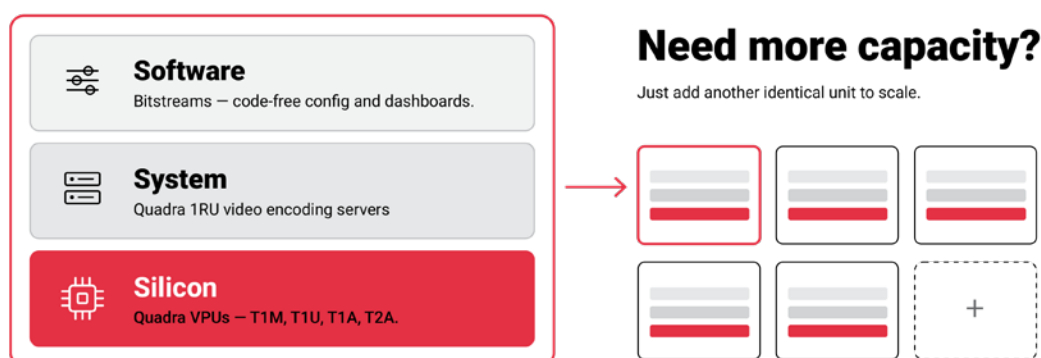


Figure 2. One architecture, three layers. Silicon, system, and a software control layer combine into a known, repeatable capacity unit.

Choose the system, then the control layer.

A VPU provides the processing engine, not the whole system. The host still has to support the deployment physically, electrically, thermally, and operationally, which is why dedicated infrastructure is an architecture decision rather than a card installation. Anyone can put a card in a server and run a test. Production is different: it means sustained workloads, predictable stream behavior, monitoring, failover planning, and a repeatable path for adding capacity. The goal is a known-good configuration that becomes the building block for future capacity.

For teams that do not want to engineer a server from components, the Quadra Video Server provides a validated, high-density multi-VPU platform, with around ten T1U cards in a 1RU chassis supporting up to roughly 320 live 1080p30 streams. For compact, on site, edge, or mobile work, the Quadra Mini Server packages a single T1M into a small system at about 138 watts with SDI capture.

Hardware alone does not make a deployment manageable. Bitstreams is NETINT's control and automation layer, turning VPU hardware into a deployable platform with code-free FFmpeg configuration, reusable templates, and dashboards for stream health and system load. It matters because operational problems usually have to be solved alongside processing problems.

Engineers can always work directly with FFmpeg, GStreamer, SDKs, and APIs. But many production deployments still need a practical way to configure, observe, and repeat encoding workflows without rebuilding them by hand. Bitstreams is most useful when a team wants a managed VPU workflow rather than a purely low-level integration, so it can see what is running, whether streams are healthy, and how capacity is being used.

If video is core to the business, build for video. Define the requirements first, then choose the silicon, system, and software that meet those requirements.

The ecosystem around the deployment.

A dedicated deployment is rarely just one product. A team may have the server but need the software layer, or have the workflow software but need validated hardware, or need cloud capacity, integration, transport, or monitoring around the VPU. This is where the broader ecosystem completes the system: hardware partners such as Advantech, integration partners such as Arcadian, and hosted capacity when a region or a deadline demands it.

Dedicated VPU infrastructure is the clearest and longest-lived path for teams that want to own and control their video processing. When the workload is clear, permanent, and central to the business, designing the system around video from the beginning gives the most control over density, power, performance, and operations.

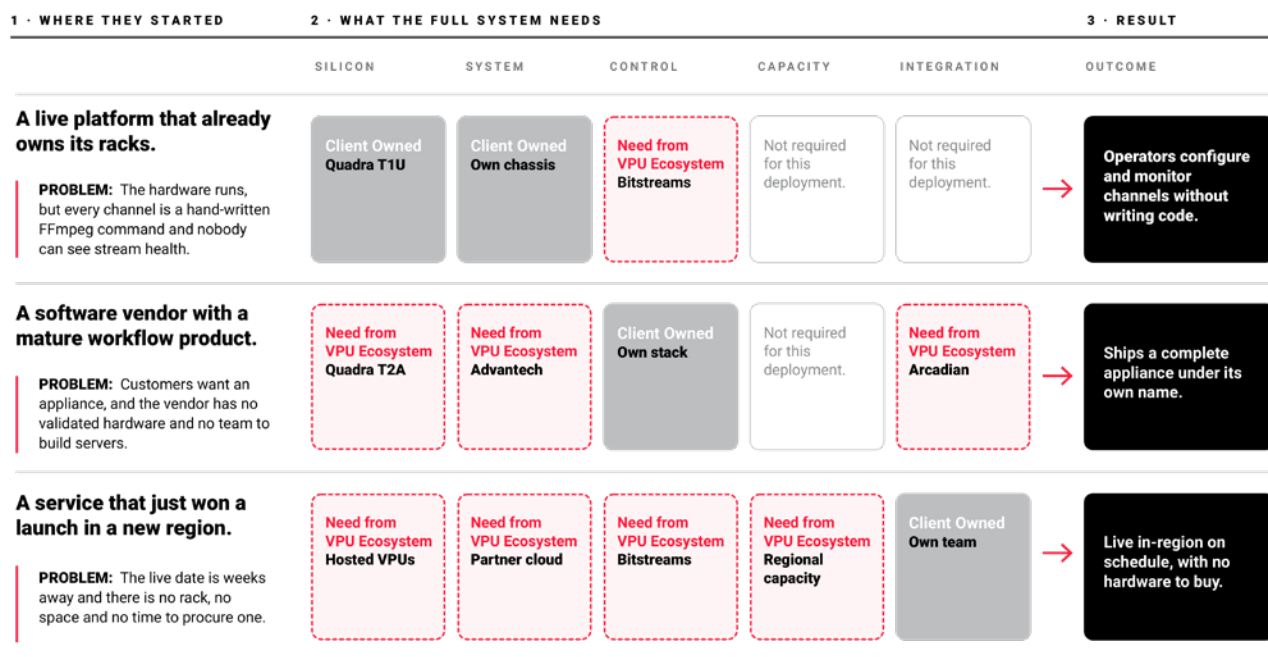


Figure 3. A deployment is rarely one product. The VPU Ecosystem fills the gaps.

Sources & further reading

1. NETINT, VPU products, Quadra family specifications. <https://netint.com/products/>
2. NETINT, Bitstreams control and automation layer. <https://netint.com/bitstreams/>
3. NETINT, Quadra Mini Server (Apr 2025). <https://netint.com/quadra-mini-server-empowering-the-next-generation-of-video-streaming-workflows/>

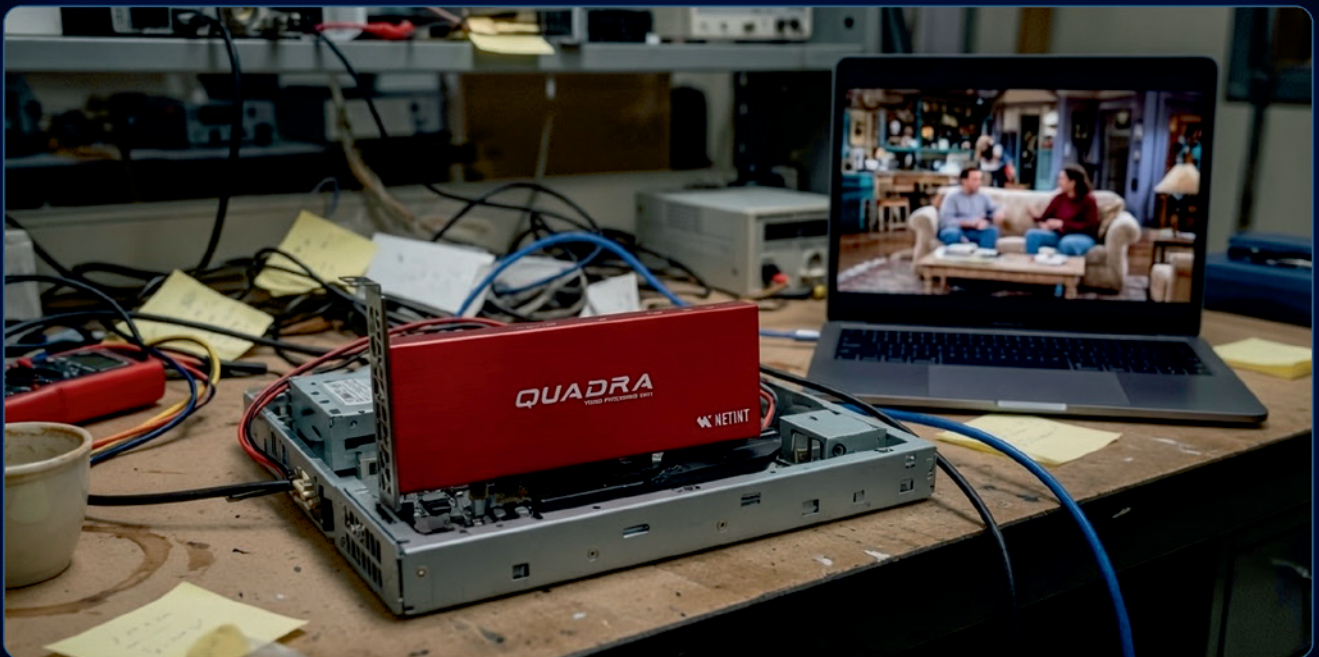
Sure, VPUs are great.

They're just not going to integrate themselves.

The silicon is the easy decision. The integration into your supply chain is the hard part, and that's the only thing we do.

A VPU on an eval board is a promise. Turning it into a production pipeline is a different discipline entirely: profiles, orchestration, failover, monitoring, a live cutover that never drops a stream. Not a hardware purchase. An integration project.

Arcadian takes media companies from “the chip looks great” to “it's running in production,” and stays through scale.



See what's between you and production.

Talk to the VPU Integrators — arcadian.la



Field-Proven Methodology for Deploying NETINT Quadra VPUs

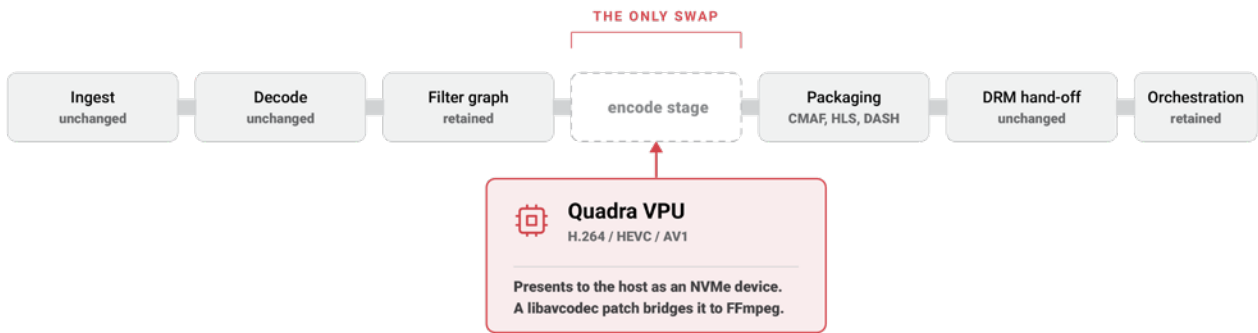
How to move from evaluating to migrating your CPU- and GPU-based transcoding to VPU infrastructure.

Why VPUs, Why Now?

Every VPU project we deploy starts with a trigger. The specifics vary, but the underlying pattern is the same: a real-time video workload has outgrown the economics of general-purpose compute. The triggers we see most often:

- **Rack space** — Rack space and density ceilings. The facility is full, and adding CPU encode capacity means adding racks, power circuits, and cooling that don't exist.
- **Power** — Power and cooling budgets under pressure. Sustained software encoding burns tens of watts per stream; at channel-count scale, the power bill becomes an architecture decision.
- **Codec efficiency** — An AV1 or HEVC mandate that CPU encoding can't economically meet. Modern codecs demand dramatically more compute per frame; real-time AV1 in software is possible but rarely affordable at density.
- **Cloud cost** — Cloud transcode spend that scales linearly with growth. Per-minute cloud encoding pricing that made sense at launch stops making sense at scale, and egress compounds it.
- **Control** — Latency, data-sovereignty, or on-premises requirements that rule out managed cloud encoding entirely.

Figure 1. One segment of your FFMpeg / GStreamer video pipeline is replaced, rest stays the same.



The Quadra family is purpose-built for exactly these conditions: one Codensity G5 silicon architecture across four form factors, selected by deployment environment and power envelope, from the M.2-based T1M for edge and space-constrained deployments to the dual-ASIC T2A, which delivers up to 64 streams of AV1, HEVC, and H.264 in roughly 40 watts. For teams that want turnkey density rather than card-level integration, the 1RU Quadra Video Server runs ten VPUs and ships with NETINT's Bitstreams management software preinstalled.

Recognizing the trigger is step one of the methodology. The remaining nine steps are how we turn that recognition into a production system without betting a channel lineup on a datasheet.

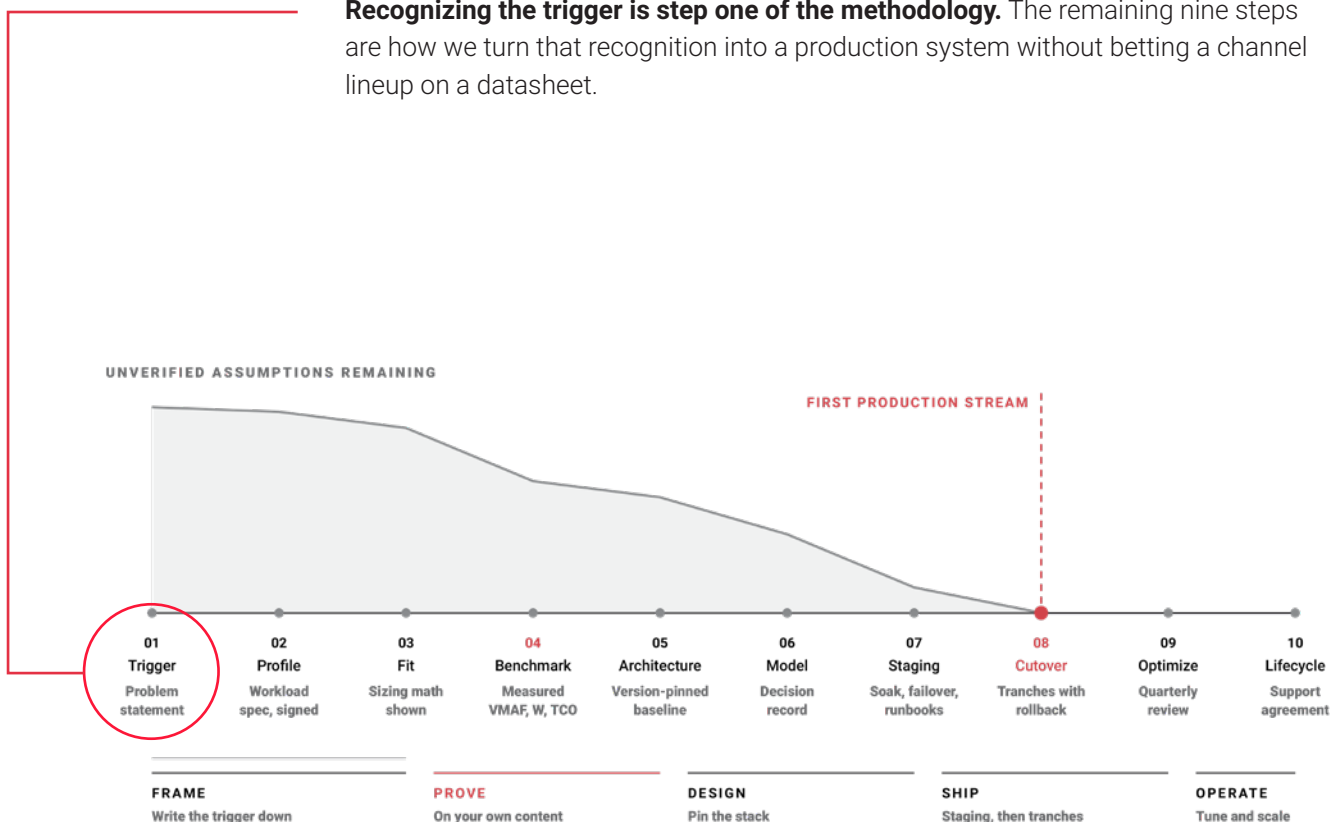


Figure 2. 10 Steps to de-risk deployment before the first stream, not after it. Every step concludes with a deliverable. The shaded area is assumed risk. After cutover, risk is controlled and therefore eliminated.

The Arcadian Deployment Methodology.

Step 1 — Problem Recognition

This step usually happens before we're in the room, but making it explicit matters: the trigger defines the success criteria for everything downstream. A power-constrained facility measures success in watts per stream. A team facing an AV1 mandate measures it in quality-per-bit at real-time speed. A cloud repatriation project measures it in three-year TCO. We begin every engagement by writing the trigger down and agreeing on the metric that will ultimately declare the project successful.

Deliverable 1:

One-page problem statement with the primary success metric and its target value.

Step 2 — Evaluation Intent & Workload Profiling

This is the step most vendors skip and the step that determines whether the hardware sizing in Step 3 is engineering or guesswork. We profile the actual workload: codec mix (H.264, HEVC, AV1, and any transition plans between them), ABR ladder profiles and renditions per channel, live-to-VOD ratio, peak and steady-state concurrency, latency budget, and, critically, the existing software stack. Whether the pipeline is FFmpeg, GStreamer, or a custom application against the SDK changes the integration plan materially. We also capture the operational envelope: available rack units, power per rack, host server inventory and PCIe generation, and the monitoring stack the new hardware must report into. The output is a workload specification that both Arcadian and the customer engineering team sign off on. Every subsequent step traces back to this document.

Deliverable 2:

Signed-off workload specification: codec mix, ladder profiles, concurrency, latency budget, software stack, and operational envelope.

Step 3 — Solution & Hardware Fit

With the workload spec in hand, hardware selection becomes arithmetic rather than persuasion. The Quadra family shares one silicon architecture across four form factors, so the question is which package fits the deployment environment: the T1M (M.2) for edge and space-constrained systems, the T1U (U.2) and T1A (PCIe add-in card) for standard data-center scale-out, and the T2A (dual-ASIC, single slot) when slot count is the binding constraint and cost per stream is the priority. Where the customer prefers a turnkey platform, the Quadra Video Server provides ten VPUs in 1RU with Bitstreams preloaded.

The sizing document shows its math: streams per card for the customer's specific ladder (not a generic 1080p30 figure), cards per host given PCIe lanes and thermals, hosts per rack given the power budget, and the resulting headroom against projected growth. We size for the workload spec plus an explicit growth margin — never to the datasheet maximum.

Deliverable 3:

Sizing document with streams-per-card, cards-per-host, and hosts-per-rack calculations shown, including growth headroom.

Step 4 — Proof of Concept & Benchmarking

The POC exists to replace every number in the business case with a measured one. We benchmark against the customer's own content such as their sports feed, their film grain, their screen-share webinars, never canned demo clips, because encoder behavior is content-dependent and the only benchmark that predicts production is one run on production-representative sources.

Quality is scored with VMAF (and PSNR where the customer's QC process requires it) against the output of the incumbent encoder at matched bitrates, reported as both mean and low-percentile scores. Density is measured as sustained concurrent ladders per host under thermal steady state, not burst. Power is measured at the wall, per stream. These measured figures feed a TCO model comparing the VPU build against the incumbent CPU, GPU, or cloud spend over a three-year horizon, and it is this validated TCO, not the raw performance data, that typically carries the project through budget approval.

For teams that want to validate before any hardware ships, NETINT VPU capacity is now available as-a-service through partners such as NetActuate, allowing benchmarks to run on rented Quadra-equipped infrastructure. Arcadian can execute the same POC protocol on rented capacity and carry the results directly into an on-premises sizing.

Deliverable 4:

Benchmark report with measured VMAF, density, and watts per stream, plus a three-year TCO model built from measured values.

Step 5 — Architecture & Integration Design

Quadra integrates at the encode stage of the pipeline: the card presents as an NVMe device, and a libavcodec patch bridges the device interface to FFmpeg and GStreamer. In practice this means filter graphs, packaging (CMAF, HLS, DASH), DRM hand-offs, and orchestration logic are retained; the encoder invocation changes. The architecture design covers where the cards sit in the media flow, load management across multiple cards and hosts, failure modes and stream-level failover behavior, and how card health, temperature, and utilization surface into the customer's existing monitoring (Prometheus, Zabbix, or Bitstreams' dashboard where the Quadra Video Server is in play).

This step also fixes the firmware and driver baseline and the supported FFmpeg/GStreamer versions for the deployment, so that the staging environment in Step 7 and production in Step 8 are built from an identical, version-pinned stack.

Deliverable 5:

Integration design document: media-flow placement, load management, failover behavior, monitoring integration, and a version-pinned software baseline.

Step 6 — Deployment Model Selection

With the architecture defined, the deployment model is a business decision made with engineering consequences on the table: fully on-premises, colocation, hybrid (VPU capacity for steady-state load with cloud burst retained for peaks), or a phased model that runs VPU infrastructure alongside existing cloud encoding while confidence builds. There is no universally correct answer — the workload spec from Step 2 and the TCO model from Step 4 make the trade-offs explicit, and the choice determines the rollout sequencing in Step 8.

Deliverable 6:

Deployment model decision record with the selected model, the alternatives considered, and the rollout sequencing.

Step 7 – Pilot & Staging Validation

Staging is where the design meets production-shaped reality. We run the pinned software stack on production-identical hardware under representative load: full ladders, real channel counts, soak tests measured in days rather than minutes, and deliberate failure injection (card removal, host loss, process kill, etc.) to confirm that failover behaves as designed and that monitoring alerts fire where and when they should. Operational runbooks are written and exercised here, by the people who will own the system, before any production stream depends on them.

Deliverable 7:

Staging validation report: soak results, failover test outcomes, alert verification, and signed-off operational runbooks.

Step 8 – Production Rollout & Migration

We never big-bang a channel lineup. Rollout is phased: a first tranche of lower-risk channels cuts over with the incumbent encode path kept warm as rollback, followed by progressively larger tranches as confidence accumulates. Each tranche has explicit entry criteria (staging sign-off, monitoring green), exit criteria (a defined observation window at production quality), and a rehearsed rollback procedure. Because staging was version-pinned and production-identical, cutover is an operational exercise rather than a discovery process — the first production stream behaves exactly like the last test stream, and if it doesn't, the process failed before the rollout did, which is exactly where failures belong.

Deliverable 8:

Migration plan with per-tranche entry/exit criteria and rollback procedures, executed and signed off tranche by tranche.

Step 9 – Optimization & Scale-Out

Once production is stable, tuning begins: capped-CRF and rate-control profiles refined against real traffic for predictable CDN cost, ladder pruning where AV1 efficiency allows renditions to be consolidated, and utilization review to reclaim headroom. Scale-out follows the same sizing math as Step 3, now calibrated by months of production telemetry rather than POC data. This is also where codec-transition plans (H.264 to HEVC to AV1) execute incrementally, since the same silicon handles all three.

Deliverable 9:

Quarterly optimization review: rate-control tuning changes, ladder adjustments, utilization and capacity forecast.

Step 10 – Ongoing Support and Lifecycle Management

VPU infrastructure is long-lived, and its operational lifecycle is where a partner earns its keep. Arcadian provides firmware and driver lifecycle management against NETINT's release stream (including validation of upgrades in staging before production), monitoring and alert response, capacity planning as channel counts grow, and a defined escalation path into NETINT engineering for issues that require it. The goal is that the customer's engineering team spends its time on their product, not on encoder plumbing.

Deliverable 10:

Support agreement covering firmware/driver lifecycle, monitoring, capacity planning, and vendor escalation paths.

The Integration Surface, Honestly Stated.

Engineering leads rightly distrust the phrase “drop-in replacement,” so here is the accurate version. If your pipeline is FFmpeg- or GStreamer-based, the Quadra SDK integrates through a libavcodec patch, and your encode invocations change while your surrounding pipeline logic does not. Supported FFmpeg and GStreamer versions are documented and version-pinned per deployment; running outside the supported matrix is where avoidable pain lives, which is why Step 5 fixes the baseline and Steps 7–8 never deviate from it.

If your pipeline is a custom application, integration is against the SDK directly and the effort is larger. This is exactly the case the workload profiling in Step 2 exists to surface early, so it is scoped in the plan rather than discovered in staging. In both cases, load management across multiple cards, stream-level failover, and health monitoring are solved problems with established patterns; the architecture design in Step 5 selects among them rather than inventing them.

Working With Arcadian.

Arcadian is a streaming media operations and technology company and a formal NETINT Technologies partner, with engineering and operations teams across the US, Europe, and Asia-Pacific. We have deployed and operated video infrastructure across live and VOD workloads for broadcasters, platform operators, and service providers, and the methodology in this paper is the one we actually run, deliverable by deliverable.

If you are evaluating VPU-based transcoding for triggers based on power, density, AV1, or cloud spend, the fastest way to a decision you can defend is a profiled workload and a benchmark on your own content. That is Steps 1 through 4, and it is where every engagement begins.

arcadian.la · Los Angeles, CA



The State of Video Encoding

Condensed insights from 286 video professionals on codec adoption, hardware acceleration, AI integration, and strategic priorities shaping the future of streaming infrastructure.

286

Industry professionals surveyed across streaming, broadcast, and enterprise video.

46%

VP-level or above, validating strategic level findings.

35%

Engineers and architects grounding operational insights.

2025

Data collected Q4 2025

3

Representing 3 industries: streaming, broadcast and enterprise in global regions: EMEA, NA, APAC, LATAM.

Source:

2026 State of Video Encoding Industry Report, NETINT Technologies, Q4 2025.

Link to unabridged report:



6 Findings define 2026 video encoding market

Cost and quality are no longer trade offs. This year's data reveals an industry navigating simultaneous pressures. The era of choosing between cost and quality is ending. Solutions that deliver both will define the next generation of encoding infrastructure. Those requiring a tradeoff face increasing resistance from buyers operating under tighter budgets with leaner teams.

1 AV1 will triple its production footprint in 2026.

Currently at 17% production deployment, AV1 will reach 57% combined reach by year end as 40% of respondents plan deployment. HEVC maintains dominance at 85 combined, but AV1 is the growth story. Device decode support (54%) and tool chain limitations (43%) remain the gating factors.

Organizations without a 2026 AV1 roadmap risk falling behind on bandwidth efficiency.

2 GPU dominance = concentration risk as alternatives gain ground.

64% use GPU acceleration, but market is actively reconsidering its hardware mix: 51% plan to evaluate GPUs and 49% will evaluate VPUs in 2026—near parity. VPUs and ASICs have already reached 28% adoption, with FPGAs at 14%.

VPU vendors have a 12–18 month window to capture organizations diversifying beyond GPUs.

3 AI is no longer experimental. Expansion is the default.

65% of respondents plan to expand AI/ML in encoding workflows next year. Current deployments concentrate in transcription (28%), scene classification (22%), and super resolution (15%). The next wave targets core encoding intelligence: content-aware ladder generation and real-time quality prediction. AI capabilities are becoming table stakes in encoder evaluation.

Vendors without AI road maps risk shortlist exclusion.

4 Budget and team constraints outweigh technology gaps.

The top roadmap blockers are organizational, not technical: budget constraints (39%) and limited team capacity (38%). Device support lags at 30%. Seventy-seven percent of respondents cite at least one organizational barrier, compared to just 19% citing a purely technical one.

Ease of integration and "time to value" will differentiate solutions more than raw performance.

5 Edge interest is strong, but conviction is uncertain.

42% expressed interest in deploying encoding at the edge, but 30% remain "not sure"—a substantial undecided segment. Live events operators show the strongest edge interest (44%), driven by latency requirements for real-time delivery.

Edge vendors must reduce ambiguity with reference architectures, use cases, and trial programs.

6 Executives & engineers align on cost; diverge on innovation.

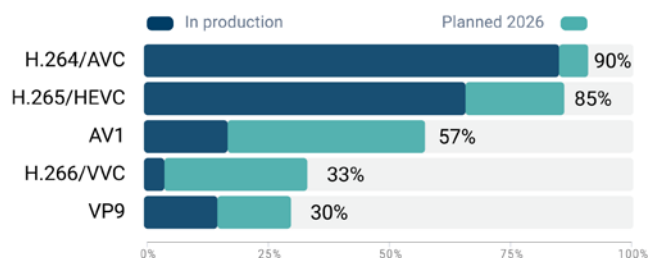
Executive priorities: revenue growth (47%), budget reduction (43%), cost per delivered hour (36%). Technical teams champion AI/ML (35%) alongside efficiency. The gap between innovation investment and executive KPIs demands that business cases connect technology to revenue enablement.

Position solutions against revenue growth and unit economics—the dominant executive KPIs.

Technology snapshot and strategic priorities:

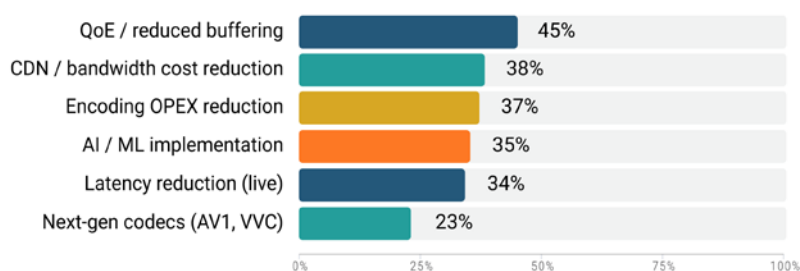
Codec Adoption Trajectory

Production deployment vs. 2026 combined reach.



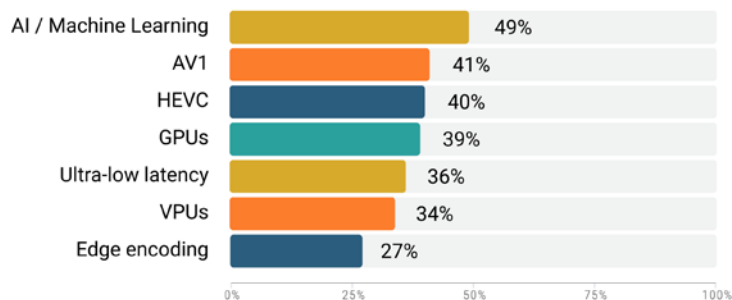
Strategic Priorities

Top technology initiatives.



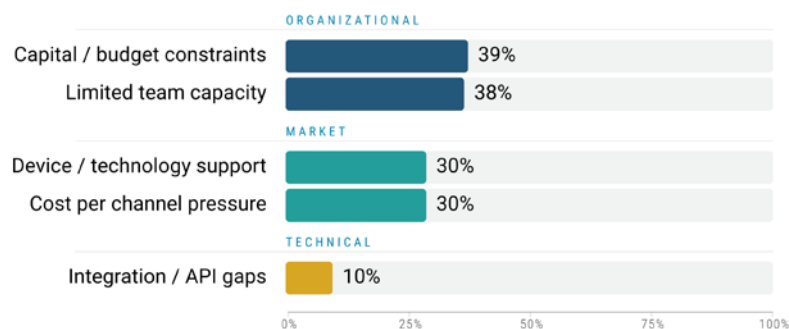
Expansion Plans

Top technology expansion area (% planning).



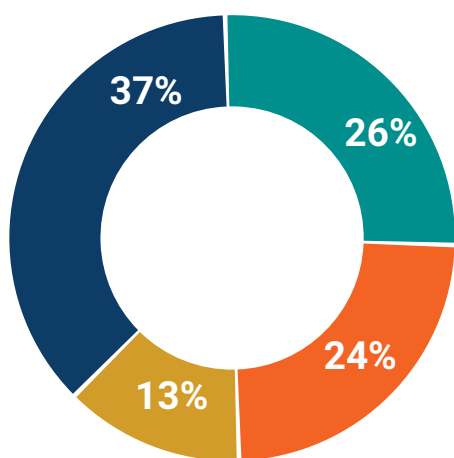
Roadmap Blockers

Top constraints limiting 2026 roadmap execution.
77% cite at least one organizational barrier.



Organizational archetypes shape market.

Analysis of operational scale, team composition, and technology posture reveals four distinct segments. Each archetype has different buying motivations, technology maturity and constraints demanding tailored approaches.



37% Small Team Operators

69% have 1–5 person video teams. Budget constrained but pragmatic; seeking turnkey solutions with minimal integration overhead.

26% Specialized / Evaluators

Bimodal scale: 40% zero VOD, 43% zero live. Pre-production or adjacent-industry evaluators representing future demand.

24% Infrastructure Engineers

Mid-scale VOD focus, 100–1,000 hours/month. Most innovation-forward: 74% plan AI/ML expansion. Cloud-centric deployment.

13% Enterprise Hyperscalers

68% run 5,000+ live channels. Most advanced codec stack but highest hardware skepticism (42%). Risk-averse, incremental optimizers.

10 Insights that Matter.

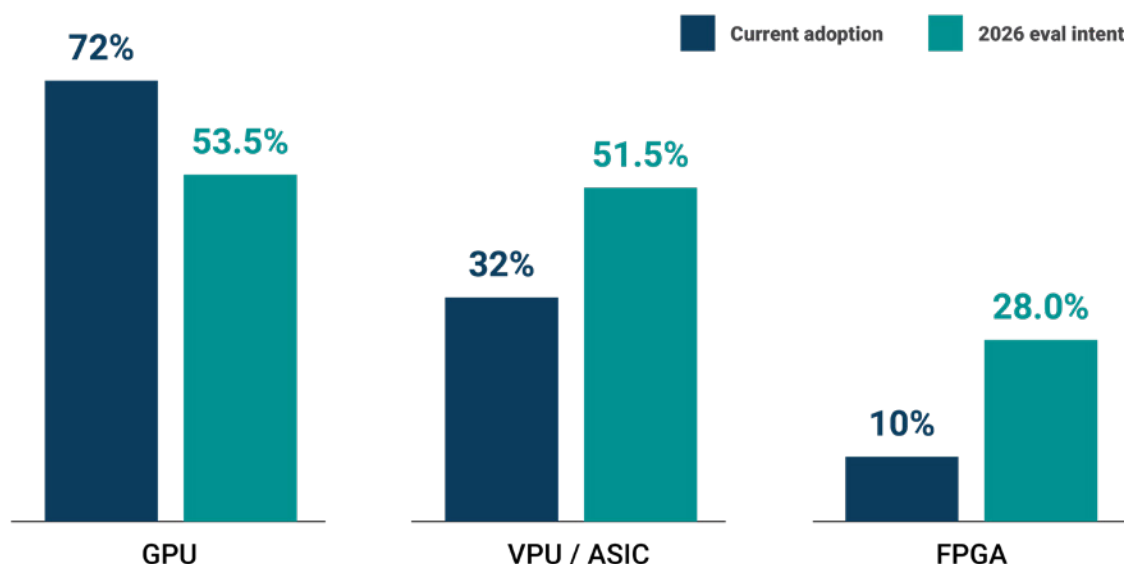
A strategic synopsis tailored for the VPU Ecosystem based on the complete 47-page: 2026 State of Video Encoding Industry Report Q4 2025. [Download full report here.](#)

1

The GPU Monoculture is Cracking.

Hardware inflection. GPU adoption sits at 72% but that number masks a structural stress fracture. GPU users report power consumption (39%), codec feature gaps (37%), and insufficient stream density (35%) as active pain points.

Meanwhile, VPU evaluation intent has reached 51.5%, nearly matching GPU evaluations at 53.6%. A 2.1-point gap. For the first time in a decade, no single hardware technology commands the market. The incumbency advantage is gone. This is the window.



2.1pp

GPU-VPU eval gap

41%

Deploy multiple HW types

39%

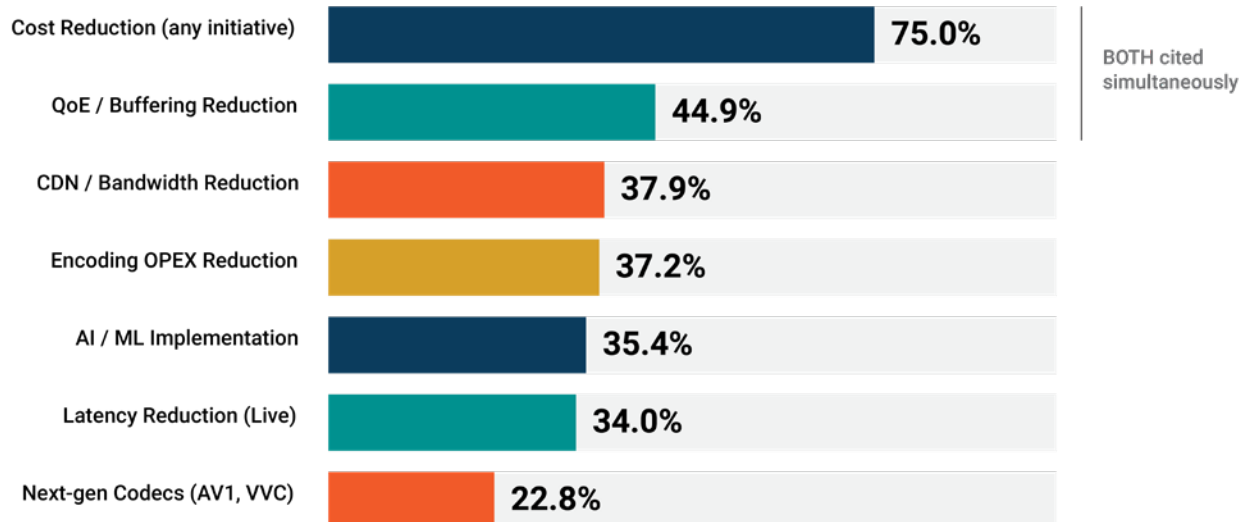
GPU users cite power pain

2

The Dual Mandate Replaced the Old Tradeoff.

Market demand signal. The old market logic used to be sacrifice quality for cost, or spend for quality. That is structurally over. 75% of respondents cite at least one cost-reduction initiative. 45% simultaneously cite quality-of-experience improvement as a top priority. Organizations are pursuing both, in parallel, right now.

Any vendor still framing this as a tradeoff is speaking a language the market has already abandoned. The either/or era is structurally over.

**2.31**

Avg initiatives / org

42%

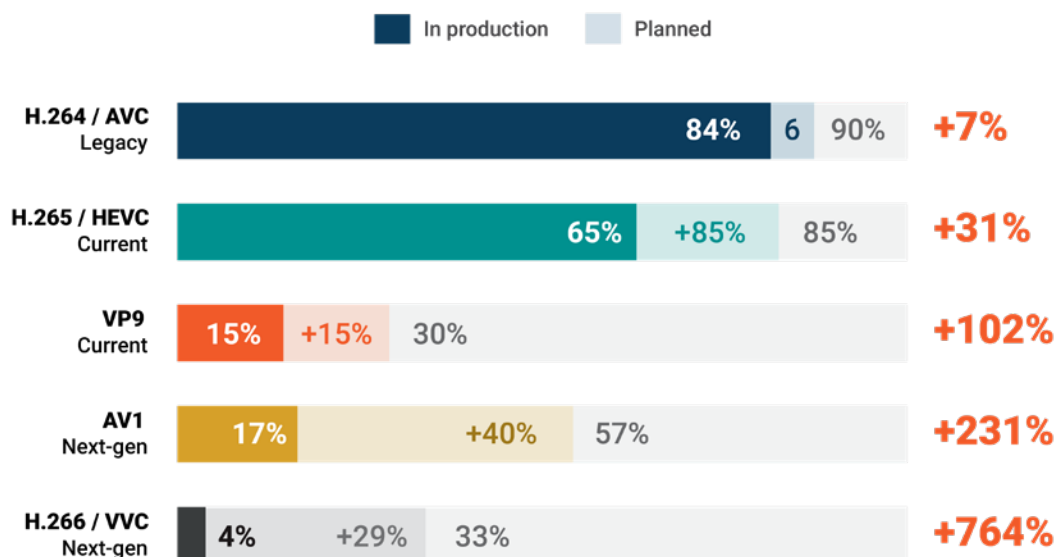
Select all 3 priorities

3

AV1 Is Not a Future Codec. It's a 2026 Infrastructure Decision.

Codec transition. AV1 sits at 17% production deployment today, but 40% of organizations plan deployment this year, representing a 231% planned growth rate, the highest of any codec. The progression ladder is predictable: organizations with VP9 in production are 27x more likely to have AV1 deployed. HEVC adopters are the natural next wave.

Hardware decode support (54% cite it as a barrier) is the gating factor which means purpose-built silicon is directly in the critical path for a solution.



57%

AV1 combined reach EOY 2026

+231%

AV1 planned growth rate

54%

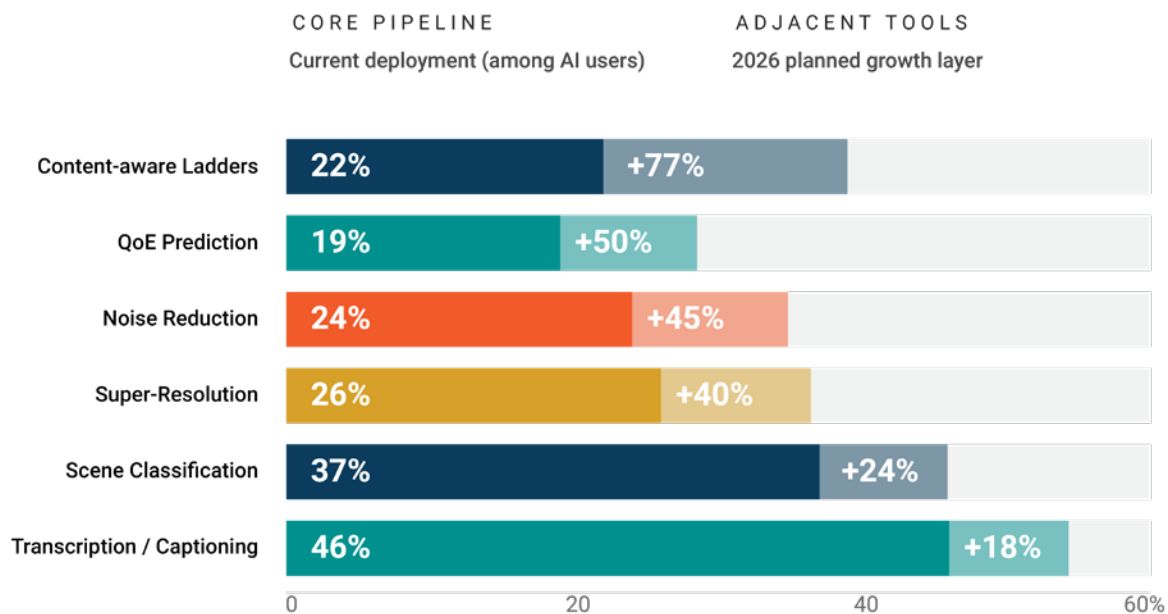
Cite HW decode as barrier

4

AI Has Crossed from Experiment to Infrastructure.

AI maturity signal. 60% of respondents already deploy AI/ML in at least one workflow. 53% plan expansion in 2026. When asked in an open-ended question what force will shape encoding's future, 43.8% named AI unprompted, more than 2x any other response.

The more important signal: AI is migrating from auxiliary applications (transcription, classification) to core encoding intelligence, content-aware ladders (+77% planned growth), QoE prediction (+50%). Encoding solutions without an AI compute story face accelerating displacement. +77% growth in content-aware ladders.

**43.8%**

Named AI as future force (unprompted)

53%

Plan AI expansion 2026

60%

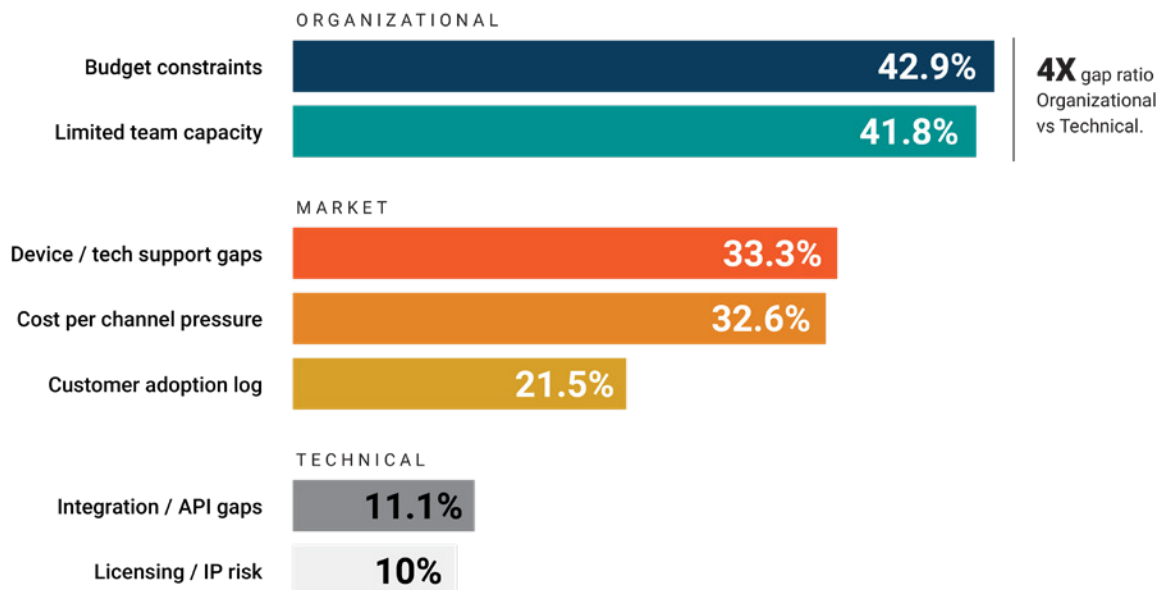
Already deploy AI / ML

5

The Binding Constraint Is Organizational, Not Technical.

The real enemy. This may be the most category-defining finding in the report. Budget constraints (42.9%) and limited team capacity (41.8%) outpace every technical barrier by a factor of four. Integration/API gaps rank last at 11%.

The market's real enemy is not inadequate technology. It is complexity, friction, and infrastructure that demands more people and capital than most teams have. 49% of all organizations, including those at companies with 5,000+ employees, run video teams of five or fewer people. Ease of deployment is the obstacle.



77%

Cite at least one org barrier

4x

Organizational vs. Technical ratio

49%

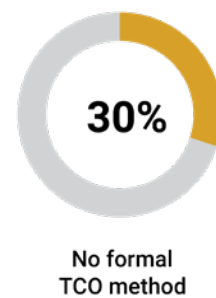
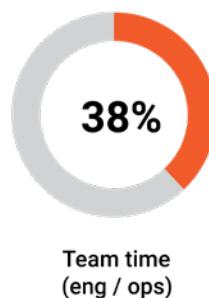
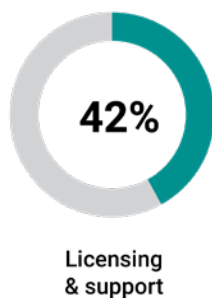
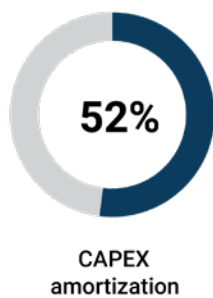
Teams of 5 or fewer

6

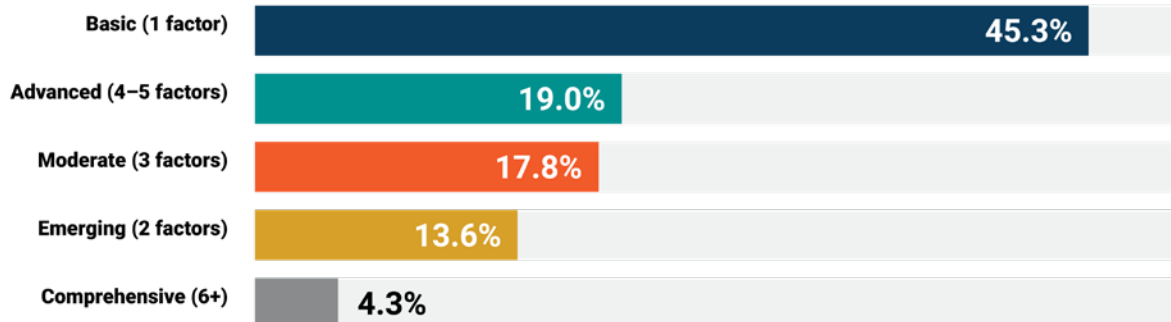
The TCO Blind Spot Is the Industry's Biggest Vulnerability.

The measurement gap. 30% of respondents have no formalized TCO methodology. 27% track zero unit economics. These are not small operators. This gap correlates directly with the largest market segment: Methodical Evaluators (37.4% of the market). They are making hardware investment decisions: codec selection, vendor choices, infrastructure architecture, without quantitative feedback.

The vendor who gives this segment a path to measure, wins their trust before any benchmark conversation begins.



TCO SOPHISTICATION DISTRIBUTION



30%

No TCO methodology

27%

Track zero unit economics

37.4%

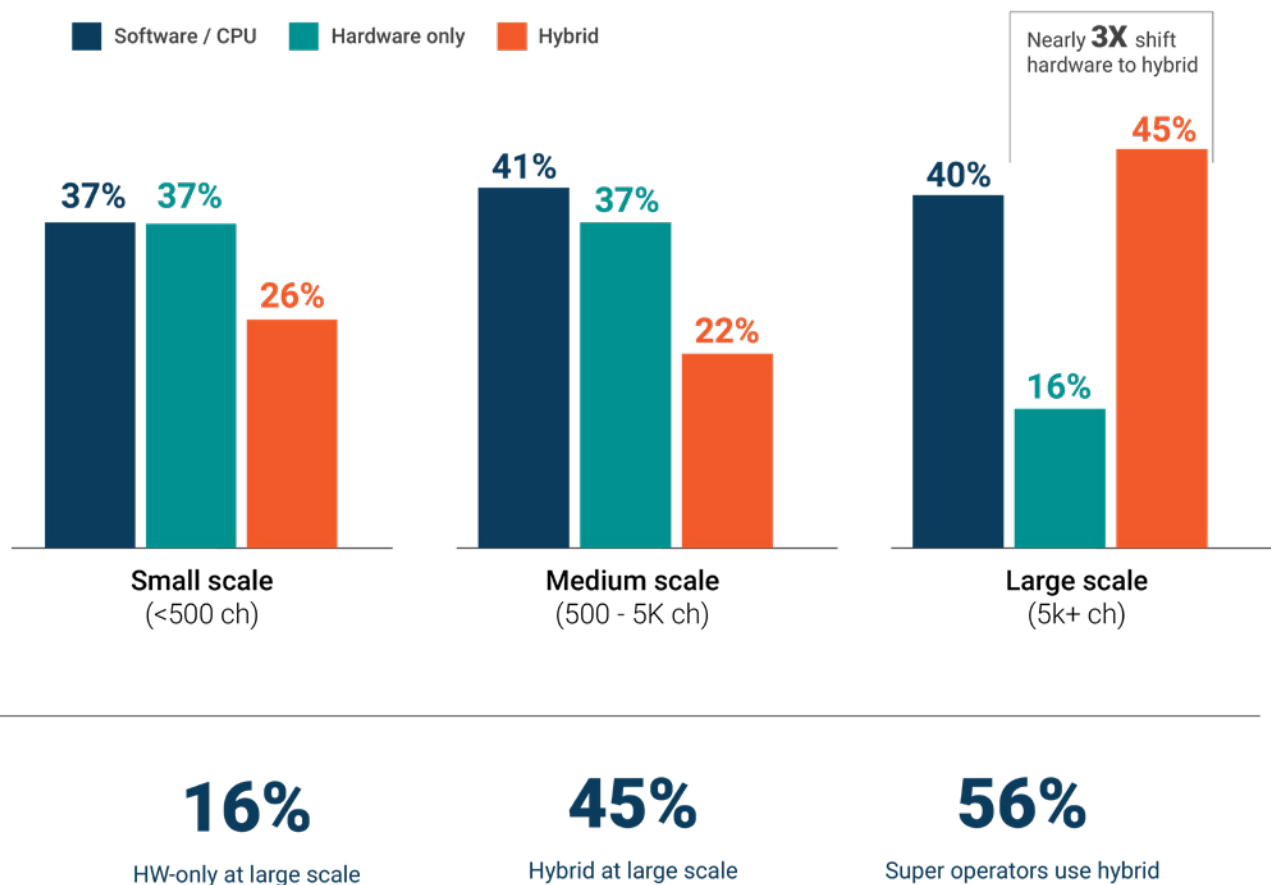
Methodical Evaluators hit hardest

7

At Scale, Hybrid Wins. Hardware-Only Loses.

Infrastructure architecture. As organizations scale to thousands of live channels, they increasingly rely on a mix of software and purpose-built hardware encoding. At large operational scale (5,000+ live channels), hardware-only encoding drops to 16% while hybrid approaches surge to 45%, nearly a threefold shift. Software handles long-tail content. Purpose-built hardware handles high-density live channels. A hybrid orchestration layer routes workloads to the most cost-effective resource.

Vendors offering only one approach face a shrinking addressable market as organizations scale.

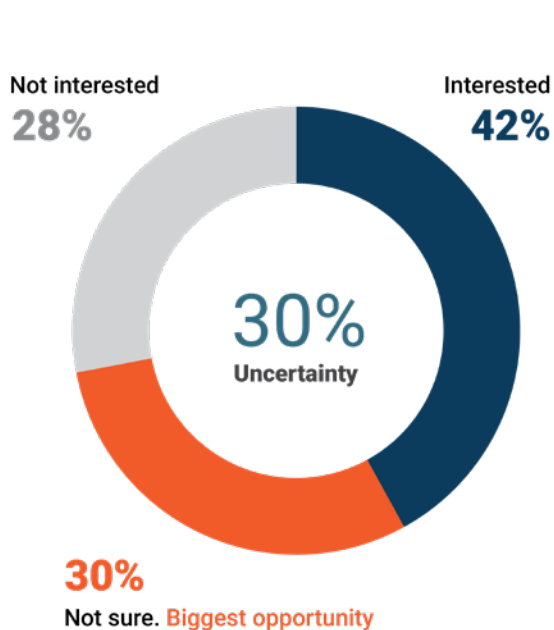


8

Edge is a Developing Market Waiting for a Reason.

Latent demand. 42% of respondents express interest in edge encoding. 30% say they are “not sure.” That state of uncertainty of nearly one-in-three professionals is not skepticism. It’s actually an absence of information.

Localized ladder generation (56%) and dynamic ad insertion (48%) lead stated use cases. Edge interest correlates with operational scale: operators running 250+ live channels show 50% edge interest versus 35% at small scale. The market exists. It needs architecture, use cases and a proof path.



TOP EDGE USE CASES

Localized ladder generation
56%

Dynamic ad insertion
48%

AR / VR applications
21%

Cloud gaming
10%

EDGE INTEREST BY OPERATIONAL SCALE

250+ live channels 50%

1–10 live channels 35%

42%

Interested in edge encoding

56%

Want localized ladder generation

30%

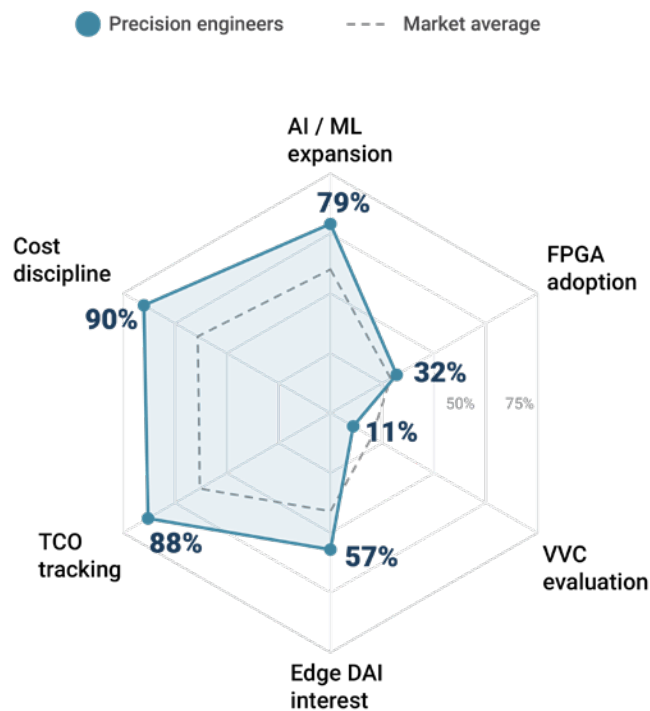
Uncertain = Educational opportunity

9

Precision Engineers: Small in Number, Yet Large in Influence.

Segment leaders. At 12.9% of the market, Precision Engineers punch well above their segment share with their hands in live sports, betting and mission-critical broadcasts. They lead on AI/ML expansion intent (78.8%), FPGA adoption, VVC evaluation, edge DAI interest, and TCO sophistication.

Precision engineers are the lighthouse customers whose infrastructure decisions navigate the rest of the market, as others watch and follow.



ARCHETYPE PROFILE

12.9%

Market share

78.8%

AI / ML expansion intent

2.6

TCO factors tracked (highest)

32.4%

Rank cost reduction as #1 priority

64.9%

Build-oriented (highest of any archetype)

0%

Outsource — none

12.9%

Share of market

78.8%

AI/ML expansion intent

0%

Outsourced. All build or hybrid.

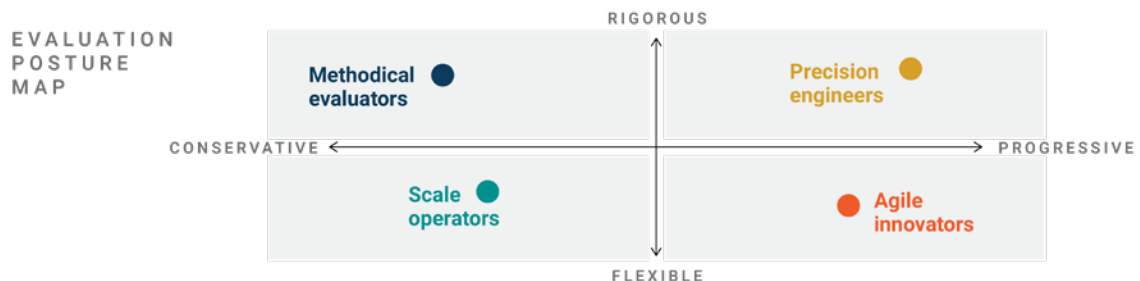
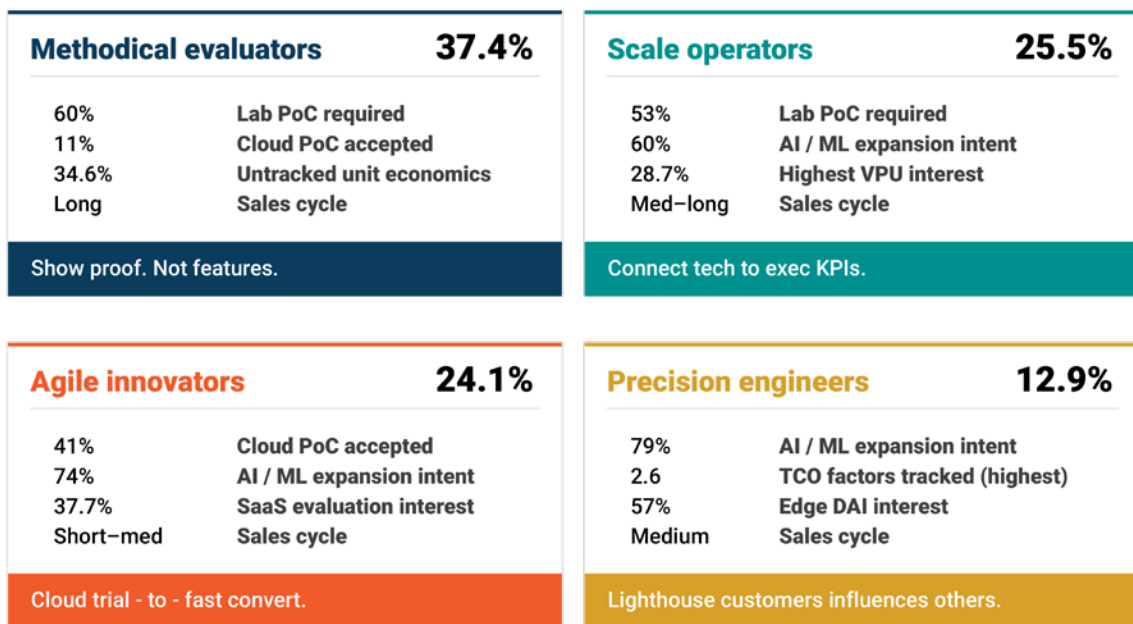
10

Four Buying Personalities. One Default Approach Misses 75%.

Decision influence. Statistical clustering reveals four archetypes:

1. Methodical Evaluators (37.4%) require lab hardware and documented proof.
2. Scale Operators (25.5%) need turnkey integration and executive ROI language.
3. Agile Innovators (24.1%) will try cloud-first and move fast.
4. Precision Engineers (12.9%) demand latency specs, flexibility guarantees, and measurable cost models.

One evaluation path, one message, one motion captures one quarter of the market.archetypes, a 4x gap. Segmentation isn't optional.



37.4%

Methodical Evaluators
Largest segment

4x

Cloud PoC gap
(10.7% - 41.3%)

12.9%

Precision Engineers
= Influencer

10 Insights That Matter

The market is not waiting for a better encoder. It's waiting for infrastructure that removes the organizational friction, delivers measurable economics, and scales from today's codec stack to tomorrow's, without forcing a choice between cost and quality.

That is the category NETINT is building.
These 10 data points are the demand proof.



2026 State of Video Encoding Industry Report
NETINT Technologies, Q4 2025

Download complete 47-page report.

You read where the industry was headed. **Now, tell us where it's going next.**

Last year, roughly 300 video engineering professionals shared what they were seeing, building and betting on. The results became a benchmark for the industry.

**Now we're asking what changed? What surprised you?
And what do you see coming next?**



Fill the survey

Scan to shape the next industry report.

The Future of Hardware Acceleration

Discover a smarter way to transcode
by lowering your cloud costs for
media-intensive tasks.





When Economics Break: A Framework for Escaping the Video Compute Trap

How Specialized Silicon and Distributed Compute are Rewriting Video Infrastructure Economics.

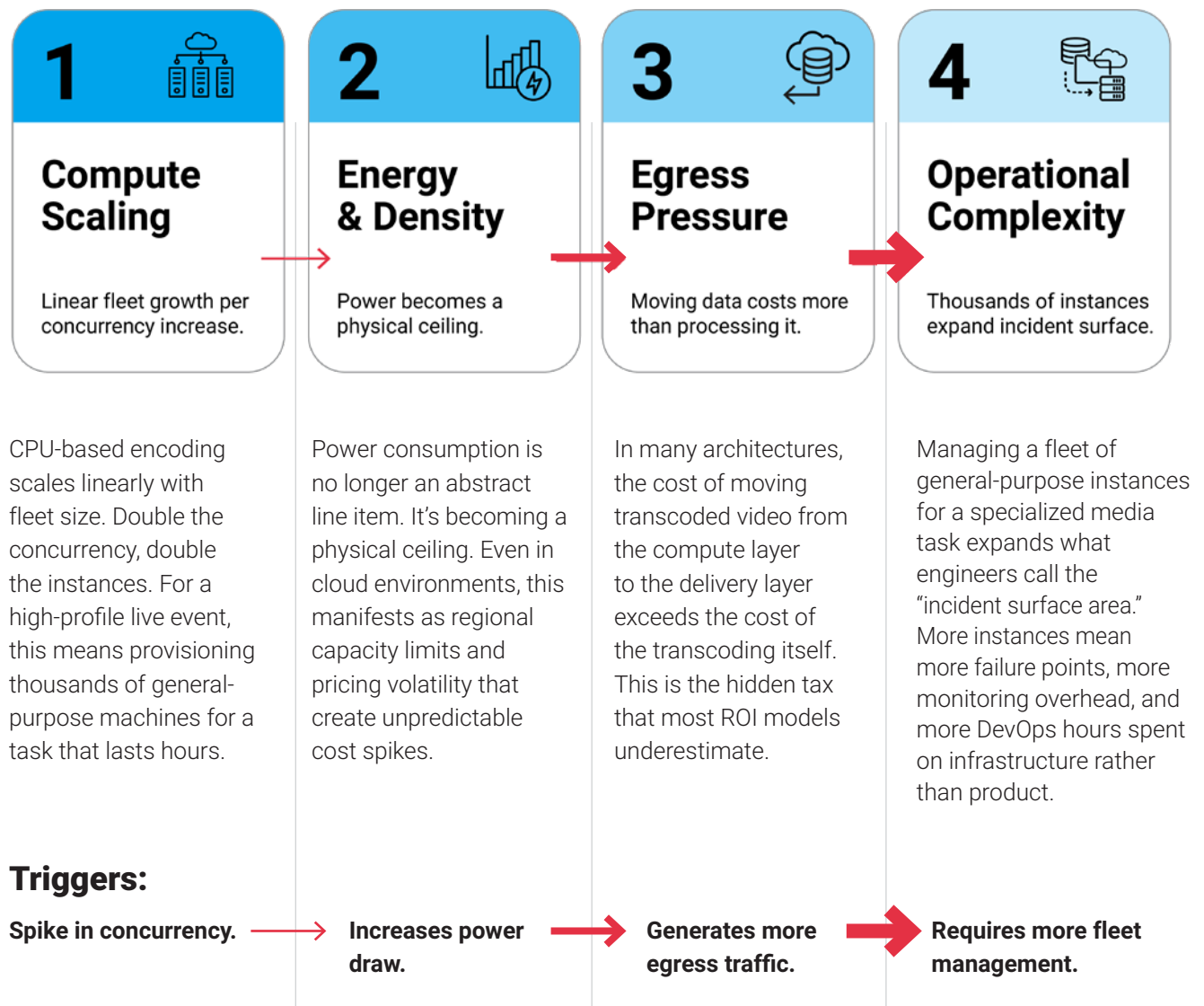
The business of delivering video at scale has a structural problem that no amount of cloud provisioning can solve. Every time a media organization adds a resolution tier, extends its ABR ladder, or pushes into a new geography, it triggers a cascade of infrastructure costs that compound in ways most planning models fail to anticipate. Compute grows. Egress bills multiply. Power ceilings tighten. And the operational burden of managing thousands of general-purpose instances quietly erodes the margins that content was supposed to generate.

For years, the industry response was straightforward: add more instances, negotiate better rates, and treat the problem as a procurement exercise. That approach worked when video was a fraction of total internet traffic. It does not work when video is the majority of it.

The organizations now gaining a structural advantage are those rethinking the architecture itself. Rather than scaling general-purpose compute horizontally, they are pairing purpose-built video processing hardware with compute environments designed for media economics. The combination of NETINT VPU-accelerated transcoding and Akamai's Distributed Cloud represents one of the most complete expressions of this shift.

Four Cost Levers That Break Video Infrastructure.

These four levers drive infrastructure costs in video and together, they create a compounding effect that makes linear scaling unsustainable.



These 4 levers do not operate independently.
The compounding nature of this cycle is what
makes incremental optimization insufficient.

Media-first architecture: From ingest to delivery.

The Akamai and NETINT VPU-accelerated VMs address these levers by replacing the general-purpose compute model with a five-stage pipeline built specifically for video workloads. Each stage maps to a specialized Akamai service, creating an end-to-end system where acceleration is applied where it matters most.

The pipeline begins with ingest and accelerated transcoding on Akamai Accelerated Compute instances equipped with NETINT Quadra T1U VPUs. Unlike general-purpose CPUs, these processors are purpose-built for video codecs (H.264, HEVC, AV1). The result is predictable performance where capacity is planned in “streams per footprint” rather than “instances per peak day.”

Transcoded fragments then move to workflow and origin management through Akamai Media Services Live (MSL). MSL serves as the bridge between compute and delivery, acting as a high-availability origin that manages stream configurations and ensures integrity before distribution.

At the packaging and manifest stage, raw transcoded rungs are wrapped into delivery formats like HLS or DASH. Manifests can be dynamically manipulated for server-side ad insertion (SSAI), multi-language audio tracks, or personalized stream routing.

Final delivery is handled by Akamai Adaptive Media Delivery (AMD), leveraging one of the world’s largest distributed edge networks for low-latency, high-quality playback regardless of geographic distance from the origin.

The final stage wraps the stream in a security and optimization layer that includes token authentication, geo-blocking, and DDoS protection. Concurrently, Common Media Client Data (CMCD) provides real-time feedback from the player to the network, improving buffer management and bitrate switching.

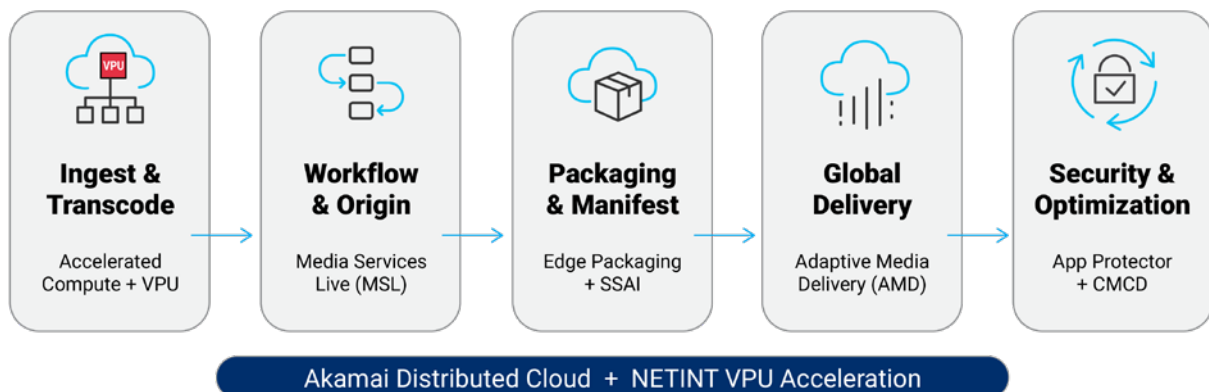


Figure 2: Five-stage reference architecture mapping to Akamai services.

Operational reality: how this fits into modern workflows.

A reference architecture is only as valuable as its deployability. This design is built to fit into modern CI/CD workflows through Kubernetes worker nodes and Terraform, allowing teams to treat their VPU-accelerated transcoding fleet as infrastructure as code.

The density of VPUs makes them well-suited for steady state workloads like 24/7 linear channels, where predictable throughput matters more than burst elasticity. For high-profile live events that demand rapid scaling, the cloud-native nature of Akamai's platform provides the burst capacity. Placing transcoding compute in the same regions as MSL origins minimizes latency and maximizes the reliability of the ingest-to-origin handoff.

A defensible decision framework.

The transition to VPU-accelerated transcoding on Akamai's Distributed Cloud is not a hardware upgrade. It is a strategic realignment of video unit economics. By combining the specialized processing power of NETINT silicon with Akamai's global scale and low-egress compute environment, media organizations can decouple their viewership growth from their infrastructure costs.

The organizations that make this shift early will not just reduce their current bills. They will build a cost structure that compounds in their favor over time, creating a competitive moat that is difficult to replicate through procurement negotiations alone. In an industry where content budgets dominate strategic conversations, the teams that quietly fix the infrastructure economics underneath will be the ones with the most room to invest in what viewers actually see.

Numbers behind the shift.

Architecture diagrams are persuasive. Unit economics are conclusive. Three specific areas of measurable improvement emerge from this design, each addressing a different cost lever.

On power efficiency, the gains are dramatic. Published benchmarks show that VPU-accelerated nodes achieve energy efficiency improvements of 4x to 6x compared to GPU or CPU alternatives. Where a GPU may consume roughly 2.5 - 3.6 watts per stream, a VPU sustains equivalent quality at approximately 0.4 - 0.7 watts per stream. In high-density environments where energy cost is the primary proxy for total cost of ownership, this difference reshapes the entire planning model.

On quality and stability, VPUs provide a predictable quality floor across the entire ABR ladder. Using VMAF (Video Multi-Method Assessment Fusion) as the benchmark metric, hardware-fixed encoding functions deliver consistent scores rather than the "best-effort" variability of software encoding. During peak traffic, this consistency translates directly to user experience.

On egress economics, the architectural alignment of Akamai's compute and delivery layers creates a decisive advantage. Accelerated Compute offers egress rates as low as \$0.005 per gigabyte, roughly 18 times lower than typical cloud egress pricing. This effectively neutralizes the "egress tax" that forces many organizations to artificially limit their ABR ladder quality.

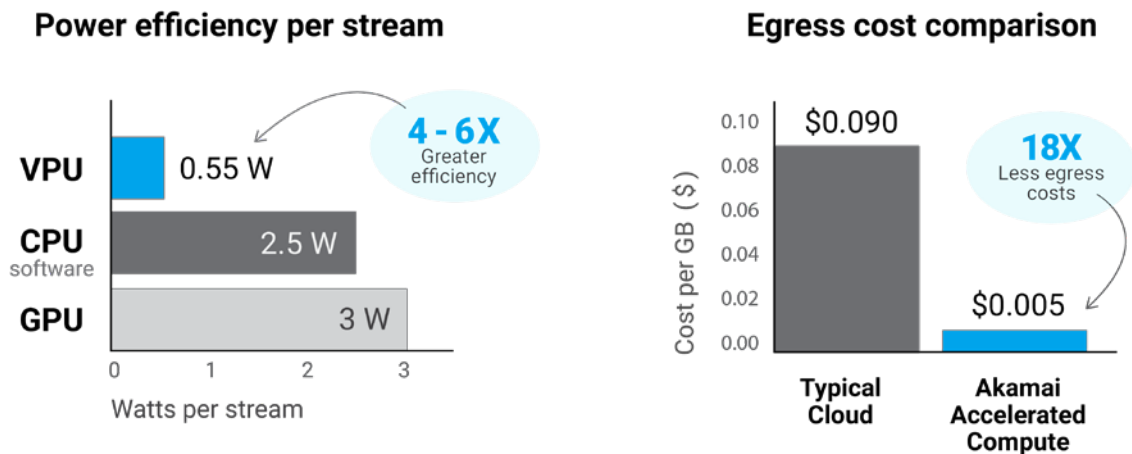


Figure 3: Power efficiency and egress cost comparison. DATA: published Cires21 + NETINT benchmarks).

When egress costs drop by an order of magnitude, the calculus of ABR ladders change entirely.

Organizations can afford to include higher-bitrate rungs that were previously cost-prohibitive, directly improving the viewer experience without proportional cost increases.



Reduce time to action on live video feeds

Wowza Video Intelligence Framework.
A real-time intelligence layer for live video that you control.

LIVE



CAM 118 · HWY

20:41:32.817

VEHICLE

type: car
speed: 78 mph
lane: 3

ACTION

🔔 Speed Alert Webhook →

LIVE



CAM 032 · FACILITY

18:33:10.662

UNAUTHORIZED ACCESS

zone: server_room_2
confidence: 0.96
badge: unknown

ACTION

🔔 Access Alert Webhook →

LIVE



CAM 023 · STADIUM

21:07:55.214

CROWD SURGE

section: 120
density: high
movement: surge

ACTION

🔔 Crowd Alert Webhook →

LIVE



CAM 009 · SUBSEB

09:12:44.905

ANOMALY DETECTED

asset: pipeline_17
anomaly: crack
severity: medium

ACTION

🔔 Inspection Alert Webhook →



Detect. Understand. Act.

Turn live video feeds into structured event signals, webhooks, and automated workflows in the systems you run today.



Designed For Model Flexibility.

Ships with production models out-of-the-box, with support for VLMs and custom-trained models. Change your models, not your architecture.



Built Into Wowza Streaming Engine.

Deploy video intelligence on-premises, at the edge, or fully air-gapped. You control the environment and what gets sent to the cloud.

200,000+ SERVER INSTANCES • 170+ COUNTRIES • 35,000 DEPLOYMENTS • NO NEW CAMERAS. NO SEPARATE AI STACK. NO REBUILD.



See Wowza and VIF live
in the NETINT booth

VPU Ecosystem Pavilion • IBC 2026 • [wowza.com](https://www.wowza.com)





Maximizing Encoding Efficiency in Live Streaming Workflows

You are not trying to win an encoding benchmark. You need to keep a streaming workflow running. That is where encoding efficiency is critical, and the practical question Wowza and NETINT are answering.

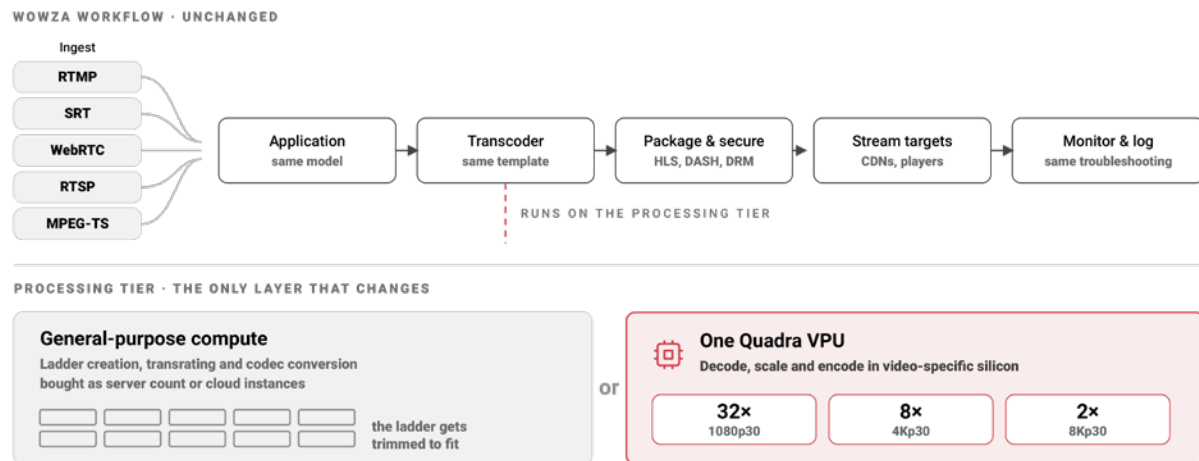
The source has to arrive. The application has to recognize it. The transcoder must produce the right renditions. The keyframes have to line up, the outputs have to behave with the endpoints, and the logs have to explain what happened when something fails. Efficiency only counts if it survives all of that.

The workflow is where efficiency becomes real.

Streaming engineers operate workflows, not chipsets. In Wowza, that workflow starts with a stream, ingested over RTMP, SRT, WebRTC, RTSP, or MPEG-TS. From there it may be passed through, packaged, transcoded, transrated, recorded, secured, and pushed to CDNs or players.

That is why encoding efficiency has to be judged in context. A faster encoder is useful only if the team can still configure the workflow, generate outputs, automate the process, monitor the stream, and troubleshoot results. If acceleration turns a known workflow into a special-case project, it creates more work, but if it fits into the workflow the team already understands, it creates value.

Figure 1. Efficient processing has to fit behind the same applications, templates, and stream targets a team already operates.



Where encoding costs actually show up.

Encoding cost is rarely a single line item. It shows up as a server count or cloud instance cost. It is the reason a team limits the number of renditions in a ladder.

If a source only needs repackaging, the efficiency gain is limited. If the workflow requires transrating, codec conversion, or ABR ladder creation, the processing tier becomes important. That is where a VPU belongs in the discussion. With a single Quadra VPU able to encode 32x 1080p30, 8x 4Kp30 or 2x 8Kp30 streams, the operational cost reduction is easily understood.

AI is the next workload.

The processing tier is about to do more than transcode. The next workload arriving in live streaming is inference: extracting frames from a live stream, running AI models against them, and turning what happens on screen into metadata, clips, and alerts while the stream is still live. That work lands in exactly the same place transcoding does at the processing tier, and it multiplies the compute attached to every channel. A workflow that could barely afford its ABR ladder cannot afford an inference pipeline on top. Efficient video processing is what creates the headroom for intelligence.

This is no longer hypothetical. Wowza Video Intelligence Framework (VIF) runs alongside Wowza Streaming Engine and connects live streams to AI inference, producing structured outputs - webhooks, ID3 tags, timestamped clips, and event signals that downstream systems can act on when required. Like the streaming workflow itself, it can be deployed on-premises, at the edge, in the cloud, or in a hybrid environment.

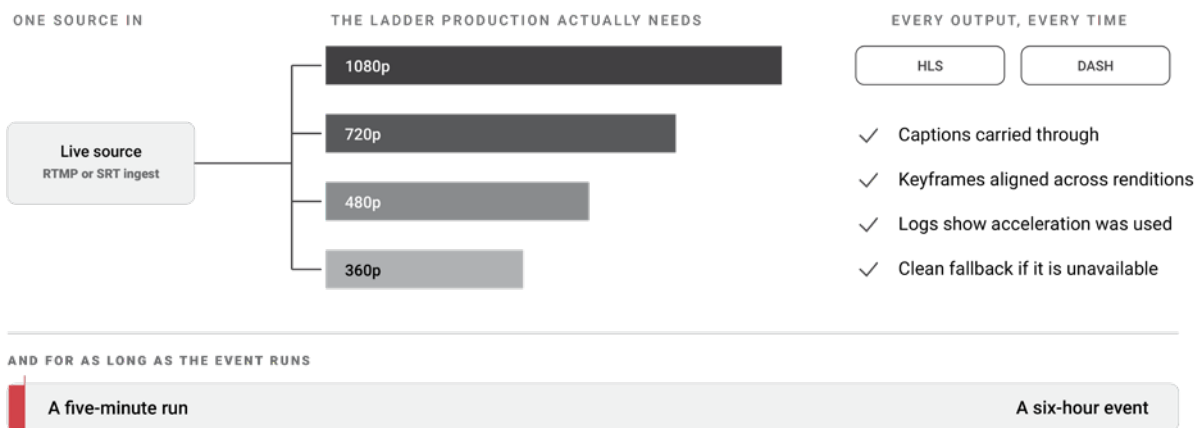
The compute question is where this article started: whether the processing tier can carry the workload without adding servers? With Wowza, the answer is yes.

The real test is the ladder.

A useful test starts with the ladder the customer needs. If production requires 1080p, 720p, 480p, and 360p, test that. If it needs HLS and DASH, test both. If the stream carries captions or requires aligned keyframes, test those. If the event runs six hours, do not rely on a five-minute run.

The benchmark should not answer only how many streams a device can encode. It should answer whether the Wowza workflow can produce the outputs the team needs, for as long as they are needed, with fewer servers or lower infrastructure cost.

Figure 2. Test the full ladder, not one stream. One source in, the renditions a real audience needs out, for the length of a real event.



What should not change.

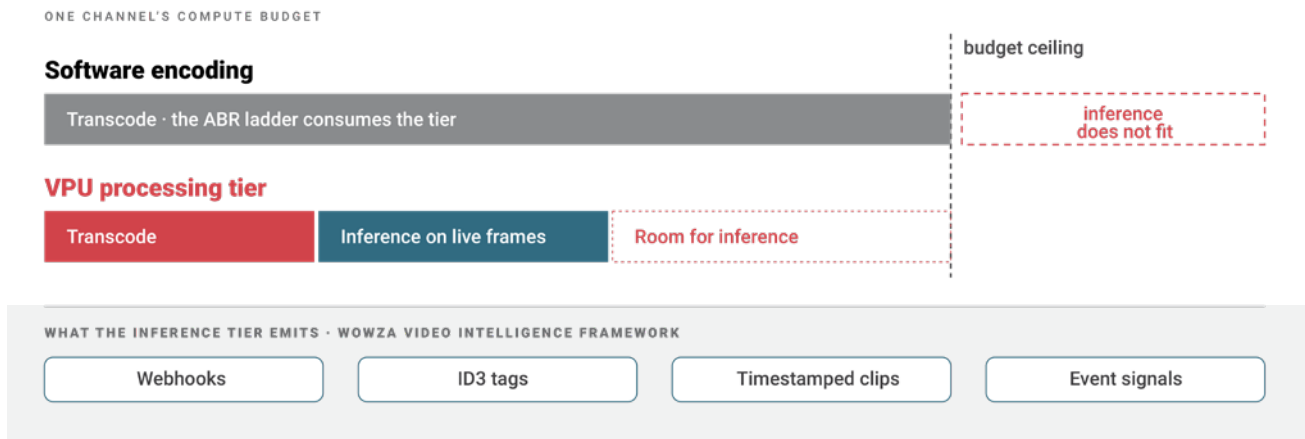
The streaming team should still think in Wowza terms: ingest, transcode, package, secure, deliver, log, and troubleshoot. A good acceleration path should not require reorganizing the entire workflow around the hardware. It should make the processing tier more efficient while preserving how the team builds and operates the service.

Several practical questions follow from that. Can the same application model be used? Can a transcoder template still define the required outputs? Can the workflow still be automated, and can the logs show whether acceleration was used? Can the system fall back cleanly if acceleration is unavailable? Those questions matter more than a peak stream count, and they should be answered before anything goes to production.

If a customer can ingest the same stream, create the same ladder, and deliver the same outputs with fewer compute resources, the value is real.

If the integration adds complexity that operations cannot support, the benchmark doesn't matter.

Figure 3. Inference lands where transcoding already lives. A workflow that barely affords it's ABR ladder cannot afford an inference pipeline on top. Efficient video processing creates headroom for intelligence.



Map the hardware to the deployment.

Wowza is often chosen because the team wants control over where the workflow runs: on-premises, private cloud, public cloud, at the edge, air-gapped, or behind a firewall. A processing tier is only useful if it exists where the workflow needs to run. So the hardware question and the deployment question belong together, and any VPU path should be evaluated for the specific Wowza deployment a team operates. When efficient, video-specific processing fits that model without forcing a new one, it can reduce the servers and instances needed to run the same streaming workload.

That pairing matters most at the edge and on-premises, where VIF deployments run inference close to the source. A device that decodes, scales, encodes, and runs inference in a couple of watts per stream is what makes intelligence-per-channel an operating cost a team can actually carry.



Sources & further reading

1. Wowza Streaming Engine documentation (Transcoder and workflows). <https://www.wowza.com/docs/about-wowza-transcoder>
2. Wowza Streaming Engine documentation (hardware acceleration overview). <https://www.wowza.com/docs/hardware-acceleration>
3. NETINT, VPU products and software integration (FFmpeg, GStreamer, SDK). <https://netint.com/products/>

VPU Capacity Without a Hardware Cycle

The question is no longer only whether to buy more servers. It is whether you can validate and deploy VPU capacity quickly where you need it.

The best way to test these models is with real workloads. NetActuate and NETINT give video teams a way to test, deploy, and scale VPU-powered encoding without requiring a traditional hardware procurement and configuration cycle.



Deploy NETINT Quadra VPUs in seconds



VMs, K8s, or Bare Metal across NetActuate's 45 Global POPs



Provides choice and a faster path from pilot to production



MEET WITH US

Book a time to meet
with us at IBC 2026.



LEARN MORE

NetActuate
powers NETINT's
VPU-as-a-Service



VPU Capacity Without a Hardware Cycle

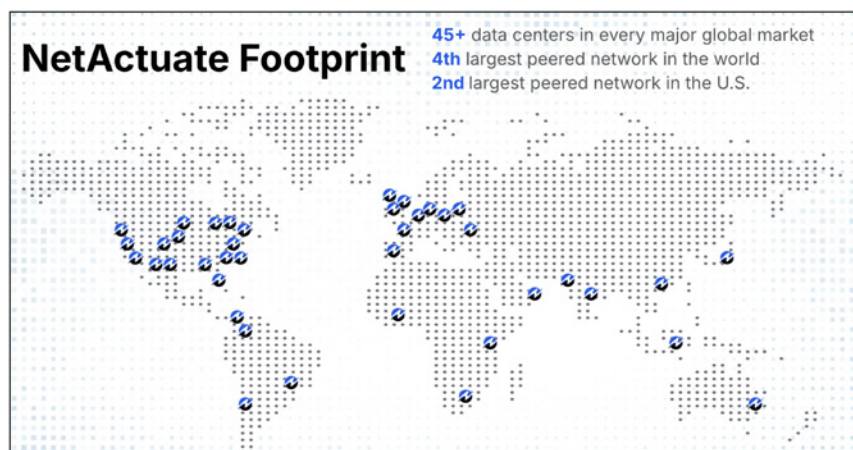
NetActuate and NETINT give video teams a way to test, deploy, and scale VPU-powered encoding without requiring a traditional hardware procurement and configuration cycle.

Encoding and transcoding workloads using VPUs are growing across the industry, but the path to capacity is rarely simple. Buying servers, validating hardware, securing power, choosing locations, and integrating into production can take six months or more. For some teams, that is acceptable. For others, the delay is a hold-up to competitiveness. That's where NetActuate and NETINT come together to provide a faster path from pilot to production.

Global capacity and delivery.

Announced in April 2026 and shown at NAB, NetActuate's VPU-accelerated infrastructure makes NETINT Quadra VPUs available across a global network. Workloads often need to run near users, near contribution sources, near exchange points, or near a customer's existing infrastructure. Latency, bandwidth cost, and routing all shape where processing should happen. NetActuate's distributed footprint makes that part of the conversation practical for reaching audiences around the world.

Instead of starting with a physical server project, a team can start with a VPU-enabled environment, deploy VM-based cloud compute, and run real encoding pipelines the same day.



VPU Deployment.

In the NetActuate portal, customers attach NETINT Quadra VPUs to virtual machines to run workflows and transcoding pipelines. After choosing compute type, firmware, operating system, and location, VPUs are attached to VMs and deployed as a service in less than one minute.

For larger, more complex deployments, public cloud, private cloud, managed Kubernetes, bare metal, colocation and custom deployments are all available.

As an incentive, NetActuate is offering qualified prospects a no-cost Proof of Concept (POC) environment for at least two weeks of testing and evaluation.

DEPLOY VIRTUAL SERVER

Deploy Location

CPU TYPE
General Compute

☒ Include Accelerators

SERVER LOCATION
North America
Raleigh, NC

Billing

BILLING CYCLE
USAGE / Usage

SERVER SIZE

RAM 8 GB

CPU 8 CPU

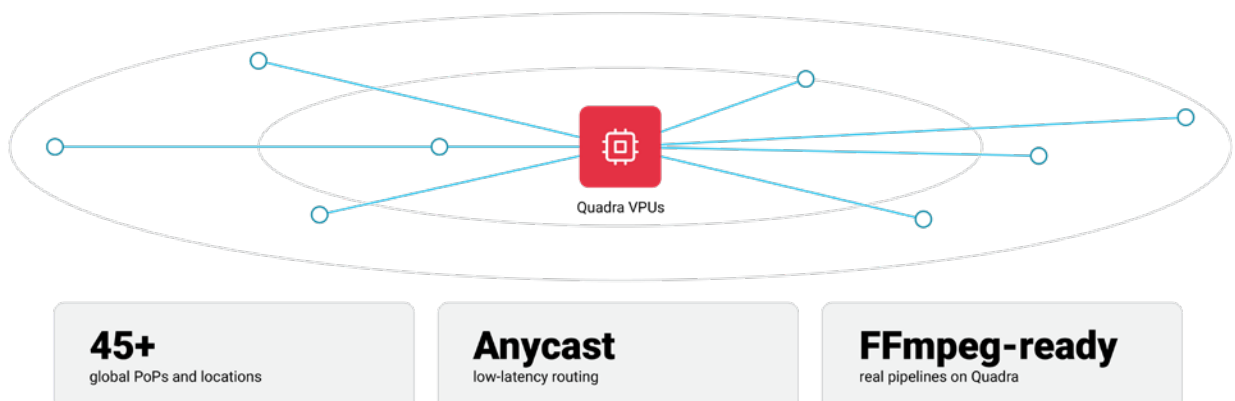
DISK 100 GB

VR8x8x100

\$173 / per month

ALSO INCLUDED

- ✓ Outbound transfer based on Contract
- ✓ Unlimited Inbound Transfer
- ✓ 10 Gbps Port Speed
- ✓ 1 IPv4 + 1 IPv6 Dedicated IP(s)



Why this is different from generic cloud transcoding.

While clouds excel at providing compute, they are not optimized for video. CPU-based encoding is up to ten times more expensive at scale, and GPU infrastructure is not always the right fit for high-volume encoding once power, cost, and availability enter the decision. Not to mention the availability issue with GPUs now that the data center has been overrun with AI workloads.

NETINT Quadra VPUs are purpose-built for video, designed to support high-density AVC, HEVC, and AV1 workflows. Delivered through distributed infrastructure, they give customers video-specific acceleration without building the physical layer themselves.

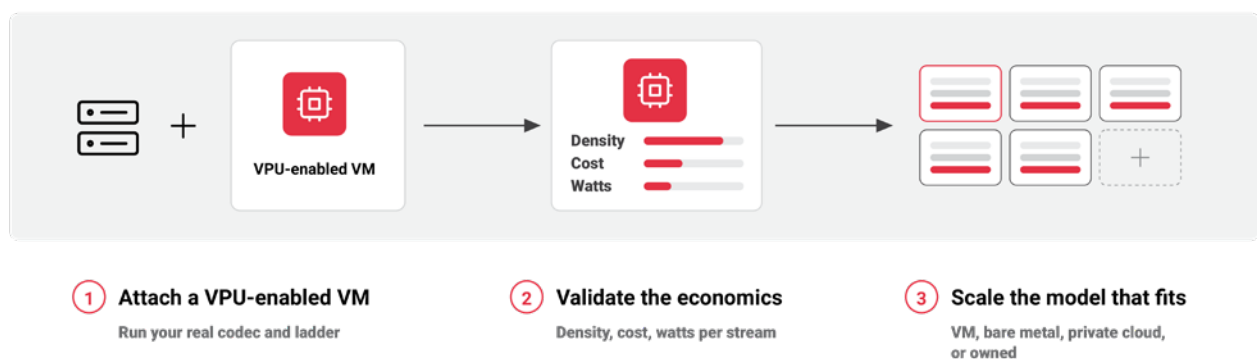
The question is no longer only whether to buy more servers.

It is whether you can validate and deploy VPU capacity quickly where you need it.

What engineers can test.

The best way to test these models is with real workloads. A meaningful evaluation does not stop at a generic benchmark. Engineers should test the actual pipeline they plan to run.

With a VPU-enabled environment, teams can measure how their workflows behave on Quadra, how many outputs they can sustain, and where the operational limits are. That evidence is valuable even if the final deployment changes, because it gives engineering a clear view of the VPU path and the business a clearer view of cost and density.



The NetActuate Media Marketplace.

To meet the demand of broadcasters, media & entertainment content owners, and video streaming platforms around the world, NetActuate has launched the NetActuate Media Marketplace. A systematic collection of media, streaming, and software appliances that can be deployed in one click to any of NetActuate's 45+ points of presence.

The Marketplace launches with media appliance images across media categories, including NETINT Bitstreams, a media processing framework that enables most of the video encoding and transcoding functions needed for modern video streaming workloads, alongside a deep open source library.

Instead of deploying a bare server, customers pick a solution, a location, VM size, and NetActuate handles the rest. The applications retain their license and run just as they would in a customized deployment over NetActuate's network.

This allows for hyperscaler-like capabilities at a fraction of hyperscaler costs. It's the perfect combination of operational flexibility and compute efficiency.

Meet NetActuate at IBC.

If you are validating hosted VPU solutions to reduce CPU transcoding costs, need to expand geographically, or are interested in new deployment models across the globe, book a time to meet with the NetActuate team at IBC:



Sources & further reading

1. NetActuate and NETINT deliver global VPU-accelerated infrastructure (PR Newswire, Apr 16, 2026). <https://www.prnewswire.com/news-releases/netactuate-and-netint-deliver-global-vpu-accelerated-infrastructure-for-cloud-scale-deployments-302744826.html>
2. NetActuate, NETINT VPU Acceleration (product page). <https://netactuate.com/netint-vpu-acceleration>
3. NetActuate, VPU Encoding with NETINT Quadra (technical guide). <https://www.netactuate.com/docs/guides/vpu-encoding-with-netint-quadra>
4. NETINT VPU Ecosystem. <https://netint.com/ecosystem/>
5. Hive Group, Future of Satellite at IBC. <https://wearehivegroup.com/ibc-show/where-does-the-signal-go-next>



Live video distribution workflows that hold **even when the network doesn't.**

Zixi is a live video platform for managing and delivering live video workflows across IP, cloud and hybrid environments.

zixi.com





Reliable Live Video Meets Efficient Processing

Live video does not fail politely. The real question is not whether a workflow can transport and transcode, but whether it can keep doing so as events, feeds, channels, and output profiles multiply.

If the contribution path drops packets, the audience sees it. If failover logic is unclear, the operations center sees it. If the transcoding tier cannot scale at event time, the business sees it. The engineering question is not whether live video can be moved or processed. It is whether the workflow holds together as it grows.

Zixi and NETINT solve real streaming workflow challenges. Zixi Broadcaster handles live ingest, transport, conditioning, failover, transmuxing, and transcoding. ZEN Master provides the control plane for orchestration, monitoring, alerting, topology, and API-driven operations. NETINT VPUs provide an efficient hardware path for the transcoding tier.

A live event cannot wait for a better instance type to free up, and a 500-event weekend cannot be run like five premium events with manual attention on every feed. The live stack has to be built for repeatability. Zixi reduces the workflow-control burden; NETINT reduces the transcoding-density burden. Zixi Broadcaster can recover from significant packet loss through adaptive error correction and hitless failover, which keeps the contribution path stable before processing even begins.

What each layer owns.

Think of Zixi Broadcaster as the media engine and ZEN Master as the operating layer above it. Broadcaster is the data plane: it receives inputs, handles transport and stream conditioning, manages failover, and can transcode or transmux as part of the workflow. ZEN Master is the control plane for managing Broadcasters, sources, channels, targets, events, telemetry, and automation.

That distinction tells engineers where to look when something breaks. Source behavior, transport state, failover, conditioning, and transcoding sit on the Broadcaster side. Orchestration, provisioning, monitoring, and alerting sit on the ZEN Master side. In production, that boundary matters; do not collapse them into one generic platform box.

Reference Architecture

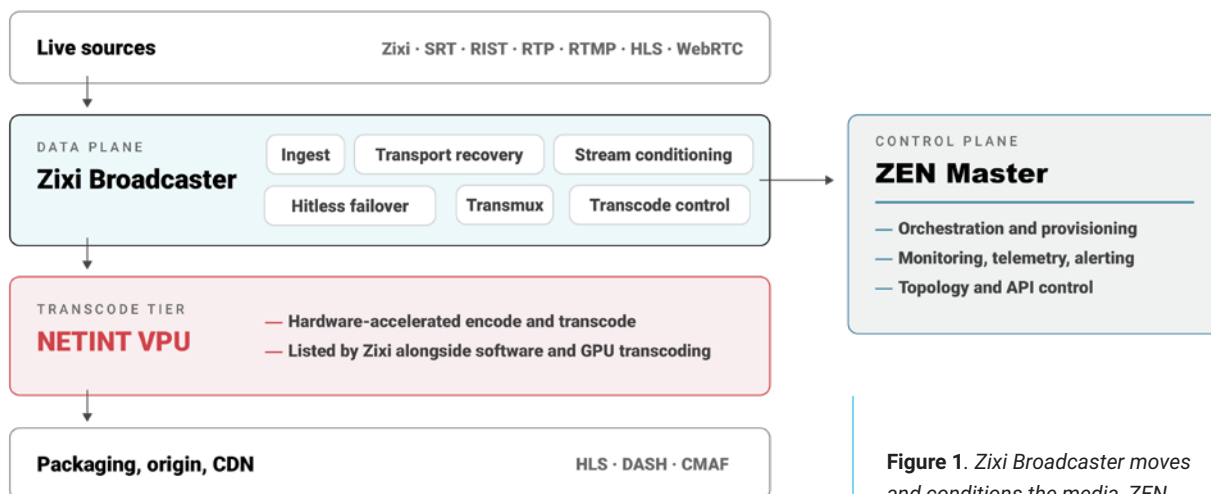


Figure 1. Zixi Broadcaster moves and conditions the media, ZEN Master controls it, and NETINT accelerates the transcoding tier.

The role of VPUs.

NETINT VPUs change the economics and density of the transcoding tier. It does not replace Zixi's workflow logic. When the workflow needs normalization, bitrate ladder generation, or event-based transcoding, the processing layer has to scale with the number of feeds. CPUs can do the work, but they are general-purpose. GPUs can also perform video processing tasks, but availability and instance sizing may not line up with live requirements, especially as cloud GPU demand is pulled toward AI.

A VPU is narrower by design. VPUs are built for video processing. If the workload is primarily encode and transcode, the question is why a live platform should consume general-purpose compute or compete for GPU and critical AI capacity when a purpose-built video processor is available with VPUs.

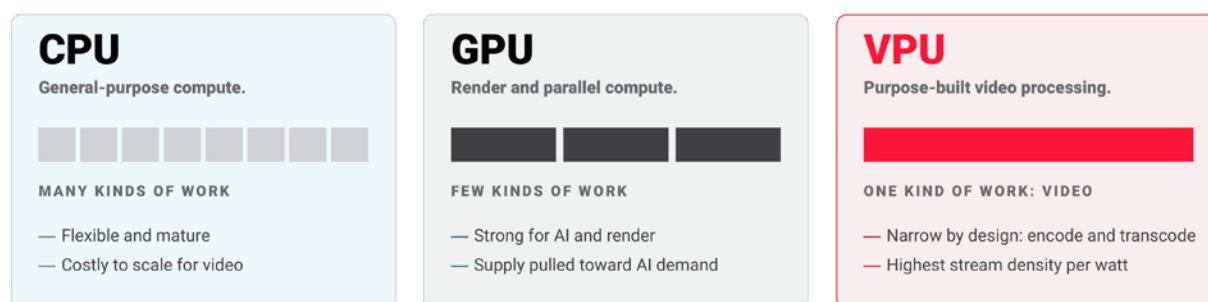


Figure 2. Choose by workload fit; if the job is mostly video transcoding, use hardware built for video transcoding.

The questions an engineer will ask.

Where does the VPU sit? In the transcoding tier. ZEN Master does not become the VPU layer; it controls and observes the workflow around the processing. Codec and profile support should follow Zixi's published Broadcaster matrix as the authority, and advanced codecs, 10-bit, and HDR behavior should conform to the specific workflow.

What should you benchmark? Not a single stream. Benchmark the ABR ladder with the deployed codec(s), resolution(s), frame rate, GOP structures, audio profiles, caption path, packaging, failover modes, and monitoring overhead. Include peak-event concurrency and warm-standby requirements, and report sustained frames per second, stream stability, recovery behavior, and cost per event or per channel.

Zixi reduces the workflow-control burden. NETINT reduces the transcoding-infrastructure burden.

Together they let live teams scale without treating transport, observability, and transcoding economics as separate projects.

If you run live events, the practical problem is not whether you have a protocol. You probably have several. The problem is whether you can ingest many feeds, normalize them, route them, monitor them, fail them over, transcode them, and deliver them without turning every new event into a custom operations project. This is exactly the optimum combination that Zixi and NETINT provide.



Sources & further reading

1. Zixi Software Platform (Broadcaster, ZEN Master, transcoding options). <https://zixi.com/zixi-software-platform/>
2. NETINT, VPU Ecosystem (Zixi contribution and transport). <https://netint.com/ecosystem/>
3. Zixi Broadcaster documentation (transcoding profiles). <https://docs.zixi.com/zixi-broadcaster-zec-current-version/creating-new-transcoding-profiles>



Stream more. Pay less.

i3D.net's private cloud and global edge network deliver:

<50ms

global latency

**Optimized
Routing**

as standard

↓ Egress

vs. public cloud
costs

EXCLUSIVE SHOW OFFER

Test a server for 30 days

Talk to our team to get started.



Scaling Performance: Critical Video Workloads

Not every video workload is defined by volume. Some are defined by latency, routing consistency, and reach. i3D.net and NETINT bring encoding into a performance infrastructure built for those demands.

Some workloads need capacity, but they also need consistency, low latency, predictable routing, and geographic reach. That is the profile of gaming, interactive streaming, real-time video, and live experiences, where the user experience depends on how fast the system responds.

NETINT provides purpose-built video processing with Quadra VPUs. i3D.net, the Europe-based infrastructure provider, brings the high-performance layer: managed private cloud, bare metal, a global low-latency network, and game services. Together they can support video workloads where encoding has to fit inside a larger performance strategy.

Capacity is not enough

For many platforms, adding capacity means adding encode or transcode throughput. That matters, but it is only part of the problem for performance-sensitive workloads. Gaming and real-time video have a different profile: traffic spikes during launches and peak windows, the workload often needs to run close to users, and routing consistency shapes the experience.

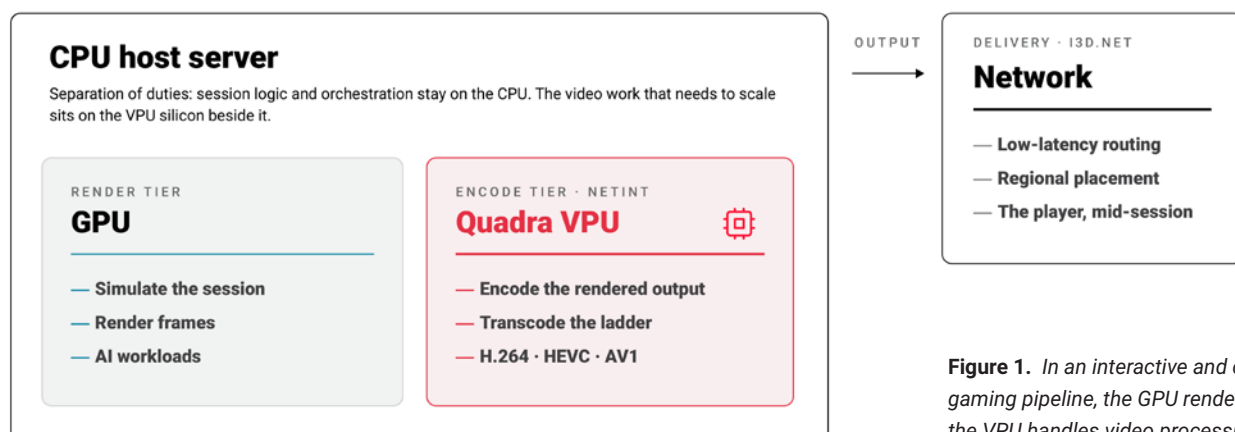


Figure 1. In an interactive and cloud-gaming pipeline, the GPU renders, while the VPU handles video processing.

In that environment, video processing cannot be treated as an isolated function. Encoding has to fit into the platform's performance model. That is why i3D.net's role differs from a general infrastructure provider. The value is not only that infrastructure exists, but that it is designed for demanding, latency-sensitive, globally distributed workloads.

Why VPUs matter in performance infrastructure.

Performance infrastructure has to use power, space, and compute efficiently. If video processing runs entirely on general-purpose CPUs, high-volume encoding consumes too much capacity relative to output. GPUs are valuable for rendering and AI, but they are not always the most efficient answer for every encode and transcode task, especially as GPU supply is pulled toward AI demand.

NETINT Quadra VPUs give infrastructure teams a video-specific processing layer. The architectural point is separation: use GPUs for rendering, simulation, or AI, and a VPU for encoding. A distributed platform can then place power-efficient encoding capacity closer to users without adding unnecessary CPU or GPU load.

i3D.net sits between public cloud and self-managed infrastructure, which is an important position for platforms that want more control than generic cloud but do not want to operate every layer themselves. A customer may need compute, managed operations, connectivity, flexible VMs, and DDoS protection, while NETINT VPUs provide the encoding engine. That lets a team treat video processing as part of a broader environment: where it runs, how it connects, how it scales, and how it is protected.

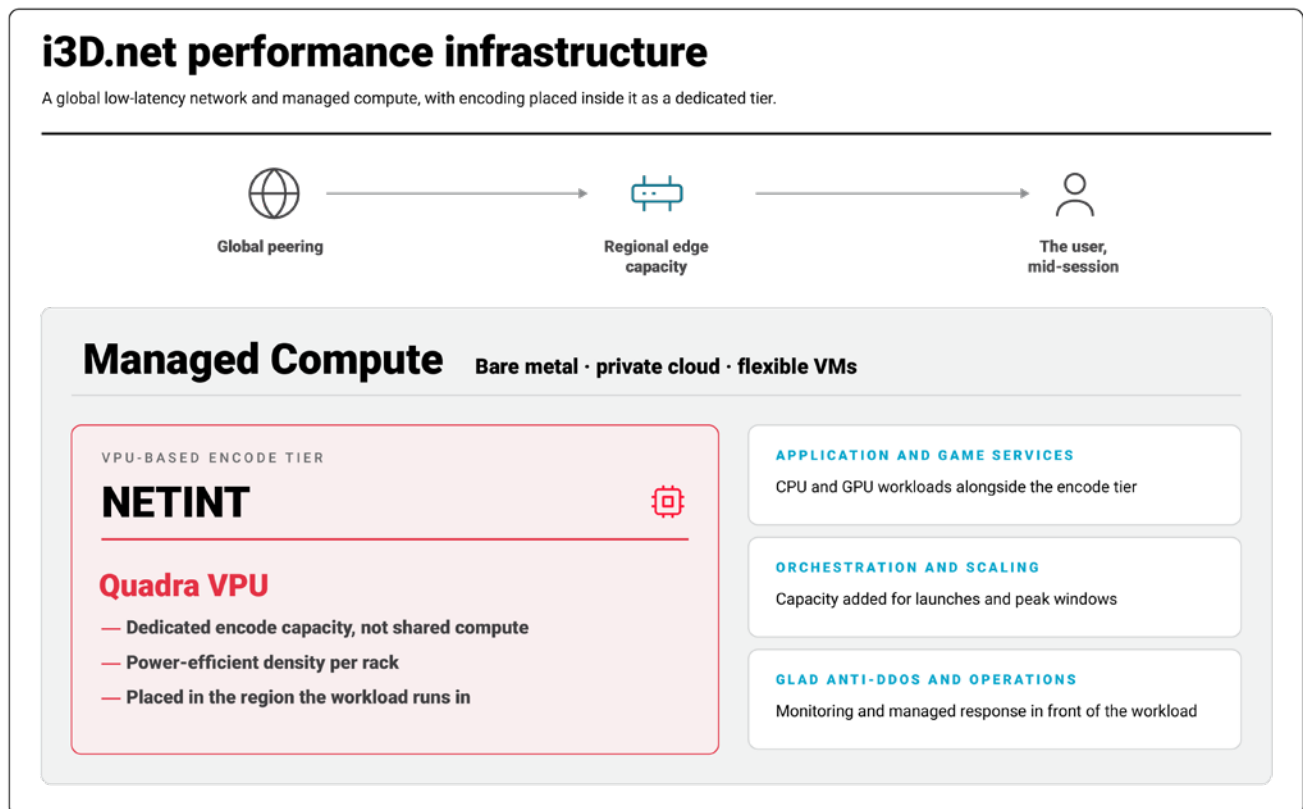


Figure 2. Encoding inside a performance infrastructure: global network and peering, managed compute, anti-DDoS and operations, with a VPU-based encode tier.

Where this model fits.

The story is strongest where video processing is tied to performance-sensitive infrastructure. A cloud-gaming provider may render gameplay, encode the output, and deliver it with low latency. An interactive platform may support real-time participation. A live experience may need regional capacity, predictable routing, and resilience under peak traffic. These are not generic transcoding scenarios. i3D.net brings compute, network, operations, and reach; NETINT brings the video processing layer that sits inside it.

Real-time and interactive use cases like cloud-gaming, a live participatory broadcast, or a real-time communications platform cannot tolerate the variability that comes from competing for shared compute. Placing a dedicated, power-efficient encode tier within managed infrastructure keeps the video path predictable even when traffic spikes during a launch or a peak window.

What engineers should evaluate.

A performance-critical workload has to be tested as a system. Encoding performance matters, but it should be measured alongside latency, network path, workload placement, scaling behavior, monitoring, and failure handling. A VPU can improve processing efficiency, but the production result depends on where and how it is deployed.

Engineers should consider the real workload: codecs, resolution, bitrate ladder, frame rate, latency target, concurrency, regional demand, routing, and orchestration. They should also define the VPU's role in the architecture, whether it handles live encoding, transcoding, cloud-gaming output, or regional media processing.

This is where the partnership gives the VPU story more weight. The question is not only whether VPUs work, but where VPU-powered processing should sit inside a performance infrastructure design. A gaming company that already trusts i3D.net for server infrastructure can evaluate whether a video-specific tier improves the encoding side of the workflow, without reorganizing the platform around it.

The user never sees the architecture. Users feel the result: lower latency, stable sessions, and consistent quality under peak load.

For performance-critical workloads, infrastructure problems surface as user-experience problems. Latency becomes gameplay delay, routing inconsistency becomes unstable sessions, and inefficient encoding becomes higher cost or lower density. The question worth answering is where video processing should live so the full experience performs the way users expect. That is the conversation i3D.net and NETINT can take on together.

Sources

1. NETINT, i3D.net in the NETINT VPU Ecosystem. <https://netint.com/vpu-ecosystem/i3d-net/>
2. i3D.net, Global network and infrastructure. <https://www.i3d.net/>
3. i3D.net, GLAD anti-DDoS solution. <https://www.i3d.net/anti-ddos-solution/>



High-Efficiency VPU Platforms Built for Live Streaming

Flexible 1U and 2U servers for reliable, scalable deployment



VEGA-7142

1U Video Edge Server

- Powered by the AMD EPYC™ 9006 SP8 Series processors
- Designed for video media servers & recorders

VEGA-7241

2U Video Edge Server

- Powered by AMD EPYC™ 9005 Series processors
- Up to 6 PCI Express accelerators or I/O cards per chassis
- Designed for field operations



Beyond General-Purpose Compute

Immersive video is about to turn every live event into millions of personalized streams. The infrastructure answer is not more servers; it is putting the right silicon in the right place.

When professional football adopted the Video Assistant Referee at the 2026 FIFA World Cup, broadcasting quietly crossed a threshold. Synchronized high-definition cameras, processed in near real time, proved capable of reconstructing physical reality with enough fidelity to overturn a referee's call. Call it the volumetric truth: not a picture of the game, but a manipulable digital twin of it. If a room full of officials can scrub through a reconstructed moment from any angle, why can't a fan?

The engineering vision taking shape across the streaming industry is essentially a personal VAR for every viewer: multi-camera capture feeding spatial AI, which uses techniques such as 3D Gaussian splatting to build a live, manipulable model of the event, which in turn delivers personalized virtual viewpoints to millions of concurrent users. Tactical overheads, courtside seats, a camera that follows one player all night.

The barrier is no longer data collection. What stands between the demonstration and the product is not imagination. It is infrastructure economics.

The immersion gap.

Here is the uncomfortable asymmetry at the heart of live sports production: the technology used to capture the game is roughly a decade ahead of the technology used to watch it. Broadcasters already track millimeter-accurate, real-time 3D physics on the field. While the fan in their living room receives a flat, directed, one-size-fits-all 2D feed. Capture has gone volumetric; delivery has not.

Closing that gap changes the arithmetic of streaming. In conventional broadcasting, adding viewers meant adding replicas of the same feed. In immersive and personalized broadcasting, one physical event generates thousands, eventually millions, of unique output streams, each of which must be reconstructed, rendered, and encoded. The multiplication effect is the opportunity. It is also the trap.

The trap of borrowed hardware.

The industry reflex, trained by a decade of cloud thinking, is to solve every scaling problem with a uniform pool of general-purpose compute. For immersive media, that reflex breaks down. If the pipeline relies on premium general-purpose GPUs to handle both complex spatial rendering and massive high-density video encoding, the total cost of ownership fractures. Expensive silicon built to understand a scene ends up spending its cycles on the routine, repetitive work of compressing one.

That is why the executive question has changed. It is no longer “Which server is fastest?” It is “Which workload belongs on which class of silicon, and how should the system be engineered so that silicon sustains its performance in production?”

This is the practical case for a third category of processor alongside the CPU and GPU: the video processing unit, or VPU.

Start with workload placement.

A modern immersive or live pipeline has three materially different jobs, and it helps to give each one a role. The CPU is the orchestrator: it ingests synchronized SDI or ST 2110 feeds, preserves timing and data integrity, and manages workflow and control logic.

The GPU is the artist: it applies computer vision, 3D Gaussian splatting or other reconstruction methods to turn camera feeds into manipulable scene data.

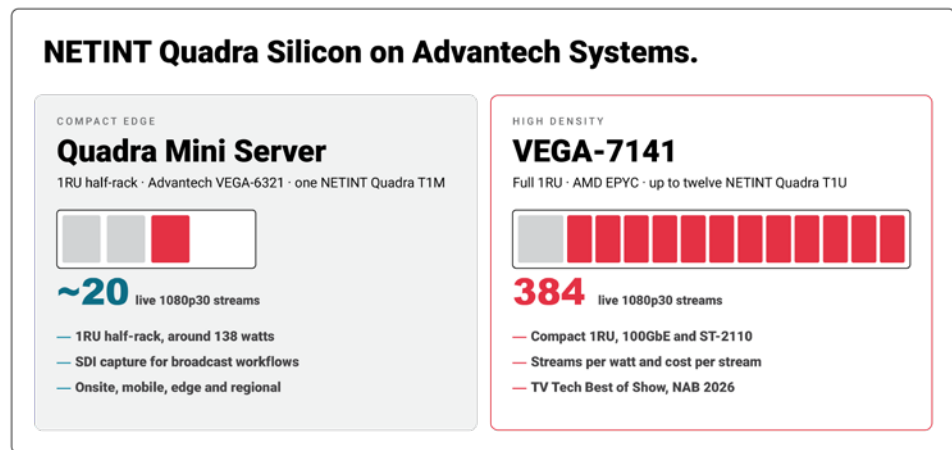
The VPU is the engine: purpose-built silicon that converts the outputs into AVC, HEVC or AV1 streams at the scale, latency and cost the service model can support.

Run all three jobs on one accelerator class, and you strand expensive resources. GPUs are superb at massively parallel mathematics, but encoding thousands of routine renditions competes directly with the reconstruction workload that creates differentiation.

CPUs remain essential for orchestration, yet general-purpose cores are not optimized for sustained, high-density encode. A heterogeneous design is more disciplined: the CPU orchestrates, the GPU reconstructs, and the VPU processes video.

Crucially, this is a system-level decision, not a shopping decision. It implicates PCIe topology, memory and network bandwidth, thermal headroom, power delivery, software interfaces and product lifecycle. An accelerator advertised for high stream counts creates little value if it is starved for data, throttled by airflow or tied to a short recertification cycle.

Figure 1.
A heterogeneous media architecture keeps orchestration, spatial compute and codec processing on the silicon best suited to each job.



Three nodes, not one monolith.

In deployment terms, the architecture resolves into three distinct nodes. Immersive edge capture lives at the venue, where synchronized feeds demand absolute precision timing, local NVMe storage and resilience; this node is the bridge between the physical camera lens and the digital twin, and its defining requirement is zero data loss.

Spatial AI compute lives in an on-premises data center or production truck, where premium GPU capacity is protected for volumetric reconstruction, inference and scene manipulation, thermally stabilized for continuous load. Media delivery lives at the outbound or cloud edge, where dense VPU capacity absorbs the repetitive codec work and personalized streams fan out to the audience.

The boundaries can move, but the jobs remain distinct, and separating them gives operators a cleaner scaling model. Capture scales with camera count. Spatial compute scales with scene complexity and latency targets. Delivery scales with renditions, viewpoints, and viewers. Decouple those axes and growth in one stage no longer forces overbuilding the entire pipeline.

Density is an operating metric.

For conventional video, streams per server was a useful measure. Under real rack, power, and cooling constraints, streams per rack unit and streams per watt are more revealing; in a power-constrained facility, a denser node can be the difference between deploying capacity and waiting for it. Advantech and NETINT illustrate the spectrum with two deployment profiles built on the same VPU silicon strategy.

At the compact end, the Quadra Mini Server packages a NETINT Quadra T1M VPU into Advantech's VEGA-6321-based 1RU half-rack platform, alongside a 13th Gen Intel Core CPU, DDR5 memory and NVMe storage, in a power envelope of roughly 138 watts. This small powerhouse is capable of processing 20 live 1080p30 streams across H.264, HEVC, and AV1, with optional SDI capture and support for FFmpeg, GStreamer, the NETINT SDK, and Bitstreams media processing framework. It answers a question many teams face: where to put dedicated VPU encoding when a full server deployment does not make sense, whether mobile, in-stadium or at the regional edge.

At the dense end, the Advantech VEGA-7141 supports up to twelve Quadra T1U VPUs on an AMD EPYC platform in a single rack unit. Capable of 384 simultaneous 1080p30 streams, with 100GbE networking and ST 2110 support, it's easy to see why the system received a TV Tech Best of Show Award at the 2026 NAB Show. The point is not the large number. It is that capacity, networking, cooling and power are engineered together as one production system, which is what makes density deployable rather than theoretical.

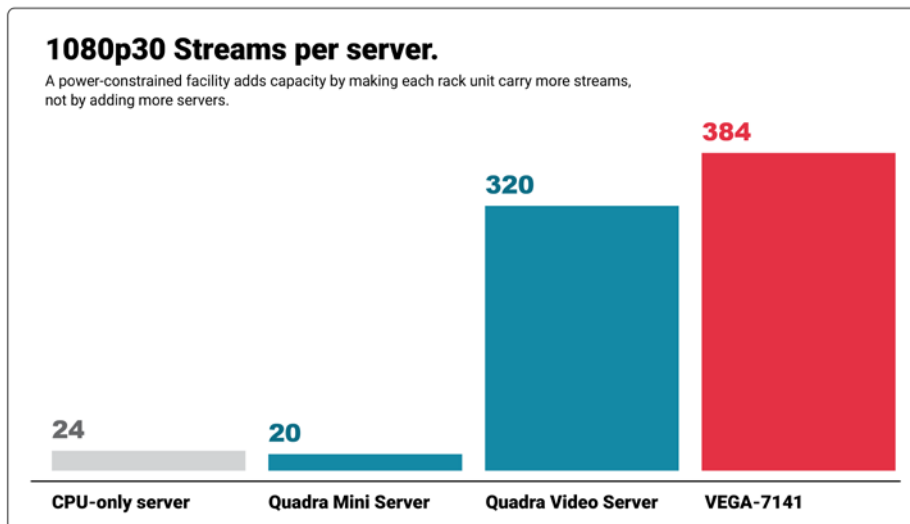


Figure 2. One VPU strategy serves both compact edge and high-density production deployments; the selection criteria are operational, not only computational.

The server determines whether the accelerator delivers.

The bottlenecks that kill real deployments rarely appear on benchmark charts. Accelerators rated for massive stream counts can be starved of data because generic PCIe topologies cannot provide sustained, non-contending bandwidth. Video processing runs 24/7, and a chassis designed for ordinary enterprise loads can recirculate heat around multiple accelerators until expensive silicon throttles and drops frames.

This is where the division of labor between the silicon topologies matters. NETINT supplies purpose-built video processing silicon; Advantech integrates it into media platforms where bandwidth, thermals, capture (SDI), and network interfaces are validated. For business leaders, that reframes the total cost of ownership conversation. Acquisition price matters, but so do rack expansion, power, cooling, recertification, deployment delay, and risks. A lower-cost server that forces more racks, or wastes GPU cycles on commodity encode, can easily be the more expensive choice.

Four questions for the next architecture review:

1. **Are scarce GPU cycles reserved for video encoding?**
If premium GPUs spend most of their time on routine encoding tasks, the workload map deserves another look.
2. **Can each node scale independently?**
Camera count, spatial complexity and output streams rarely grow at the same rate. Be ready to increase GPU capacity to accommodate new AI features. At the same time, be sure to leave room in the rack for VPUs when more channels or camera angles are needed.
3. **Are performance claims sustainable?**
Ask about data paths, thermals and power under continuous 24/7 load, not only peak lab throughput. The lab is never production. Don't assume "it worked in the lab" will carry through to the stadium or data center.
4. **Is the platform lifecycle aligned to the service?**
Broadcast and streaming products may outlive commercial IT refresh cycles. This means a server that would normally be depreciated over a 5-year time period could last 7 years or longer. Qualification costs must be modeled explicitly.

From accelerator choice to media system design.

The broader lesson is that immersive media is unlikely to scale through indiscriminate acceleration but through silicon neutrality, where each workload is routed to the compute engine built for it. VPUs remove a high-volume, repeatable workload from general-purpose CPUs and premium GPUs, making room for richer experiences without collapsing delivery economics. The winners of the immersive era will not simply own more compute. They will place compute more intelligently.



M P E G - 5

LC = VC

LCEVC integrates into existing workflows while extending their performance.

**MOST CODECS FORCE
STEP CHANGES IN
INFRASTRUCTURE.**

46%

BITRATE SAVINGS
WITHIN EXISTING PIPELINES

**FASTER
ENCODING**

LOWER COMPUTE OVERHEAD

**DROP-IN
COMPATIBLE**

STANDARD PIPELINES & PLAYERS

Mandated in DTV+ (TV 3.0)

and included in ATSC 3.0 specifications.

V-NOVA



The Next Layer of Video Efficiency

Processing efficiency and delivery efficiency are two levers on the same scaling problem. This edition adds the strongest published evidence, the strongest published objections, and the boundaries where each holds.

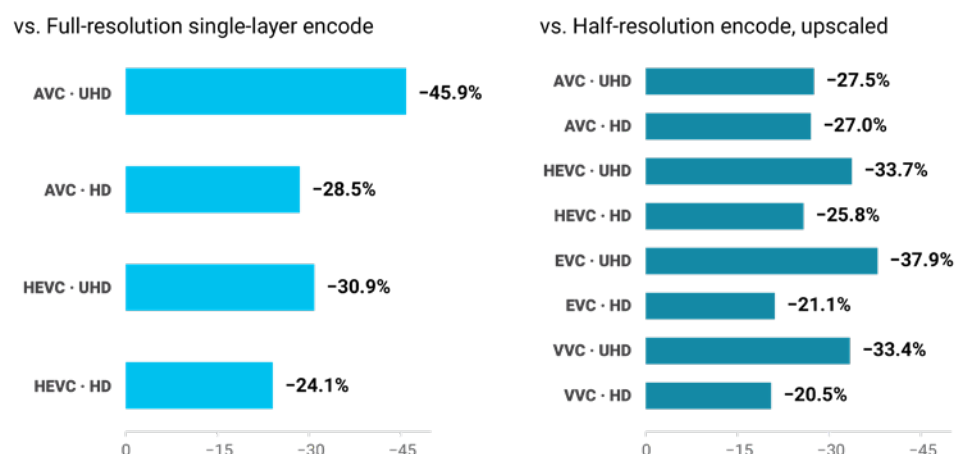
Video platform OPEX analysis starts with the visible costs: transcoding servers, cloud instances, power, cooling. NETINT addresses this layer with VPUs. Meanwhile, bitrate drives the second cost center consisting of CDN invoices, cache efficiency, time to first picture, rebuffering rates, and video quality on constrained networks. V-Nova addresses this with MPEG-5 LCEVC.

MPEG-5 LCEVC (ISO/IEC 23094-2) is a standardized enhancement layer, where the base codec, H.264, HEVC, AV1, or VVC, encodes at roughly one quarter of the output resolution. Lightweight residual sub-layers restore full-resolution detail as the enhancement exists as SEI metadata or a CMAF track. This enables stream packaging, CDN mechanisms, and DRM to remain relatively untouched, while devices without LCEVC player support simply render the base layer.

MPEG's own verification testing with BT.500 DSIS viewing (50 screened viewers) found that LCEVC saved 45.9 percent for AVC at UHD, 28.5 percent at HD, 30.9 percent for HEVC at UHD, and 24.1 percent at HD for HEVC in MOS BD-rate.

Figure 1. MPEG verification tests: subjective BD-rate against two anchors.

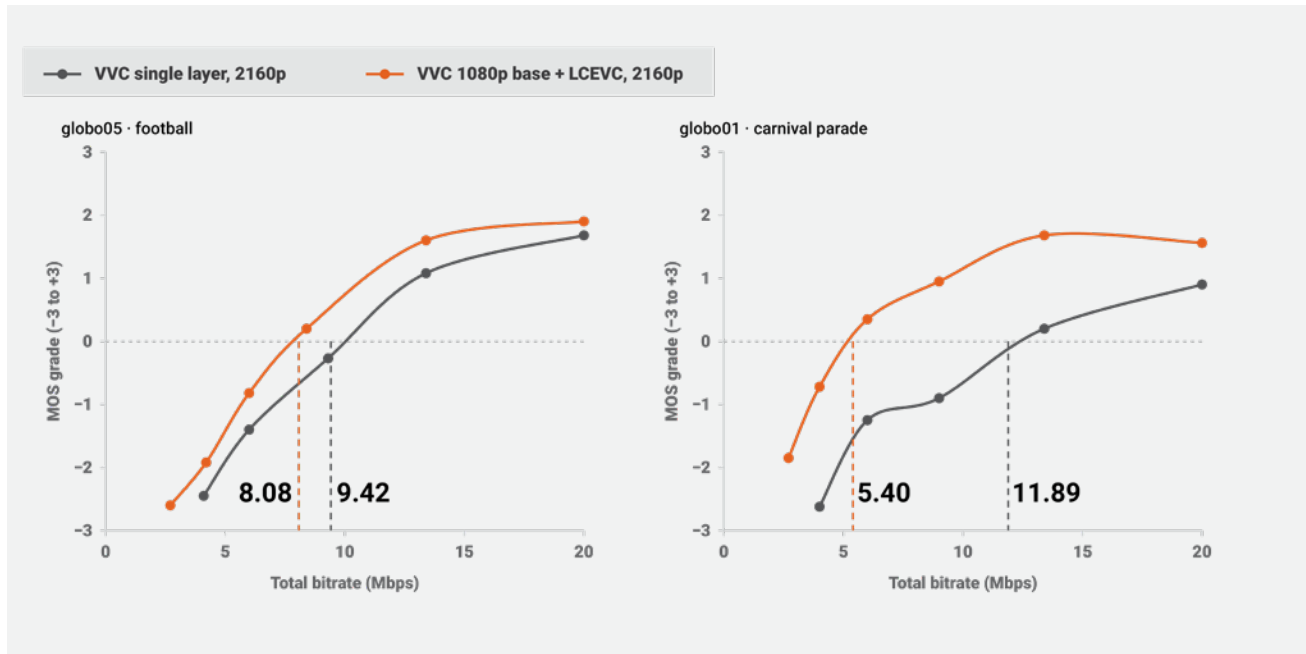
Data: ISO/IEC JTC 1/SC 29/WG 04 N0076 (2021), Tables 13 and 14. Charts recreated.



Brazil's SBTVD Phase 3 testing by the University of Brasília found that a single-layer 4K stream encoded with VVC needed 12.54 Mbps to match a 1080p reference. While, VVC plus LCEVC reached the same 4K resolution and quality target using just 8.08 Mbps, representing a 36 percent drop in bandwidth.

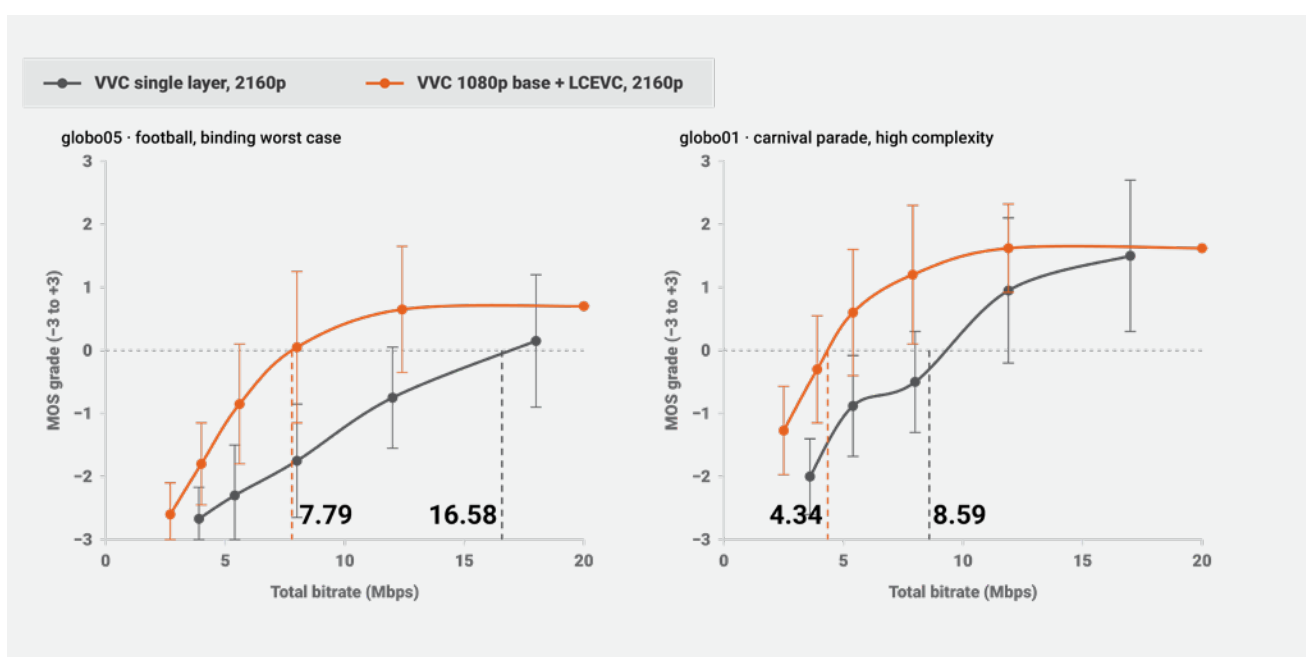
Figure 2. The consortium round that raised the right question.

Data: SBTVD Forum Phase 3, February 2024. Grade 0 equals the 1080p reference; confidence intervals overlap at adjacent points.



Further testing commissioned by V-Nova and performed by UnB using a MainConcept live encoder and a 120-frame GOP in both arms delivered even better results. Single-layer 4K VVC needed 16.58 Mbps to match the reference while VVC plus LCEVC required just 7.79 Mbps, 53 percent less at equal subjective quality, with per-content savings of 46 to 53.5 percent across five sequences.

Figure 3. Matched encoders: equal 4K quality at less than half the bitrate. Data: UnB Additional Tests for TV 3.0, August 2024. Error bars are published confidence intervals.



Newer and independent results agree with the SBTVD testing runs. V-Nova's peer-reviewed QoMEX 2026 VVenC study measured 23 percent savings at equal subjective quality with a 2.6 times faster encoding speed. A January 2026 West of Scotland preprint found LCEVC-enhanced VVC comparable to multilayer VVC, both ahead of base-layer upsampling. Kingston and TU Berlin measured minus 42 percent VMAF BD-rate versus x264 and minus 39 percent versus x265 for 1080p60 gaming. Surrey measured 18.7 and 15.8 percent savings over H.264 and HEVC at CRF 19, holding under packet loss.

TV 3.0, branded DTV+, became Brazilian law on August 27, 2025, and V-Nova reports commercial 4K HDR service under 10 Mbps in Rio de Janeiro, São Paulo, and Brasília since June 2026. FFmpeg, GStreamer, and Shaka Player carry LCEVC upstream; ATSC A/345 lists it as a NextGen TV option.

Table 1. *The published record, including adversarial results. Sources are listed at the end of this article.*

STUDY	YEAR	METHOD	HEADLINE RESULT	INDEPENDENCE
MPEG verification tests (WG04 N0076)	2021	BT.500 DSIS, 50 viewers, 2 labs	AVC -45.9% UHD / -28.5% HD; HEVC -30.9% / -24.1% (MOS BD-rate)	MPEG-supervised, independent labs
SBTVD Forum Phase 3 (Brazil)	2024	BT.500-15, 30 viewers per test	4K parity at 8.08 vs 12.54 Mbps; encoders differed between arms	Government-funded, run by UnB
UnB re-test, matched encoders	2024	BT.500-15, 30+ viewers per test	4K parity at 7.79 vs 16.58 Mbps, -53%; per-content -46 to -53.5%	V-Nova commissioned, UnB executed
Kingston Univ. + TU Berlin (ACM MHV)	2022	VMAF BD-rate + P.809 panel	-42.1% vs x264, -38.9% vs x265, 1080p60 gaming	Independent; SDK from V-Nova
University of Surrey (Springer MTA)	2024	VMAF, SSIM, PSNR; packet loss	-18.7% (AVC) / -15.8% (HEVC) at CRF 19; reverses at CRF 27/35 on AVC	Independent
Univ. of the West of Scotland	2026	DCR, 25 viewers, 15 UHD clips	LCEVC ≈ multilayer VVC; both beat base upsampling at the stronger operating point	Independent; bitstreams from V-Nova
V-Nova VVenC study (QoMEX)	2026	DCR at independent facility + metrics	BD-MOS -23%; VMAF -30%; PSNR +77%; 2.6× faster encode	Vendor-authored, peer-reviewed venue
MediaTek (Picture Coding Symposium)	2024	PSNR under JVET CTC + expert viewing	PSNR losses vs VTM/HM references; no visual benefit seen at 4K	Independent, adversarial
Leibniz Univ. Hannover (ICASSP)	2025	PSNR-domain BD-rate sweep	Optimized LCEVC still trails full-res VTM by ≈15% (PSNR)	Independent, adversarial

Objections, taken seriously.

“The gains are a metric artifact. Show me PSNR.”

Partly true: LCEVC optimizes perceptual reconstruction, and PSNR punishes it. The same VVenC encodes that saved 23 percent with human viewers and 30 percent on VMAF lost 77 percent on PSNR; VMAF correlated with those viewers at 0.89, PSNR at 0.49. MediaTek (PCS 2024) found PSNR losses against reference VTM and HM encoders and no visual benefit at 4K, and Leibniz University Hannover (ICASSP 2025) found optimized LCEVC trailing full-resolution VTM by roughly 15 percent in PSNR. Both use reference encoders, operate above delivery bitrates, and optimize the one metric LCEVC does not optimize. With human viewers at delivery bitrates, the record for LCEVC is consistent and validated.

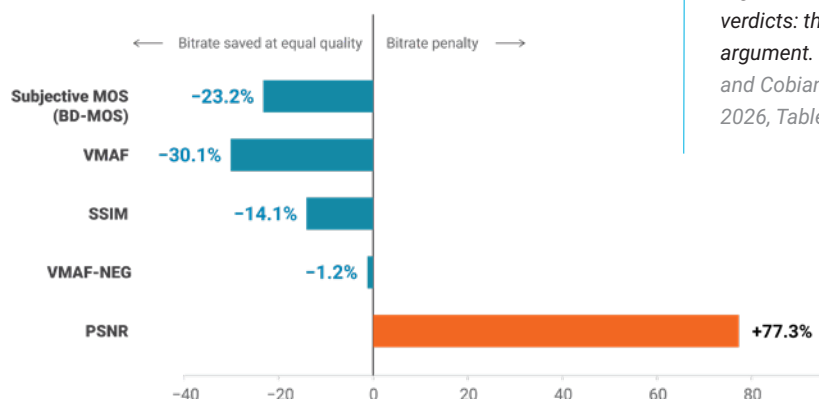


Figure 4. Same encodes, five verdicts: the metric decides the argument. Data: Razaak, Tahir, and Cobiachi (V-Nova), QoMEX 2026, Tables II to IV.

“Layered coding lost this argument once already. SVC and SHVC went nowhere.”

They did. SVC added overhead and complexity, and no ecosystem carried the cost. LCEVC inverts each failure mode: the enhancement is 9 to 14 percent of the stream, not a parallel prediction structure; encoding gets cheaper, 2.6 times faster for VVenC and about 40 percent fewer SVT-AV1 cycles (SPIE, 2022); and legacy decoders play the base layer. SVC’s costs landed before its benefits, while LCEVC runs the other direction.

“A good encoding ladder already downscales. This is upsampling with extra steps.”

MPEG tested this as a second anchor: the same codec at half resolution, upsampled. LCEVC beat it by 21 to 38 percent depending on codec and resolution (Figure 1, right panel); the Scotland study repeats the finding for VVC. Ladders that drop resolution have the right instinct. LCEVC adds a standardized channel for the detail the downscale threw away.

“Nothing wins everywhere. Where are the losses?”

Three boundaries are published. Operating point: Surrey’s AVC gains at CRF 19 reversed at CRF 27 and 35, and near-10-percent enhancement layers sometimes matched plain upsampling in the Scotland data. Content: per-content gaps across the Brazil rounds ranged from 14 to 55 percent.

“Are these objective or subjective benchmarks?”

The QoMEX study is V-Nova-authored, the UnB re-test V-Nova-commissioned, and the academic studies used V-Nova’s SDK since no other implementation exists. Against that: the verification tests ran under MPEG supervision, Brazil’s Phase 3 was government-funded, and the field publishes adversarial results. Formal subjective tests at delivery-relevant operating points land between 23 and 53 percent savings across five years. It is right to demand public protocols; but be careful not to ignore replicated results across disparate studies.

Two layers, two proofs.

The pairing is arithmetic: the base layer is the expensive part of an LCEVC stream, and the base layer is what a VPU accelerates. Silicon that encodes N 4K channels encodes roughly 4N 1080p base layers.

Adopting encoding-efficient hardware that delivers bandwidth savings must be at the top of every video team's technical agenda. Video platforms that only chase bitrate savings will struggle to produce efficient outputs economically. The stronger path evaluates both layers (compute and bandwidth efficiency) and holds each to the same priority. Compute efficiency gains that are given away by excessive delivery costs can no longer be defended.

Processing efficiency helps platforms create more streams with fewer resources.

Compression efficiency helps them move those streams with fewer bits.

Workflow efficiency decides whether either improvement can be used in production.



Sources

1. ISO/IEC JTC 1/SC 29/WG 04 N0076, Verification Test Report on the Compression Performance of LCEVC, 2021. [Report PDF](#). Peer-reviewed as Battista et al., IEEE Transactions on Circuits and Systems for Video Technology, 2023. ieeexplore.ieee.org
2. SBTVD Forum, TV 3.0 Project, Phase 3: Real-Time Video Coding Subjective Quality Assessment, February 29, 2024. forumsbtvd.org.br
3. University of Brasília, Additional Tests for TV 3.0, August 2024. Commissioned by V-Nova; report circulated through the SBTVD Forum ecosystem.
4. Razaak, Tahir, and Cobiainchi (V-Nova), Compression Efficiency and Complexity Evaluation of MPEG-5 LCEVC-enhanced VVenC for UHD Video Delivery, QoMEX 2026.
5. Barman, Schmidt, Zadtootaghaj, and Martini, Evaluation of MPEG-5 Part 2 (LCEVC) for Live Gaming Video Streaming Applications, ACM Mile-High Video 2022. doi.org/10.1145/3510450.3517279
6. Roberts and Adeyemi-Ejeye (University of Surrey), LCEVC performance and packet-loss study, Multimedia Tools and Applications, January 2024. doi.org/10.1007/s11042-023-17931-0
7. Ramzan et al. (University of the West of Scotland), subjective comparison of LCEVC-enhanced VVC and multilayer VVC for UHD, arXiv preprint, January 2026. arxiv.org/abs/2601.10448
8. Chubach et al. (MediaTek), LCEVC evaluation under JVET common test conditions, Picture Coding Symposium 2024. doi.org/10.1109/PCS60826.2024.10566455
9. Benjak and Ostermann (Leibniz University Hannover), LCEVC parameter optimization study, IEEE ICASSP 2025. doi.org/10.1109/ICASSP49660.2025.10889268
10. Cobiainchi et al. (V-Nova, Intel, Meta), Enhancing SVT-AV1 with LCEVC, SPIE Applications of Digital Image Processing XLV, 2022. doi.org/10.1117/12.2633882
11. Eyevin Technology, Evaluating LCEVC Decoding Performance on the Android Platform, March 9, 2023. eyevinntechnology.medium.com
12. TV Technology, Brazil Makes It Official: New DTV Standard Leverages ATSC 3.0 Tech, August 27, 2025. tvtechnology.com. V-Nova, DTV+ commercial launch report, June 11, 2026 (vendor source). v-nova.com
13. FFmpeg 7.1 release notes, September 2024 ffmpeg.org; GStreamer 1.26 release notes, March 2025 gstreamer.freedesktop.org; ATSC A/345:2026, VVC Video. atsc.org
14. V-Nova, Licensing Programs for MPEG-5 LCEVC, August 8, 2025 (vendor source). v-nova.com
15. Jan Ozer, Streaming Learning Center and Streaming Media coverage of LCEVC, 2020 to 2021, including a disclosed V-Nova consulting engagement for the earliest tests. streaminglearningcenter.com

The market isn't waiting for a better encoder.

It's waiting for infrastructure that removes the friction, delivers the economics and scales from today's codec stack to tomorrow's, without forcing a choice between cost and quality.



Test VPU encoding in the cloud. **Today.**

1 Create an account for a VPU-accelerated environment.

Deploy Quadra VPU VMs and run a sample of your real workloads in parallel with your current workflow. Setup in minutes and start today.

2 Run for two weeks. Measure your own numbers.

See for yourself the encoding efficiencies and economic gain between your two, identical and parallel workflows. From here, safely pick the deployment path that fits your stack.

3 Supported by NETINT and partner engineers.

You'll get immediate support plus an engineer on call if the pilot stalls. Use our VPU Deployment Playbook to guide your testing to migration process.

Start today: netint.com/cloud-pilot



netint.com